ENHANCING DOWNBEAT DETECTION WHEN FACING DIFFERENT MUSIC STYLES

Simon Durand, Bertrand David, Gaël Richard

Institut Mines-Telecom, Telecom ParisTech, CNRS-LTCI, 37/39, rue Dareau, 75014 PARIS – France

ABSTRACT

This paper focuses on the automatic rhythm analysis of musical audio at the bar level. We propose a novel approach for robust downbeat detection. It uses well-chosen complementary features, inspired by musical considerations. In particular, a note accentuation model and a detection of pattern changes are introduced. We estimate the time signature by examining the similarity of frames at the beat level. The features are selected through a linear SVM model or a weighted sum. The whole system is evaluated on five different datasets of various musical styles and shows improvement over the state of the art.

Index Terms— Downbeat-tracking, Music Information Retrieval, Music Signal Processing

1. INTRODUCTION

Cooper sees the pulse as one of a series of regularly recurring equivalent stimuli [1]. Pulses with different accentuation, can be grouped into regular patterns to form bars. The downbeat, simply defined as the first beat of each bar, is a high-level piece of musical information, the detection of which implies the human perception ability to organize musical content. Its automatic detection would be a useful tool for many applications. For example, automatic music-to-sheet transcription [2], track-to-track alignment for studio post-production, music indexing [3] and media to music playback synchronisation [4].

Specific effort has been dedicated to design efficient methods that include downbeat detection in a rhythm analysis framework. For instance, in [4,5] a deterministic hierarchical framework is used. Others rely on classification methods, as support vector machine or k-nearest-neighbour [6, 7]. Lastly, probabilistic frameworks (Hidden Markov Models or bayesian graphical models) are developed for a better adaptation to data [8–11]. The aforementioned methods mainly rely on rhythmic pattern recognition and/or chord transition detection. Using both cues often leads to better results. It can likely be linked to our multi-level perception of downbeats, taking into account time indicators but also harmony and melodic lines [12].

An important issue in the field is the absence of large enough indexed datasets [13] when compared to language processing for example. This sometimes leads the algorithms to learn and to be tested on sets where the genre is not diverse enough. This work also faces these issues but tends to focus on challenging excerpts from different music genres. To enhance the robustness of the detection in that context, we propose a novel approach, taking into account rhythm perception [1] [12], that uses well chosen complementary features, inspired by musical considerations and classifies them to extract the downbeats of musical excerpts. This allows the model to adapt to a wide variety of musical pieces, ranging from Classical to Pop/Rock music, with or without drums.

This paper is organized as follows. We first present the three main parts of our system in section 2. It includes feature extraction, time signature estimation and features fusion/selection. In section 3, the features related to harmony, percussive events, melody and musical structure that are used in our work are described. The way downbeats are detected is presented in section 4. The proposed algorithm is finally evaluated on five datasets and compared to two state of the art systems in section 5 before some concluding remarks.

2. MODEL OVERVIEW

Our system, illustrated in figure 1, is decomposed in three parts. We first extract the musically inspired features from the audio signal and the beat positions. We then estimate the time signature: the number N_b of beats per bar. We assume it is constant along the track. We have therefore N_b different ways of placing the downbeats, i.e N_b downbeat sequences ds. For example, the i^{th} downbeat sequence ds_i contains all the beats that follow two hypotheses. The first downbeat is the i^{th} annotated beat and then every $N_b{}^{th}$ beat is a downbeat. Each feature gives a weight to all possible downbeat sequences according to their likeliness. Finally, a fusion or selection strategy is used to obtain one downbeat sequence from the features.

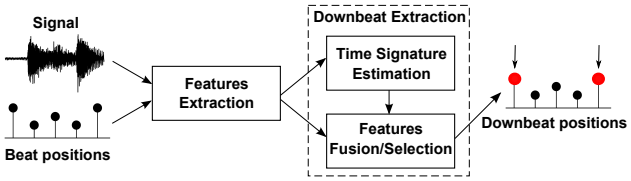

Fig. 1. Model overview.

Our chord changes, harmonic balance, melodic accents and pattern changes features are linked to harmony, rhythm patterns, melody and musical structure. The downbeat is inherently a high-level musical concept and it is then important to rely on high-level semantic features to estimate its position in the audio signal. The first assumption is that chords are more likely to change near a downbeat than near another beat. Several authors have used this kind of feature to estimate downbeats [14, 15]. The second assumption is that drummers usually use a bass drum on the first and third beats and snare drum on the second and fourth beats of the bar. As suggested in [4, 8], the oscillation between high frequency and low frequency content in the signal near beat positions can thus be used as a feature to distinguish beat positions within the bar. The third assumption is that some notes are more accented than others in a song and that accentuation is more likely to happen near a downbeat. To our knowledge, this property has never been used for downbeat detection from audio signals. In this work, we take advantage of the results of perceptive studies carried out by [16–19] among others in order to define and model note accent. In that regard, we follow Ellis statement that using both melodic and rhythmic cues play an important role in the interpretation of meter and we use these two

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

cues [19]. Beats near accented notes are weighted accordingly and act as a feature for downbeat detection. The last assumption considering feature extraction is that a pattern change is more likely to happen near a downbeat than near another beat of the bar. A pattern change is a significant change in the musical content. We present two methods that have not been used previously for downbeat detection. The low-level feature used to model pattern changes is the similarity matrix, introduced by [20].

