A BASS LINE TRANSCRIPTION ALGORITHM FOR POLYPHONIC MUSIC

Fabian Seipel
Zplane GmbH, 10823 Berlin, Germany
fabian.seipel@gmail.com

ABSTRACT

This paper presents an algorithm to detect bass lines in polyphonic music pieces. Several analysis methods from various research work in the field of music information retrieval and signal processing were assembled to an overall system. A database including tracks from various genres, tempo and key with bass line annotations was utilized for evaluation purpose. The algorithm achieves recall and precision note tracking scores situated around 60-75% for 50/100 ms note onset ranges and frame level evaluation scores of 80/78% voiced/total pitch detection as well as 85/82% for voiced/total chroma detection.

1. INTRODUCTION

Automatic music transcription has significantly gained importance over the last 15 years. Due to larger data storage possibilities and improved computational power, algorithms in this field of research have improved but still cannot compete with human experts and consequently demand future research and further development [1]. Automated notation can provide powerful insights for sophisticated music information retrieval applications, involving topics like music education, live performance as well as recording or production contexts, musicological research, large scale annotation of legacy content and more [2].

Within this matter Goto [3] has addressed the issue of adequate representation systems. Most common existing symbolic notations such as musical scores or the MIDI standard define a single note by discrete pitch, an onset time and duration but discard additional information regarding expressiveness in performance. However, for certain tasks, abstracted notation systems contain all relevant information and therefore justify their employment.

Bass lines are part of the melodic construct of nearly all musical pieces in western tradition throughout a variety of genres. According to Goto [3] they are closely related to a piece's tonality and thus can give insights to these chord structures. Nevertheless there are several criteria to distinguish those from other harmonic and melodic content of a musical piece. Unlike lead melodies, a bass line is characterized by a monophonic combination of notes in lower register, mostly played by a single instrument like bass guitar, double bass, tuba, or a bass synthesizer [4]. Depending on the piece, the tonal scope of the temporal trajectory is limited to a certain range of two or three octaves at most [5]. The harmonic content of the bass instrument however can include overtones in higher ranges. These assumptions serve as criteria for estimating parameters and conditions throughout the algorithmic development.

Several different concepts for bass line estimation from polyphonic audio have already been proposed. Goto [6] [7] has presented methods employing an Expectation-Maximization algorithm on the basis of harmonic frequency content combined with a multiple agent tracking system which relies on temporal continuity of the fundamental frequency (F0). Hainsworth and Macleod's system [5] first locates potential onsets by peak detection and estimates corresponding F0's with a harmonic comb and confidence measure. Hereupon tracking methods select promising F0 candidates if according partial frequencies are found before finally selecting the fundamental by maximizing note confidences based on the previous found onset intervals. Ryynänen and Klapuri's [8] approach employs a frame-wise pitch-salience estimator for feature extraction. These features serve as an input to derive both an acoustic model for bass and rest notes by hidden-markov- and gaussian-mixture-approaches as well as a musicological model based on note bigrams and variable-order-markov-models to estimate the musical key. Transcription is finally obtained by Viterbi decoding. Salamon and Gómez [9] apply salience functions derived on chroma features by adapting harmonic pitch class profiles to extract mid-level representations to determine discrete pitches of the bass line.

This paper proposes an alternative method for bass line detection, whereas the procedure does not introduce new algorithmic concepts but rather assembles already existing ones from different authors to an improved overall system. This includes methods like key and tuning frequency estimation, beat tracking, constant-q-transformation (CQT) [10][11], harmonic percussive separation (h-p-Sep) using median filtering [12], harmonic sum spectrum (HSS) [13], adapted tracking algorithms [14][6][7] and final post processing steps before rendering the results to a midi formatted output.

2. METHODS

Figure 1 displays the processing steps of the algorithm. Regarding data reduction and enhancement of computational time, multichannel audio material was summarized to monophonic content and normalized in a pre-processing step in the time domain. Key and tuning frequency estimation provide tonal information for CQT frequency analysis and optional filtering by key before applying the tracking algorithm. Beat tracking offers the possibility of mapping important tracking parameters to the bpm count as well as creating a grid for quantization. Both algorithms were implemented as state-of-the-art SDKs [1].

