LOW-LATENCY BASS SEPARATION USING HARMONIC-PERCUSSION DECOMPOSITION

Ricard Marxer, Jordi Janer[*]

Music Technology Group
Universitat Pompeu Fabra
Barcelona, Spain
ricard.marxer@upf.edu, jordi.janer@upf.edu

ABSTRACT

Many recent approaches to musical source separation rely on model-based inference methods that take into account the signal's harmonic structure. To address the particular case of low-latency bass separation, we propose a method that combines harmonic decomposition using a Tikhonov regularization-based algorithm, with the peak contrast analysis of the pitch likelihood function. Our experiment compares the separation performance of this method to a naive low-pass filter, a state-of-the-art NMF-based method and a near-optimal binary mask. The proposed low-latency method achieves results similar to the NMF-based high-latency approach at a lower computational cost. Therefore the method is valid for real-time implementations.

1. INTRODUCTION

In the rhythm section of popular western music, the bass line often fulfills the role of anchoring the harmonic framework and laying down the beat. The sound produced by the bass is predominantly harmonic with a low fundamental frequency and usually has an impulsive excitation. Bass line estimation is a relevant case in musical source separation, since it can improve the separation of drums or the predominant melody from the mix.

In comparison to other instruments, bass line separation is difficult due to the low frequency and the presence of the bass drum that shares a similar spectrum distribution. [1] demonstrates the use of Non-negative Tensor Factorization for the isolation of the bass guitar among other instruments in multichannel synthetic mixtures. [2] employ their general source separation framework for the isolation of the bass line in professionally recorded music.

Nowadays, with increasing availability of music in online streaming services, it is often necessary to process audio data as it is received by the system. And with the increase of embedded devices in our everyday lives, limiting memory requirements is often important. These factors motivate the development of low-latency methods. We propose an extension to the method presented in [3] with modifications to the signal model in order to better represent bass line components. An evaluation is conducted where the proposed method is compared to a baseline naive method and to FASST, a state-of-the-art high latency and computationally expensive method.

[*] This work was supported by the Yamaha Corporation

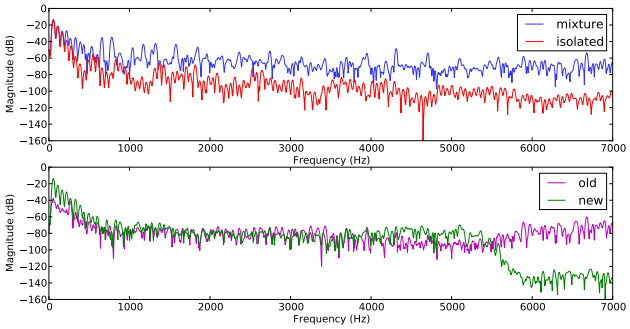

Figure 1: Example of a spectrum of the bass in a mixture and the bass in isolation *(top)*. Separated bass using the old signal model presented in [4] and with the new proposed signal model *(bottom)*.

2. METHOD

In [3] we introduced a low-latency drum separation method based on harmonic decomposition using single-frame Non-negative Matrix Factorization (NMF). An alternative to NMF in source separation is Tikhonov regularization, where the non-negativity constraint and some flexibility are sacrificed in favor of a having a closed-form non-iterative solution. Applying Tikhonov regularization to the factoring of the spectrogram requires fixing the basis matrix to constant spectrum templates. In [4] we present a signal model that contains spectrum patterns to represent both wideband and narrowband pitched components.

The bass guitar is mainly a pitched instrument, and at first sight the narrowband components in the signal model would seem to be sufficient. However bass drums quite often present a narrowband spectrum with a resonance of high magnitude and low frequency similar to that of the first partial of the bass guitar. Due to the low pitch of the bass guitar and the limited size analysis window of the STFT, the partials in its spectrum are often very close together (see Figure 1). This leads to a harmonic comb with less contrast. These components are similar to certain wideband components such as drums or sustained background noise. This causes problems especially in the high frequency range, where the bass spectrum has very low energy. To solve this issue, the signal model contains specific basis components for non-harmonic wideband spectra and the bass guitar components are constrained to represent their specific timbre.

