A CHROMA-BASED SALIENCE FUNCTION FOR MELODY AND BASS LINE ESTIMATION FROM MUSIC AUDIO SIGNALS

Justin Salamon
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
justin.salamon@upf.edu

Emilia Gómez
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
emilia.gomez@upf.edu

ABSTRACT

In this paper we present a salience function for melody and bass line estimation based on chroma features. The salience function is constructed by adapting the Harmonic Pitch Class Profile (HPCP) and used to extract a mid-level representation of melodies and bass lines which uses pitch classes rather than absolute frequencies. We show that our salience function has comparable performance to alternative state of the art approaches, suggesting it could be successfully used as a first stage in a complete melody and bass line estimation system.

1 INTRODUCTION

With the prevalence of digital media, we have seen substantial growth in the distribution and consumption of digital audio. With musical collections reaching vast numbers of songs, we now require novel ways of describing, indexing, searching and interacting with music.

In an attempt to address this issue, we focus on two important musical facets, the melody and bass line. The melody is often recognised as the *'essence'* of a musical piece [11], whilst the bass line is closely related to a piece's tonality [8]. Melody and bass line estimation has many potential applications, an example being the creation of large databases for music search engines based on Query by Humming (QBH) or by Example (QBE) [2].

In addition to retrieval, melody and bass line estimation could facilitate tasks such as cover song identification and comparative musicological analysis of common melodic and harmonic patterns. An extracted melodic line could also be used as a reduced representation (thumbnail) of a song in music applications, or on limited devices such as mobile phones. What is more, a melody and bass line extraction system could be used as a core component in other music computation tasks such as score following, computer participation in live human performances and music transcription systems. Finally, the determination of the melody and bass line of a song could be used as an intermediate step towards the determination of semantic labels from musical audio, thus helping to bridge the *semantic gap* [14].

Much effort has been devoted to the extraction of a score representation from polyphonic music [13], a difficult task even for pieces containing a single polyphonic instrument such as piano or guitar. In [8], Goto argues that musical transcription (i.e. producing a musical score or piano roll like representation) is not necessarily the ideal representation of music for every task, since interpreting it requires musical training and expertise, and what is more, it does not capture non-symbolic properties such as the expressive performance of music (e.g. vibrato and ornamentation). Instead, he proposes to represent the melody and bass line as time dependent sequences of fundamental frequency values, which has become the standard representation in melody estimation systems [11].

In this paper we propose an alternative mid-level representation which is extracted using a salience function based on chroma features. Salience functions provide an estimation of the predominance of different fundamental frequencies (or in our case, pitch classes) in the audio signal at every time frame, and are commonly used as a first step in melody extraction systems [11]. Our salience function makes use of chroma features, which are computed from the audio signal and represent the relative intensity of the twelve semitones of an equal-tempered chromatic scale. As such, all frequency values are mapped onto a single octave. Different approaches to chroma feature extraction have been proposed (reviewed in [5]) and they have been successfully used for different tasks such as chord recognition [4], key estimation [6] and similarity [15].

Melody and bass line extraction from polyphonic music using chroma features has several potential advantages – due to the specific chroma features from which we derive our salience function, the approach is robust against tuning, timbre and dynamics. It is efficient to compute and produces a final representation which is concise yet maintains its applicability in music similarity computations (in which an octave agnostic representation if often sought after, such as [10]). In the following sections we present the

SMC 2009, July 23-25, Porto, Portugal

proposed approach, followed by a description of the evaluation methodology, data sets used for evaluation and the obtained results. The paper concludes with a review of the proposed approach and consideration of future work.

2 PROPOSED METHOD

2.1 Chroma Feature Computation

The salience function presented in this paper is based on the *Harmonic Pitch Class Profile* (HPCP) proposed in [5]. The HPCP is defined as:

$$HPCP(n) = \sum_{i=1}^{nPeaks} w(n, f_i) \cdot a_i^2 \qquad n = 1 \dots size \quad (1)$$

where a_i and f_i are the linear magnitude and frequency of peak i, $nPeaks$ is the number of spectral peaks under consideration, n is the HPCP bin, $size$ is the size of the HPCP vector (the number of HPCP bins) and $w(n, f_i)$ is the weight of frequency f_i for bin n. Three further pre/post-processing steps are added to the computation. As a preprocessing step, the tuning frequency is estimated by analyzing frequency deviations of peaks with respect to an equal-tempered scale. As another preprocessing step, spectral whitening is applied to make the description robust to timbre. Finally, a post-processing step is applied in which the HPCP is normalised by its maximum value, making it robust to dynamics. Further details are given in [5].