2.1. Constant-Q-Transform

Frequency representation was obtained by constant-q-transform (CQT). Opposed to the Fourier transform, the constant-q-approach creates bins of geometrically (instead of linearly) spaced center frequencies which produces a much higher resolution towards lower tonal content. Additionally the geometrical spacing of frequencies aligns with the musical scale of equal temperament and consequently qualifies the CQT especially for audio analysis. Compared

[1] http://licensing.zplane.de/technology

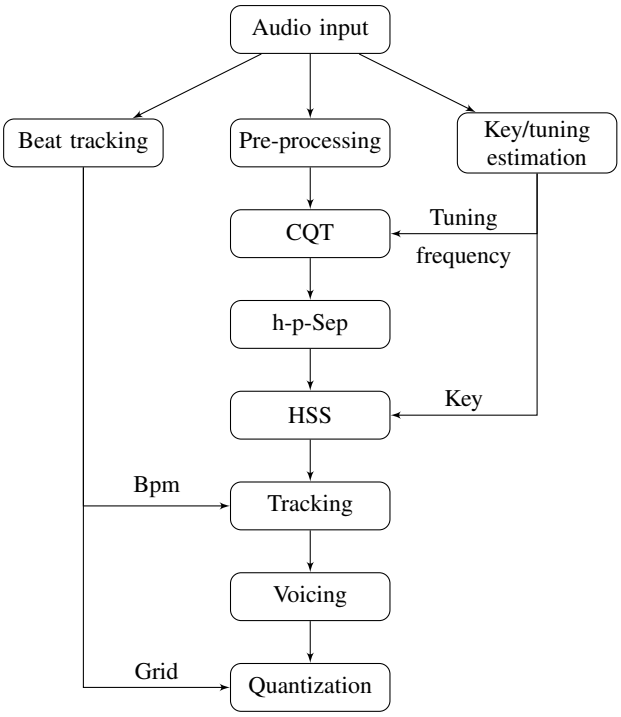

Figure 1: Flowchart of bass line estimation algorithm

to the widespread discrete Fourier transform (DFT), the CQT comes with the disadvantage of increased computational expense and lacks an inverse transform for perfect signal reconstruction. For mere analysis purposes the above mentioned CQT advantages predominate for this task.
Given a discrete time-domain signal $x(n)$, the CQT in [10] is defined as:

$$X^{CQ}(k,n) = \sum_{j=n-[N_k/2]}^{n+[N_k/2]} x(j)\ a_k^*\ (j-n+N_k/2) \quad (1)$$

where $k = 1, 2, ..., K$ indexes the frequency bins and $a_k^*(n)$ denotes the complex conjugate of $a_k(n)$, the complex valued waveform basis function:

$$a_k(n) = \frac{1}{N_k}\ w(\frac{n}{N_k})\ exp[-i2\pi n \frac{f_k}{f_s}] \quad (2)$$

including the sample rate f_s, a continuous window function $w(t)$, the window length N_k:

$$N_k = \frac{q f_s}{f_k(2^{\frac{1}{B}} - 1)} \quad (3)$$

and the the center frequency f_k of bin k:

$$f_k = f_1\ 2^{\frac{k-1}{B}} \quad (4)$$

where f_1 is the center frequency of the lowest bin, $0 < q \leq 1$ denotes a scaling factor and B defines the number of bins per octave, which also determines the time frequency resolution tradeoff. [10] [11]

2.2. Harmonic-percussive separation

A main challenge in obtaining transparent representations of bass notes in the CQT spectrum are interfering low frequency signals like bass or snare drums. In order to separate those transient proportions of the spectrum, a harmonic-percussive separation algorithm[2] was employed [12]. Median filtering in the horizontal direction of the CQT magnitude spectrum removes impulse like signal components along the frequency axis:

$$H(k,n) = median(|X^{CQ}(k, n - l_{Harm})|, ... \\ |X^{CQ}(k, n + l_{Harm})|) \quad (5)$$

whereas the pendant in the vertical direction equalizes steady harmonic signal components along the time axis:

$$P(k,n) = median(|X^{CQ}(k - l_{Perc}, n)|, ... \\ |X^{CQ}(k + l_{Perc}, n)|) \quad (6)$$