2.1. Bass Specific Signal Model

We employ the same signal model used in [4], in which pitched sources are modeled as various components of band-filtered harmonic oscillators. Non-pitched sources are incorporated into the model as wideband noise components. The main modification to the signal model is to account for the usual spectral shape characteristics of the bass guitar and bass line in western music. The lowest note in a bass guitar is E1 (41.20Hz) and usually the pitch rarely goes higher than 120Hz. The harmonic envelope of the bass guitar is mainly restricted to the frequency range from 0Hz to 5000Hz.

To achieve this behaviour in our signal model, the pitch components that would correspond to the bass are limited in frequency by setting the magnitude of high frequency partials to zero. Using the same notation as in [4] we can redefine the source-filter model of the pitched components of the basis matrix by adding a function $a[l,\omega]$ that serves as an excitation envelope:

$$\begin{aligned} \vartheta[l,n] &= f_{low} H N_L \frac{2^{\frac{iH-N_T/2+n}{HN_L}}-1}{S_r \ln(2)} \\ E_l[\omega] &= \mathcal{F}\left\{\sum_{h=1}^{N_h} a[l,hf_l] sin\left(2\pi h \vartheta[l,n] n\right)\right\} \\ \boldsymbol{W}_{l,i}[\omega] &= \begin{cases} U_i[\omega]E_l[\omega] & \text{if } l \leq N_L \\ U_i[\omega] & \text{if } l = N_L+1 \end{cases} \end{aligned} \quad (1)$$

with $H = (1-\alpha)N_T$. Where α is a coefficient to control the frequency overlap between the components, N_T is the frame size, S_r the sample rate, $\mathcal{F}$ is the Discrete Fourier Transform (DFT), N_h is the number of harmonics of our components, $\boldsymbol{W}_{l,i}$ is the spectrum of the component of l^{th} pitch filtered by i^{th} filter. U_i is the spectrum of the i^{th} filter in our filterbank. U_i is constructed as a sequence of N_I Hann windows, linearly distributed in the Mel scale and with a 50% overlap. $f_l = \vartheta[l, N_T/2]$ is the center fundamental frequency of the l^{th} pitch. $\vartheta[l,n]$ is the instantaneous frequency function of the l^{th} pitch component.

In order to restrict the use of bass pitched components during the decomposition, we force their excitation envelope to a function decreasing to zero after a given cutoff frequency f_{cut}. The bass pitched components are defined as those having a fundamental frequency lower than $f_{0_{bass}}$:

$$a[l,\omega] = \begin{cases} r(\omega)/\omega & \text{if } f_l \leq f_{0_{bass}} \\ 1/\omega & \text{else} \end{cases} \quad (2)$$

where $r(\omega)$ is a function of ones that ramps down linearly to 0 from f^s_{cut} to f^e_{cut}:

$$r(\omega) = \begin{cases} 1 & \text{if } \omega \leq f^s_{cut} \\ 1 - \frac{\omega - f^s_{cut}}{f^e_{cut} - f^s_{cut}} & \text{if } f^s_{cut} < \omega \leq f^e_{cut} \\ 0 & \text{if } \omega > f^e_{cut} \end{cases} \quad (3)$$

In our experiments we fix the size of the ramp and only control the start frequency $f^s_{cut} = f_{cut}$ and $f^e_{cut} = 1.3 f_{cut}$.

2.2. Bass Source Estimation

Using Tikhonov regularization as in [4] with the modified signal model, we can derive the pitch likelihood L from the gains vector $\hat{\boldsymbol{H}}^{TR}$. The next step is the selection of the components belonging to the bass line.

Instead of using a pitch tracking algorithm as in [5] that would add complexity and latency to the method, we rely here on a simple peak detection and picking algorithm. The proposed method is simple and has a low computational cost. In order to select the pitch of the bass line, at every frame we select the highest peak in the pitch likelihood function under a certain frequency value $f_{0_{bass}}$. We assume that only one pitched source will be present in this low frequency range, and that this source will be the targeted bass guitar or bass line.

The peak picking is performed by selecting local maxima in the pitch likelihood $\boldsymbol{L}$:

$$\omega_i \in \left\{\omega | \underset{j=\omega-W_\omega \ldots \omega+W_\omega}{\arg\max} \boldsymbol{L}(j) \text{ and } \boldsymbol{L}(\omega-1) < \boldsymbol{L}(\omega) \geq \boldsymbol{L}(\omega+1)\right\} \quad (4)$$

where $2W_\omega$ is the size of the local neighborhood for the peak local maxima.