In the following sections we detail how the HPCP computation is configured for the purpose of melody and bass line estimation. This configuration allows us to consider the HPCP as a salience function, indicating salient pitch classes at every time frame to be considered as candidates for the pitch class of the melody or bass line.

2.2 Frequency Range

Following the rational in [8], we assume that the bass line is more predominant in the low frequency range, whilst the melody is more predominant in the mid to high frequency range. Thus, we limit the frequency band considered for the HPCP computation, adopting the ranges proposed in [8]: 32.7Hz (1200 cent) to 261.6Hz (4800 cent) for bass line, and 261.6Hz (4800 cent) to 5KHz (9907.6 cent) for melody. The effect of limiting the frequency range is shown in Figure 1. The top pane shows a chromagram (HPCP over time) for the entire frequency range, whilst the middle and bottom panes consider the melody and bass ranges respectively. In the latter two panes the correct melody and bass line (taken from a MIDI annotation) are plotted on top of the chromagram as white boxes with diagonal lines.

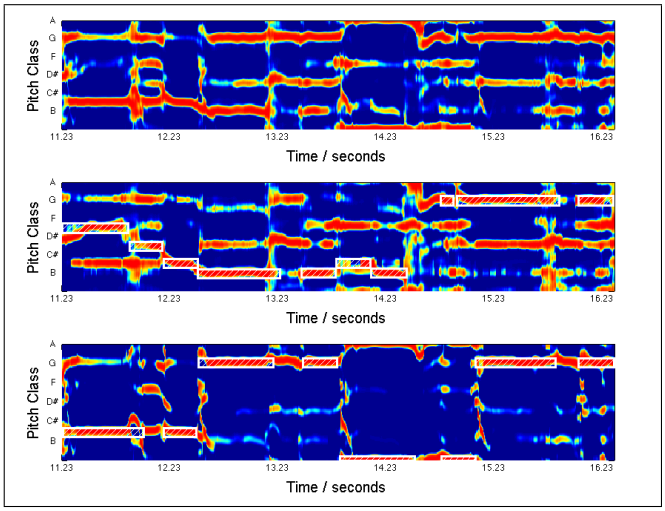

Figure 1. Original (top), melody (middle) and bass line (bottom) chromagrams

2.3 HPCP Resolution and Window Size

Whilst a 12 or 36 bin resolution may suffice for tasks such as key or chord estimation, if we want to properly capture subtleties such as vibrato and glissando, as well as the fine tuning of the singer or instrument, a higher resolution is needed. In Figure 2 we provide an example of the HPCP for the same 5 second segment of *train05.wav* from the MIREX 2005 collection, taken at a resolution of 12, 36, and 120 bins. We see that as we increase the resolution, elements such as glissando (seconds 1-2) and vibrato (seconds 2-3) become better defined. For the rest of the paper we use a resolution of 120 bins.

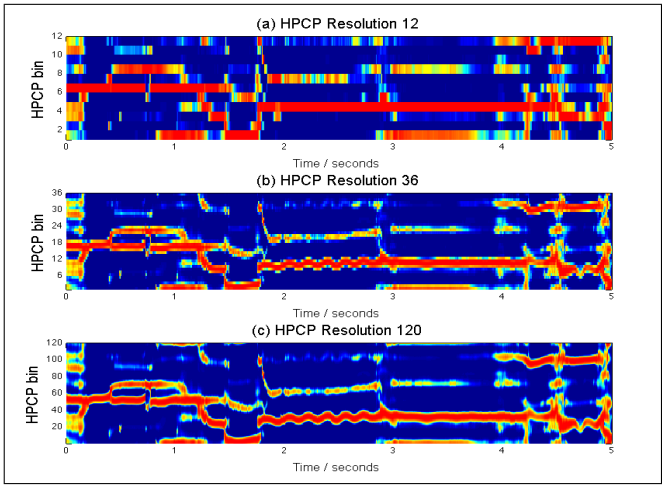

Figure 2. HPCP computed with increasing resolution

Another relevant parameter is the window size used for the analysis. A smaller window will give better time resolution hence capturing time-dependent subtleties of the melody,