The filter length is determined by $2l_{Harm} + 1$ or respectively $2l_{Perc} + 1$ for the two cases. Given these filtered spectra, soft masks based on Wiener Filtering are generated as follows:

$$M_H(k,n) = \frac{H(k,n)^p}{H(k,n)^p + P(k,n)^p} \quad (7)$$

$$M_P(k,n) = \frac{P(k,n)^p}{H(k,n)^p + P(k,n)^p} \quad (8)$$

where p denotes the power to which each individual element of the spectrogram is raised. Complex overall spectra are then recovered by element wise multiplication of mask and original spectrum X^{CQ}:

$$\hat{H}(k,n) = X^{CQ}(k,n) \cdot M_H(k,n) \quad (9)$$

$$\hat{P}(k,n) = X^{CQ}(k,n) \cdot M_P(k,n) \quad (10)$$

Subtracting a certain amount of the resulting percussive spectrum from the harmonic equivalent as a possible post processing step to clear remaining transient components can in some cases help to prepare the overall spectrum for further analysis procedures:

$$H_{post}(k,n) = \hat{H}(k,n) - w \cdot \hat{P}(k,n) \quad (11)$$

where w describes the factor of the subtraction amount. Negative results were discarded and set to zero. Low pass filtering the H_{post} spectrum was used to suppress higher harmonic content in preparation for the harmonic sum spectrum method. The cut off frequency was dynamically determined based on the power distribution in H_{post}. Figure 2 shows the spectrum of H_{post} and $\hat{H}$ for an example track.

[2] derived from:
https://www.audiolabs-erlangen.de/resources/MIR/TSMtoolbox/

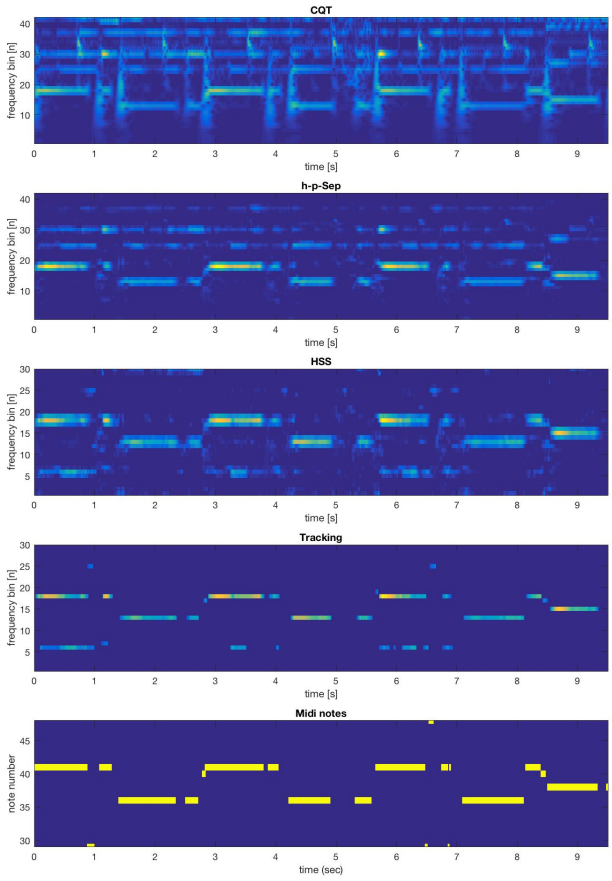

Figure 2: Results of individual processing steps: CQT, harmonic-percussive-separation, HSS, tracking, midi notes