However as we previously explained, the basis components of the bass described in Section 2.1 are similar to those of wideband components such as the bass drum or other background sources present in the low frequency range. This leads to pitch likelihood functions with a high energy distribution in the low pitch components that do not necessarily correspond to pitched instruments.

Pitched sources in the spectrum can be modeled as gaussians in the pitch likelihood function $\boldsymbol{L}$. The width of the gaussian is related to the chirp ratio of the fundamental frequency of the pitch. If the source is wideband and not pitched (e.g. drums), it can be regarded as a limit case of the partials widening and forming a smooth spectrum with no harmonic structure. Empirical observations show that wideband non-pitched sources that are not decomposed into the wideband components of our signal model, appear as wide noisy gaussians in the pitch likelihood function.

To distinguish between pitch likelihood peaks corresponding to a pitched bass and those related to other wideband sources, for each pitch likelihood peak p we define a measure of peak contrast c_p. The peak contrast feature is computed using the difference between the height of the peak and the likelihood of the local minima around it:

$$c_p = \max\left(\boldsymbol{L}(\omega_p) - \boldsymbol{L}(\omega^l_p), \boldsymbol{L}(\omega_p) - \boldsymbol{L}(\omega^r_p)\right) \quad (5)$$

where ω_p is the position of the p^{th} peak, ω^l_p is the first local minima under ω_p and ω^r_p is the first local minima over ω_p.

The bass component in the pitch likelihood ω_b is defined as the position of the highest peak, with frequency under $f_{0_{bass}}$ and whose contrast is over a given threshold $\boldsymbol{L}_{th}$.

As in [3] we create a new vector $\hat{\boldsymbol{H}}_b$ containing non-zero values only at those bins corresponding to the selected bass pitch.

$$\hat{\boldsymbol{H}}_b[\omega] = \begin{cases} \hat{\boldsymbol{H}}[\omega] & \text{if } |\omega_b - \omega| < \Delta\omega \\ 0 & \text{otherwise.} \end{cases}$$

where $\Delta\omega$ controls the amount of selected pitch components around ω_b. Therefore, we can compute the bass signal estimation as $|\hat{\boldsymbol{X}}_b| = \boldsymbol{W}\hat{\boldsymbol{H}}_b$. In a complementary fashion, the reconstruction of the non-harmonic part takes a gains vector $\hat{\boldsymbol{H}}_{nb}$ containing non-zero values for the unselected bass pitch plus the wideband filter banks. The non-harmonic source estimation is computed as $|\hat{\boldsymbol{X}}_{nb}| = \boldsymbol{W}\hat{\boldsymbol{H}}_{nb}$.

With the estimated magnitude spectra $|\hat{\boldsymbol{X}}_b|$ and $|\hat{\boldsymbol{X}}_{nb}|$ we perform a Wiener filtering to obtain the mask that isolates the bass

component:

$$\boldsymbol{m}_b = \frac{|\hat{\boldsymbol{X}}_b|^2}{|\hat{\boldsymbol{X}}_b|^2 + |\hat{\boldsymbol{X}}_{nb}|^2} \tag{6}$$

Finally the estimated bass spectrum is simply the result of multiplying the input complex spectrum with the previously presented mask $\hat{\boldsymbol{X}}_b = \boldsymbol{m}_b \otimes \hat{\boldsymbol{V}}$. The output time-domain signal is recovered by means of the inverse STFT and an overlap-add process.

3. EVALUATION

We employ the evaluation techniques used in community evaluation campaigns such as SiSEC [6] to measure the performance of the proposed method. We compute the following measures using the BSSEval toolbox: SDR (Signal to Distortion Ratios), SIR (Source to Interference Ratios) and SAR (Sources to Artifacts Ratios). Evaluation material consists of a dataset of 12 multi-track recordings containing bass guitar or a bass line compiled from publicly available resources (MASS[1], SiSEC[2] and two in-house professional recordings. The audios were downmixed to mono to avoid using pan information in the separation, since that is out of the scope of this work. The sampling rate of the audio examples is 44.1 kHz, and the spectral analysis uses a frame size of 4096 and a hop-size of 512 samples.

The proposed method, Tikhonov Regularization Bass Separation (TRBS), is compared to several existing techniques. A low frequency filter (LOWP) is used as a baseline trivial method. The publicly available implementation of FASST[3] [2] serves as a state-of-the-art high-latency option. Finally an oracle separation [7] using a binary mask is tested as a glass ceiling for spectral bin classification techniques [5]. We compared each method to a reference obtained with the soft mask oracle separation. All values presented are error measures: the difference between the soft mask oracle estimation measure and the measure of each algorithm. Thus, the lower the value the closer it is to the oracle estimator meaning better quality.