whilst a bigger window size gives better frequency resolution and is more robust to "noise" in the analysis (single frames in which the melody is temporarily not the most salient). We empirically set the window size to 186ms (due to the improved frequency resolution given by long windows, their use is common in melody extraction [11]).

2.4 Melody and Bass Line Selection

Given our salience function, the melody (or bass line depending on the frequency range we are considering) is selected as the highest peak of the function at every given time frame. The result is a sequence of pitch classes (using a resolution of 120 HPCP bins, i.e. 10 cents per pitch class) over time. It is important to note that no further post processing is performed. In [11] a review of systems participating in the MIREX 2005 melody extraction task is given, in which a common extraction architecture was identified. From this architecture, we identify two important steps that would have to be added to our approach to give a complete system: firstly, a postprocessing step for selecting the melody line out of the potential candidates (peaks of the salience function). Different approaches exist for this step, such as streaming rules [3], heuristics for identifying melody characteristics [1], Hidden Markov Models [12] and tracking agents [8]. Then, voicing detection should be applied to determine when the melody is present.

3 EVALUATION METHODOLOGY

3.1 Ground Truth Preparation

For evaluating melody and bass line estimation, we use three music collections, as detailed below.

3.1.1 MIREX 2004 and 2005 Collections

These collections were created by the MIREX competition organisers for the specific purpose of melody estimation evaluation [11]. They are comprised of recording-transcription pairs, where the transcription takes the form of timestamp-F0 tuples, using 0Hz to indicate unvoiced frames. 20 pairs were created for the 2004 evaluation, and another 25 for the 2005 evaluation of which 13 are publicly available [1]. Tables 1 and 2 (taken from [11]) provide a summary of the collection used in each competition.

3.1.2 RWC

In an attempt to address the lack of standard evaluation material, Goto et al. prepared the *Real World Computing* (RWC) Music Database [7]. It contains several databases of different genres, and in our evaluation we use the Popular Music Database. The database consists of 100 songs performed in the style of modern Japanese (80%) and American (20%) popular music typical of songs on the hit charts in the 1980s and 1990s.

At the time of performing the evaluation the annotations were in the form of MIDI files which were manually created and not synchronised with the audio [2]. To synchronise the annotations, we synthesised the MIDI files and used a local alignment algorithm for HPCPs as explained in [15] to align them against the audio files. All in all we were able to synchronise 73 files for evaluating melody estimation, of which 7 did not have a proper bass line leaving 66 for evaluating bass line estimation (both collections are subsets of the collections used for evaluating melody and bass line transcription in [13] [3]).

Category	Style	Melody Instrument
Daisy	Pop	Synthesised voice
Jazz	Jazz	Saxophone
MIDI	Folk, Pop	MIDI instruments
Opera	Classical Opera	Male voice, Female voice
Pop	Pop	Male Voice

Table 1. Summary of data used in the 2004 melody extraction evaluation

Melody Instrument	Style
Human voice	R&B, Rock, Dance/Pop, Jazz
Saxophone	Jazz
Guitar	Rock guitar solo
Synthesised Piano	Classical

Table 2. Summary of data used in the 2005 melody extraction evaluation

3.2 Metrics

Our evaluation metric is based on the one first defined for the MIREX 2005 evaluations. For a given frame n, the estimate is considered correct if it is within $\pm\frac{1}{4}$ tone (± 50 cents) of the reference. In this way algorithms are not penalised for small variations in the reference frequency. This also makes sense when using the RWC for evaluation, as the use of MIDI annotations means the reference frequency is discretised to the nearest semitone. The concordance error for frame n is thus given by:

$$err_n = \begin{cases} 100 & \text{if } |f_{cent}^{est}[n] - f_{cent}^{ref}[n]| > 50 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

[1] http://labrosa.ee.columbia.edu/projects/melody/

[2] A new set of annotations has since been released with audio synchronised MIDI annotations.