2.3. Harmonic sum spectrum

Estimation of fundamental frequencies (F0) is based on the principle that every bass note is represented by its individual harmonic overtone series in the processed harmonic spectrum H_{post}. The salience of a F0 is equal to the accumulated power magnitudes of all m-th harmonic components, weighted by a factor $g(k, m)$:

$$HSS(k, n) = \sum_{m=1}^{M} g(k, m) H_{post}(m \cdot k, n) \qquad (12)$$

An efficient way to calculate this salience has been presented by Klapuri [13] as iterative estimation and cancellation approach, a scheme which was also adapted herein. With a given set of partial weights $g(k, m)$, combinations of F0 plus overtones which contain the most power in H_{post} are identified. In a second step, those frequency bins $m \cdot k_{F0}$ are reduced by a certain share of its magnitude according to the partial weights $g(k, m)$. Finally, the subtracted amount of power is saved to the HSS spectrum at the F0 frequency bin k_{F0}. In other words, the algorithm gradually punches out a grid of harmonics from the H_{post} spectrum to calculate the harmonic sum spectrum. Optionally, H_{post} bins which do not belong to the key of the piece were discarded.

2.4. Tracking

Even though the processing steps up to now have tried to single out a transparent bass line, the HSS spectrum is still spoiled by various interferences due to the occurrence of other harmonic instruments in lower register, bass drum hits as well as fluctuations due to intonation of the bass instrument or vibrato effects. The concept of tracking has been developed to estimate continuous note events, which permits small frequency deviations in a certain range Δ. McAulay and Quartieri [14] have introduced a frame-to-frame peak matching system which has been adapted and applied to the HSS output. The system is divided into five tasks that are executed for every frame.

1. Peak localization:
Peak frequency locations in each frame n are defined as local maximums given the condition of a minimum peak value.

2. Start tracks:
If a peak is found in frame n at frequency bin k, also denoted $k_{P,i}(n)$, but there is no track existing at frequency bin $k_{T,j}(n) \pm \Delta$ in the previous frame, then start a new track.

3. Continue tracks:
Given the case, a peak has been found under the circumstance $|k_P(n) - k_T(n-1)| \geq \Delta$, then continue the existing track.

4. Put tracks to sleep:
If no peak exists in a frame that matches the condition $|k_P(n) - k_T(n-1)| \geq \Delta$ for a given track, put this track to sleep. Sleeping tracks are continued a certain amount of frames n_{Sleep} and become active again if there is a matching frequency in a following frame.

5. Kill/Save tracks:
If a track exceeds a maximum number n_{Sleep} of frames in sleeping mode, this track is terminated. Tracks are saved if the number of total frames they have lived exceeds a minimum value of frames n_{min}, otherwise they are discarded.

This routine is repeated for all frames. Important parameters like n_{Sleep} or n_{min} are mapped to the tempo of the audio material.

2.5. Voicing and Quantization

By optimizing the parameters, tracking methods already indicate unvoiced sections of the bass line in many cases (see 2). An additional criteria based on power threshold was employed to further sort out low energy tracks in unvoiced parts. Since bass line notes often occur on fractions of a beat, quantization to a certain beat grid seems desirable. Nevertheless bass line melodies especially in genres like jazz build upon expression and improvisation, leading to occurrence of note events in between subdivisions of beats. Therefore, the utilization of quantization needs to be considered case wise depending on the context, therefore it has been implemented optionally. Figure 2 displays the results of the individual processing steps from CQT analysis to midi output format.

	NTS (50ms)			NTS (100ms)			FLE	
	R	P	F	R	P	F	Voiced	Total
Pitch	60.90	65.53	63.13	70.10	74.74	72.35	80.48	77.61
Chroma	62.56	67.11	64.76	72.28	76.98	74.56	85.08	81.85

Table 1: Overview of different performance scores for pitch and chroma prediction in percent: Note tracking score for 50ms and 100ms onset range and Frame Level Evaluation for voiced as well as the total number of frames