In a first experiment we perform a parameter exploration for the LOWP and TRBS methods. For the low pass filter we studied the effect of the cutoff frequency. For the TRBS method we studied the effect of varying parameter $f_{0_{bass}}$ that controls the threshold under which a pitch may be considered as belonging to the bass. A second experiment consisted of a comparative study of all the selected methods, where the best parameters for the LOWP and TRBS methods were used.

4. RESULTS

In Figure 4 we observe that the artifacts error (SAR) of the LOWP method is very low. This is expected because the low pass filter does not add new components such as musical noise. The frequency response of the LOWP method is very smooth compared to all other methods, including the oracle soft mask that is used as reference, which explains the negative value of this error measure. We also see that the interference error (SIR) of the LOWP method is high. This method is used as a trivial baseline, and in fact it does not target the bass guitar source, it simply separates low frequency components without making any discrimination. Another observation to be made is that the cutoff frequency parameter controls the

[1] `http://www.mtg.upf.edu/static/mass`
[2] `http://sisec.wiki.irisa.fr/`
[3] `http://bass-db.gforge.inria.fr/fasst/`

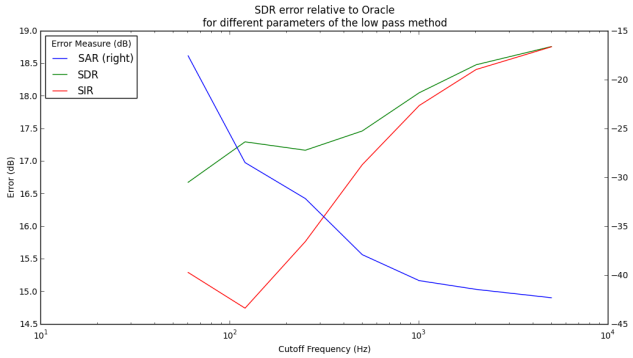

Figure 2: Average error measures for various values of the cutoff frequency parameter (in Hz) of the LOWP method.

tradeoff between artifacts and interferences. A local minimum of distortion error (SDR) is found at 250Hz, even though the average error continues to descend to 75Hz. The results of the individual excerpts, not presented here, show that the SDRs of some songs increase significantly between 250Hz and 120Hz. For that reason 250Hz was chosen as the optimal parameter value.

Table 1 and Figure 3 show best performance for both artifacts errors (SAR) and interference errors (SIR) when the $f_{0_{bass}}$ parameter is around 100Hz. Therefore 100Hz was selected as the best parameter value for $f_{0_{bass}}$ for the comparative study. We see a significant difference in the errors ($\approx$ 3 dB) depending on the parameter value. This leads us to think that a more evolved pitch selection method could further reduce separation error.

	SAR	SDR	SIR
60	12.89	13.14	14.03
70	10.41	13.08	13.54
80	8.43	12.87	12.37
90	6.85	13.01	12.41
100	6.13	12.95	12.40
110	6.43	12.96	12.34
120	6.63	13.05	12.50
130	7.22	13.37	13.18
140	7.73	13.46	13.65
160	8.15	13.83	14.13
180	8.54	13.98	14.60

Table 1: Average error measures for various values of $f_{0_{bass}}$ parameter (in Hz) of the TRBS method.

Several conclusions can be drawn from the results of the comparative study. Table 2 and Figure 4 show that the proposed method performs similarly to state-of-the-art techniques such as FASST. While FASST achieves a lower artifact error (SAR) separation, TRBS has less interference error (SIR). Another observation is that the oracle binary mask scores a slightly negative SIR error measure. This means that on average the binary mask produces less interference than the soft mask oracle. However, this improvement is balanced by the artifacts error (SAR), where the oracle binary mask reveals the highest error level of all methods.