[3] With the exception of *RM-P034.wav* which is included in our evaluation but not in [13].

The overall transcription concordance (the score) for a segment of N frames is given by the average concordance over all frames:

$$score = 100 - \frac{1}{N}\sum_{n=1}^{N} err_n \quad (3)$$

As we are using chroma features (HPCP) to describe melody and bass lines, the reference is mapped onto one octave before the comparison (this mapping is also used in the MIREX competitions to evaluate the performance of algorithms ignoring octave errors which are common in melody estimation):

$$fchroma_{cent} = 100 + mod(f_{cent}, 1200) \quad (4)$$

Finally it should be noted that as voicing detection is not currently part of our system, performance is evaluated for voiced frames only.

4 RESULTS

In this section we present our melody and bass line estimation results, evaluated on the three aforementioned music collections. For comparison we have also implemented three salience functions for multiple-F0 estimation proposed by Klapuri in [9] which are based on the summation of harmonic amplitudes (henceforth referred to as the Direct, Iterative and Joint methods). The Direct method estimates the salience $s(\tau)$ of a given candidate period τ as follows:

$$s(\tau) = \sum_{m=1}^{M} g(\tau, m)|Y(f_{\tau,m})| \quad (5)$$

where $Y(f)$ is the STFT of the whitened time-domain signal, $f_{\tau,m} = m \cdot f_s/\tau$ is the frequency of the m^{th} harmonic partial of a F0 candidate f_s/τ, M is the total number of harmonics considered and the function $g(\tau, m)$ defines the weight of partial m of period τ in the summation. The Iterative method is a modification of the Direct method which performs iterative estimation and cancellation of the spectrum of the highest peak before selecting the next peak in the salience function. Finally the Joint method is a further modification of the Direct method which attempts to model the Iterative method of estimation and cancellation but where the order in which the peaks are selected does not affect the results. Further details are given in [9]. The three methods were implemented from the ground up in Matlab, using the parameters specified in the original paper, a window size of 2048 samples (46ms) and candidate periods in the range of 110Hz-1KHz (the hop size was determined by the one used to create the annotations, i.e. 5.8ms for the MIREX 2004 collection and 10ms for the MIREX 2005 and RWC collections).

4.1 Estimation Results

The results for melody estimation are presented in Table 3.

Collection	HPCP	Direct	Iterative	Joint
MIREX04	71.23%	75.04%	74.76%	74.87%
MIREX05	61.12%	66.64%	66.76%	66.59%
RWC Pop	56.47%	52.66%	52.65%	52.41%

Table 3. Salience function performance

We note that the performance of all algorithms decreases as the collection used becomes more complex and resemblant of real world music collections. A possible explanation for the significantly decreased performance of all approaches for the RWC collection could be that as it was not designed specifically for melody estimation, it contains more songs in which there are several lines competing for salience in the melody range, resulting in more errors when we only consider the maximum of the salience function at each frame. We also observe that for the MIREX collections the HPCP based approach is outperformed by the other algorithms, however for the RWC collection it performs slightly better than the multiple-F0 algorithms.

A two-way analysis of variance (ANOVA) comparing our HPCP based approach with the Direct method is given in table 4.

Source	SS	df	Mean Squares	F-ratio	p-value
Collection	11,971.664	2	5,985.832	41.423	0.000
Algorithm	75.996	1	75.996	0.526	0.469
Collection* Algorithm	705.932	2	352.966	2.443	0.089
Error	29,768.390	206	144.507		

Table 4. ANOVA comparing the HPCP based approach to the Direct method over all collections

The ANOVA reveals that the collection used for evaluation indeed has a significant influence on the results (p-value $< 10^{-3}$). Interestingly, when considering performance over all collections, there is no significant difference between the two approaches (p-value 0.469), indicating that overall our approach has comparable performance to that of the other salience functions and hence potential as a first step in a complete melody estimation system [4].