3. EVALUATION

The performance of the detection algorithm was evaluated on a self-constructed dataset created by the band-in-a-box software[3], containing 74 pieces of various genres, keys, tempo and bass line instruments with midi bass annotation. In total the test data contained approximately four hours of polyphonic music. Metrics for evaluation are adapted from the MIREX competition for multiple fundamental frequency estimation and tracking. Different scores were employed for rating the algorithm performance. The Note Tracking Scores (NTS) include recall (R), precision (P) and f-Score (F) measure:

$$R = \frac{\#\text{correctly transcribed notes}}{\#\text{reference notes}} \tag{13}$$

$$P = \frac{\#\text{correctly transcribed notes}}{\#\text{transcribed notes}} \tag{14}$$

$$F = \frac{2RP}{R+P} \tag{15}$$

Reference notes were extracted from the annotated midi bass line of the given ground truth. The onset of a correctly transcribed note must lie within a range of $\pm$50ms or respectively $\pm$100ms and must not deviate more than half a semitone in pitch in comparison to its reference.
The Frame Level Evaluation score (FLE) compares the active pitch of the detection system for every frame with its according ground truth. A returned pitch is assumed to be correct if it is within half a semitone of the ground-truth pitch for that frame. Only one detected bass line pitch per frame can be associated with the ground truth. The score was calculated once only for voiced frames but also for the total amount of frames including the unvoiced parts. Both metrics (NTS and FLE) were evaluated for total pitch as well as chroma. Table 2.4 displays the overall mean results for each score for the whole dataset.

4. RESULTS

Both recall and precision note tracking scores for an onset range of 50ms show accuracies of 60-65%, whereas the chroma measures for this onset setting are about two percent higher. By widening the tolerance to 100ms, all scores rise about ten percent. For different algorithm parameters either recall or precision may be improved at the expense of the other result, while the f-score stays at the same level. The presented method further achieved frame level evaluation outcomes situated around 80 percent. Differences from chroma to pitch values denote five percent points and the discrepancy from total to voiced frames amounts to four percent. The latter gives insights to the voicing detection accuracy which can be estimated by these numbers.
The algorithm was implemented with the Matlab 2016a software on a 2,3 GHz Intel Core i7 processor. To compute the results of a one minute audio track, this machine needed nearly 22 seconds. For a two minute track, required time was about 40 seconds. Hence, execution speed scales linearly.

[3] http://www.bandinabox.com/

5. DISCUSSION

The interpretation of the results on hand needs to take several possible reasons for performance decrease into account. Most common misclassification errors are ascribable to octave confusions, low bass line power distribution, interference by other harmonic instruments in lower register, bass/snare drum occurrence and notes that due to intonation lie in between frequency analysis bins of CQT and HSS.
First of all, bass line sounds can vary immensely in several facets. Bass instruments differ in terms of timbre and their temporal envelope, especially regarding onsets and the decay of harmonics. Playing techniques and melodies further widen the spectrum of possible bass sounds. Developing a single detection algorithm which covers all these possibilities is challenging. The HSS method detects harmonic structures depending on the given weighting of partials $g(k, m)$. This function employs a pre-defined set of weights for all bass lines to analyze. A dynamic approach to this problem by analyzing the bass instrument's sound structure beforehand to adapt the weights individually could improve HSS results.
Octave confusions, as a second major source of errors, may as well be reduced by this idea. Since the HSS iteratively subtracts partial structures depending on their total amount of power, fundamental frequencies may falsely be denoted to an octave below or above due to disadvantageous pre-defined harmonic weight sets. Introducing a case distinction in advance to categorize the music piece towards genre or type of bass instrument would again be one way to enhance the system's performance. Another possible solution may include source separation methods. Analyzing the harmonic spectrum of a segregated time representation of the bass line could directly inform and define the weighting function $g(k, m)$ and improve HSS performance.
The octave confusion error can be quantified as the difference of chroma and pitch scores, which is nearly five percent for FLE scores and about two percent for all NTS recall and precision measures.
Another main factor that led to performance decrease especially in terms of NTS scores are onset misses. As the discrepancy in NTS