In Figure 5 we can see that this behavior is consistent on all

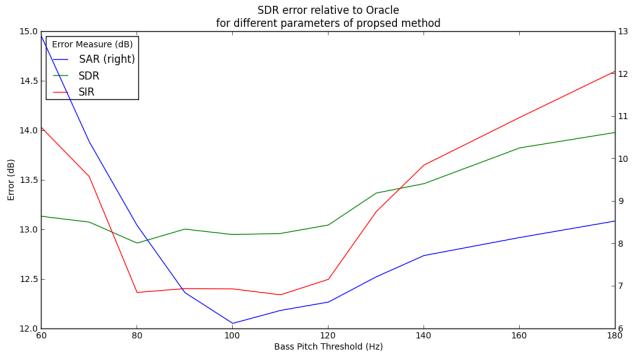

Figure 3: Average error measures for various values of $f_{0_{bass}}$ parameter (in Hz) of the TRBS method.

	SAR	SIR	SDR
LOWP-250	-32.14	15.77	17.17
FASST	4.35	15.02	14.33
TRBS-100	6.13	12.40	12.95
ORACLEBIN	9.26	-1.98	6.53

Table 2: Average error measures for the evaluated algorithms.

the individual excerpts. In listening to the separated sources, we found that these quantitative results seem to correctly reflect the perceived differences between the methods. A web page[4] with audio examples illustrate the results obtained with our method.

5. CONCLUSION

We have shown that the Tikhonov regularization spectrum decomposition method can be successfully used to perform low latency bass guitar/base line separation of western music signals. Furthermore the use of pitch likelihood peak contrast and specific bass timbre models allows us to produce separation results comparable to state of the art high latency methods, such as FASST. Quantitative results show the accuracy of the separation in contrast to baseline trivial methods such as low pass filters. The proposed method was also compared to approximations of the best possible performance of binary masks by using BSS oracle techniques.

6. REFERENCES

[1] Derry FitzGerald, Matt Cranitch, and Eugene Coyle, "Non-negative tensor factorisation for sound source separation," in *PROCEEDINGS OF IRISH SIGNALS AND SYSTEMS CONFERENCE*, 2005.

[2] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sept. 2010.

[3] J. Janer, R. Marxer, and K. Arimoto, "Combining a harmonic-based NMF decomposition with transient analysis for instantaneous percussion separation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 281–284.

[4] Ricard Marxer and Jordi Janer, "A Tikhonov regularization method for spectrum decomposition in low latency audio source separation," in *Proceedings IEEE Int. Conference on Acoustics, Speech and Signal Processing (ICASSP'2012)*, Mar. 2012.

[5] Ricard Marxer, Jordi Janer, and Jordi Bonada, "Low-Latency instrument separation in polyphonic audio using timbre models," *Latent Variable Analysis and Signal Separation*, pp. 314–321, 2012.

[6] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[7] E. Vincent, R. Gribonval, and M.D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.

[4] `http://www.dtic.upf.edu/~rmarxer/dafx13/bass`

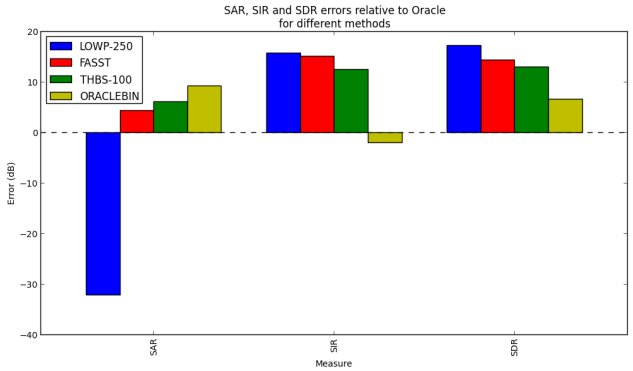

Figure 4: Average error measures (x-axis) for the evaluated algorithms. The LOWP-250 method presents very low artifacts (SAR) but a worse global separation (SDR).

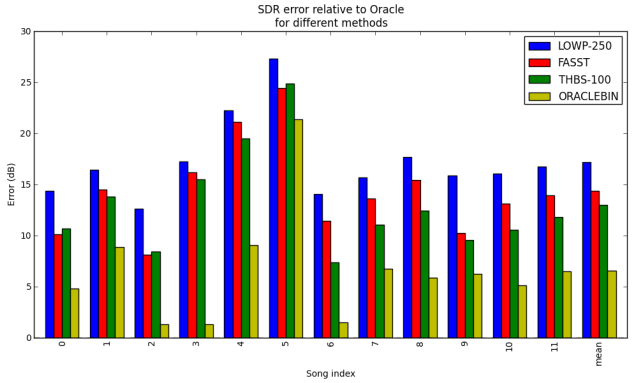

Figure 5: SDR error measures of individual audio examples for the methods.