We next turn to the bass line estimation results. Given that the multiple-F0 salience functions proposed in [9] are not specifically tuned for bass line estimation, only the HPCP based approach was evaluated. We evaluated using the RWC

[4] When comparing the results for each collection separately, only the difference in performance for the RWC collection was found to be statistically significant (p-value 0.016).

collection only as the MIREX collections do not contain bass line annotations, and achieved a score of 73%.

We note that the performance for bass line is significantly higher. We can attribute this to the fact that the bass line is usually the most predominant line in the low frequency range and does not have to compete with other instruments for salience as is the case for the melody.

In Figure 3 we present examples in which the melody and bass line are successfully estimated. The ground truth is represented by o's, and the estimated line by x's. The scores for the estimations presented in Figure 3 are 85%, 80%, 78% and 95% for *daisy1.wav* (MIREX04), *train05.wav* (MIREX05), *RM-P014.wav* (RWC, melody) and *RM-P069.wav* (RWC, bass) respectively.

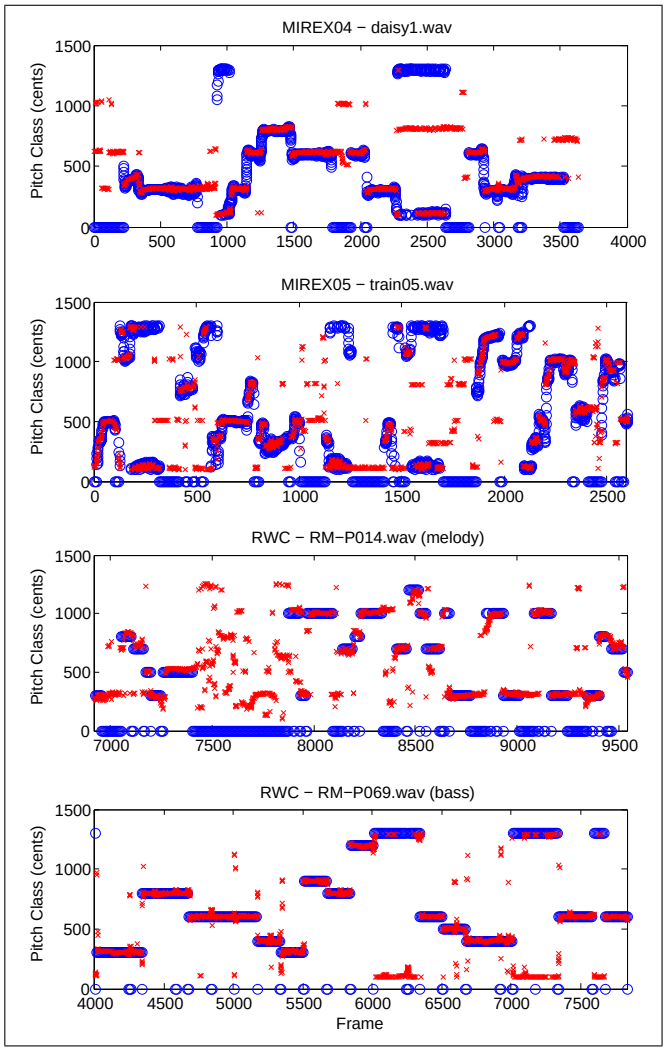

Figure 3. Extracted melody or bass line (x's) against its reference (o's) for each of the collections

In order to evaluate what are the best possible results our approach could potentially achieve, we have calculated estimation performance considering an increasing number of peaks of the salience function and taking the error of the closest peak to the reference frequency (mapped onto one octave) at every frame. This tells us what performance could be achieved if we had a peak selection process which always selected the correct peak as long as it was one of the top n peaks of the salience function. The results are presented in Figure 4.