scores for difference onset intervals (50ms and 100ms) shows, the algorithm obviously fails to detect a high percentage of note onset in the given time frame. On a closer look, these errors almost always occur because of late detection. Bass line notes often take place on drum beats instead of in-between. Because of their transient behavior, kick and snare drums are represented by short and vertical lines in the frequency spectrum. In those parts the HSS analysis fails to detect correct harmonic sums due to the overlap of these percussive elements. Filtering harmonic components helps to avoid those disambiguities and hence improves bass line detection. However, by dividing the power spectra in those two components, transient parts of the bass line are sometimes falsely ascribed to the percussive spectrum. For this reason onsets may be detected too late which leads to reduced NTS recall scores.
Other approaches for bass line detection have proposed separate onset modeling methods independent from the actual pitch detection to combine those information at the end for final transcription to prevent the above described onset misses [4] [5].
Further ideas for improvement include a more sophisticated Q-transform that builds on variable instead of constant Q factors leading to sharper time resolutions for lower frequency bins [15] and the opportunity to include bass line transition probabilities [6] [4].

6. CONCLUSION

This paper presented an algorithm for bass line detection and transcription. Employed methods include key and tuning frequency estimation, beat tracking, constant-Q-transform, harmonic-percussive-separation, harmonic sum spectrum, fundamental frequency tracking concepts, voicing and quantization. Performance was evaluated on a self constructed database containing nearly 4 hours of music from different genres, key and tempo. Scores and accuracies were calculated according to the MIREX standard for multiple fundamental frequency estimation and tracking. Results show decent performance in comparison to other efforts in this domain. The algorithm is supposed to be embedded in an overall system for music transcription. In this means the bass line detection may especially support and improve chord estimation and melody extraction.

7. REFERENCES

[1] Emmanouil Benetos, Simon Dixon, Dimitrios Giannoulis, Holger Kirchhoff, and Anssi Klapuri, "Automatic music transcription: challenges and future directions," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 407–434, Dec. 2013.

[2] Anssi Klapuri and Manuel Davy, *Signal processing methods for music transcription*, Springer Science & Business Media, 2007.

[3] Masataka Goto, "A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-world audio signals," *Speech Communication*, vol. 43, no. 4, pp. 311–329, 2004.

[4] Matti P. Ryynänen and Anssi P. Klapuri, "Automatic transcription of melody, bass line, and chords in polyphonic music," *Computer Music Journal*, vol. 32, no. 3, pp. 72–86, 2008.

[5] Stephen W. Hainsworth and Malcolm D. Macleod, "Automatic bass line transcription from polyphonic music," in *Proceedings of the International Computer Music Conference*. 2001, Citeseer.

[6] Masataka Goto and Satoru Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," in *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, 1999, pp. 31–40.

[7] Masataka Goto, "A robust predominant-F0 estimation method for real-time detection of melody and bass lines in CD recordings," in *International Conference on Acoustics, Speech, and Signal Processing*. 2000, vol. 2, pp. II757–II760, IEEE.

[8] Matti Ryynanen and Anssi Klapuri, "Automatic bass line transcription from streaming polyphonic audio," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*. 2007, vol. 4, pp. IV–1437, IEEE.

[9] Justin Salamon and Emilia Gómez, "A chroma-based salience function for melody and bass line estimation from music audio signals," in *Proc. of Sound and Music Computing Conference (SMC)*. 2009, pp. 331–336, Citeseer.

[10] Christian Schörkhuber and Anssi Klapuri, "Constant-Q transform toolbox for music processing," in *7th Sound and Music Computing Conference, Barcelona, Spain*, 2010, pp. 3–64.

[11] Judith C Brown, "Calculation of a constant q spectral transform," *The Journal of the Acoustical Society of America*, vol. 89, no. 1, pp. 425–434, 1991.

[12] Derry Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of the 13th Int. Conference on Digital Audio Effects (DAFx-10)*, 2010.

[13] Anssi Klapuri, "Multiple Fundamental Frequency Estimation by Summing Harmonic Amplitudes.," in *ISMIR*, 2006, pp. 216–221.

[14] Robert McAulay and Thomas Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[15] Christian Schörkhuber, Anssi Klapuri, Nicki Holighaus, and Monika Dörfler, "A matlab toolbox for efficient perfect reconstruction time-frequency transforms with log-frequency resolution," in *Audio Engineering Society Conference: 53rd International Conference: Semantic Audio*. Audio Engineering Society, 2014.