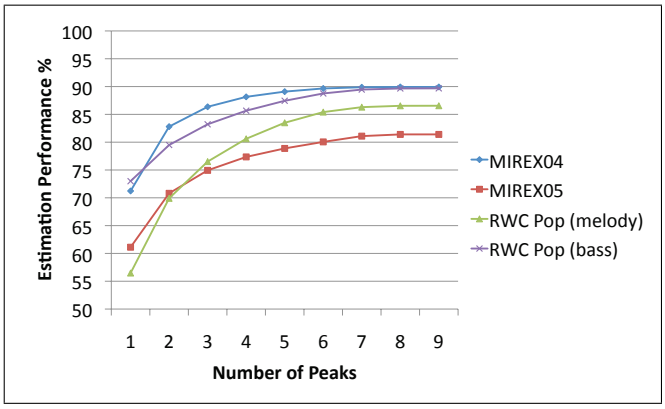

Figure 4. Potential performance vs peak number

The results reveal that our approach has a "glass ceiling" – an inherent limitation which means that there are certain frames in which the melody (or bass line) is not present in any of the peaks of the salience function. The glass ceiling could potentially be "pushed up" by further tuning the pre-processing in the HPCP computation, though we have not explored this in our work.

Nonetheless, we see that performance could be significantly improved if we implemented a good peak selection algorithm even considering just the top two peaks of the salience function. By considering more peaks performance could be improved still, however the task of melody peak tracking is non trivial and we cannot assert how easy it would to get close to these theoretical performance values.

5 CONCLUSION

In this paper we introduced a method for melody and bass line estimation using chroma features. We adapt the Harmonic Pitch Class Profile and use it as a salience function, which would be used as the first stage in a complete melody and bass line estimation system. We showed that as a salience function our approach has comparable performance to that of other state of the art methods, evaluated on real world music collections. Future work will involve the implementation of the further steps required for a complete melody and bass line estimation system, and an evaluation of the extracted representation in the context of similarity based applications.

6 ACKNOWLEDGEMENTS

We would like to thank Anssi Klapuri and Matti Ryynänen for sharing information about the test collections used for the evaluation and for their support; and Joan Serrà for his support and assistance with the HPCP alignment procedure.

7 REFERENCES

[1] P. Cancela. "Tracking Melody in Polyphonic Audio", In *Proc. MIREX*, 2008.

[2] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis. "A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed", *Journal of the American Society for Information Science and Technology*, February 2007.

[3] K. Dressler. "Extraction of the melody pitch contour from polyphonic audio", *Proc. 6th International Conference on Music Information Retrieval*, Sept. 2005.

[4] T. Fujishima. "Realtime Chord Recognition of Musical Sound: a System using Common Lisp Music", *Computer Music Conference (ICMC)*, pages 464–467, 1999.

[5] E. Gómez. *Tonal Description of Music Audio Signals*. PhD thesis, Universitat Pompeu Fabra, Barcelona, 2006.

[6] E. Gómez. "Tonal Description of Polyphonic Audio for Music Content Processing", *INFORMS Journal on Computing, Special Cluster on Computation in Music*, 18(3), 2006.

[7] M. Goto, H. Hashiguchi, T. Nishinura, and R. Oka. "Rwc music database: Popular, classical, and jazz music databases", *Proc. Third International Conference on Music Information Retrieval ISMIR-02*, Paris, 2002. IRCAM.

[8] M. Goto. "A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals", *Speech Communication*, 43:311–329, 2004.

[9] A. Klapuri. "Multiple fundamental frequency estimation by summing harmonic amplitudes", *Proc. 7th International Conference on Music Information Retrieval*, Victoria, Canada, October 2006.

[10] M. Marolt. "A mid-level representation for melody-based retrieval in audio collections", *Multimedia, IEEE Transactions on*, 10(8):1617–1625, Dec. 2008.

[11] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Steich, and O. Beesuan. "Melody transcription from music audio: Approaches and evaluation", *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1247–1256, 2007.

[12] M. Ryynänen and A. Klapuri. "Transcription of the singing melody in polyphonic music", *Proc. 7th International Conference on Music Information Retrieval*, Victoria, Canada, Oct. 2006.

[13] M. Ryynänen and A. Klapuri. "Automatic transcription of melody, bass line, and chords in polyphonic music", *Computer Music Journal*, 32(3):72–86, 2008.

[14] X. Serra, R. Bresin, and A. Camurri. "Sound and Music Computing: Challenges and Strategies", *Journal of New Music Research*, 36(3):185–190, 2007.

[15] J. Serrà, E. Gómez, P. Herrera, and X. Serra. "Chroma binary similarity and local alignment applied to cover song identification", *IEEE Transactions on Audio, Speech and Language Processing*, 16:1138–1151, August 2008.