The time signature is regarded in this article as the number N_b of beats per bar and its estimation is a problem in itself. We choose to only consider 3, 4, 5 or 7 beats per bar as these are some of the main numbers of beats per bar and these are easily distinguishable from each other as opposed to 2 and 4 beats per bar for example [21, 22]. We aim for a fixed number of beats per bar on a given portion of the song which is a valid assumption in most cases. Missing or added beats in the annotation are dealt with by observing the local tempo and avoiding local double or halve tempi. Time signature estimation can be summarized by seeking the best beat periodicity in the signal. We thus examine the periodicity of a similarity matrix at the beat level.

The goal of the final step is to obtain the most probable downbeat sequence from the $N_f N_b$ hypotheses, where N_f is the number of features and N_b the number of beats per bar. This is performed following either a fusion strategy (using a weighted sum of features sequences) or a selection strategy (select the best hypothesis using a SVM classifier).

3. MUSICAL FEATURES EXTRACTION

Further details on the musical feature extraction are provided in this section.

3.1. Chord changes

The method used here for chord estimation is largely inspired from the work of [23]. This approach combines chroma feature extraction, median filtering and pattern matching with a set of chord templates. We only considered major and minor chords, as well as chords lasting more than twice the tempo to improve the precision. For a given excerpt, the chord changes feature has N_b dimensions where N_b is the number of beats per bar, equal to the number of possible downbeat sequences ds. Each dimension of this feature is equal to the number of chord changes instants that fit the considered downbeat sequence.

3.2. Harmonic balance

The square of the absolute value of the short term Fourier Transform $F(w,t)$ is first computed for each frequency bin w and temporal bin t. We consider specific bandwidths for the snare drum $B_{snare} = [1400-7500Hz]$ and the bass drum $B_{bass} = [0-150Hz]$ proposed in [4]. The maximum values e_{snare} and e_{bass} on an interval $2L$ of two tenth of the tempo $T(b_e)$ at beat b_e are computed for the snare and the bass drum on their specific bandwidths.

$$e_{snare}(b_e) = \max_t \left(\sum_{w \in B_{snare}} |F(w,t)|^2 \right), \quad t \in [b_e - L, b_e + L] \tag{1}$$

$$e_{bass}(b_e) = \max_t \left(\sum_{w \in B_{bass}} |F(w,t)|^2 \right), \quad t \in [b_e - L, b_e + L] \tag{2}$$

Each value fitting a downbeat sequence ds is then summed across the entire audio signal and the bass-to-snare ratio HB is computed.

$$HB(ds) = \frac{\sum_{b_e \in ds} e_{bass}(b_e)}{\sum_{b_e \in ds} e_{snare}(b_e)} \tag{3}$$

Each of these choices was made after a thorough optimization process. For example, it is important to sum the snare and the bass drum values across the entire signal otherwise a lack of bass at one beat, that is likely to happen at least once in a song, negatively impacts the semantic meaning of this ratio.

3.3. Note accent

Before we can compute the note accent feature, we have to estimate, at least roughly, the melody of the song. In this work, we rely on a rather straightforward approach which exploits multiple fundamental frequency estimation based on harmonics and spectral smoothness (inspired by [24]), followed by a dynamic programming module for the melody estimation. As a result, the three most probable melodies with no overlapping notes are kept. Since the previous method is likely to estimate too many notes, two consecutive notes are tied if they have the same pitch and a note is kept if its duration is longer than a quarter of the tempo to improve the precision.

Although rhythmic and melodic information is useful for notes accent, it is not clear whether these cues combine additively or interactively [19]. We calculate the note accent as the product of the rhythmic and the melodic cues. The rhythmic cue is equal to the normalized square of the note duration and the melodic cue is computed according to weightings found in [18]. These models emphasize long notes between descending and ascending melodies. The closest beat to a note onset is weighted by the corresponding note accent and each dimension of the note accent feature is equal to the sum of the weighted beats that fit the considered downbeat sequence.

3.4. Pattern changes

To estimate pattern changes, we first build a similarity matrix where each entry $S(i,j)$ is obtained by the euclidean distance between the twelve Mel Frequency Cepstral Coefficients of the signal frames i and j. Similar patterns in the signal are thus represented by homogeneous bloc matrices. Pattern changes can be regarded in an absolute or relative way. In the absolute case, the exact time instant where pattern changes occur acts as a downbeat feature and in the relative case, the consecutive beats that fit pattern changes best is sought after.

3.4.1. Absolute pattern changes

Absolute time instant of a pattern change (termed herein absolute pattern changes) are obtained as the corners of a homogeneous bloc in the similarity matrix. The homogeneous bloc corners are obtained whenever the similarity measure s_m exceeds a predefined threshold. The similarity measure is the sum of two individual measures s_1 and s_2 described below.

The measure s_1 is a comparison between the similarity of the first N frames and the similarity of the first $N+1$ frames. It emphasizes global changes in the musical content. The measure s_2 is a comparison between the similarity of the first N frames and the similarity of the $(N+1)^{th}$ frame with the first N frames. It emphasizes sudden changes in the musical content. An illustration of these measures for $N = 2$ is made in the figure 2.

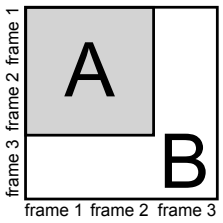

Fig. 2. Bloc matrix from the similarity matrix. s_1 is a comparison between A and A+B and s_2 is a comparison between A and B.

Each comparison is made using the Mahalanobis distance between the two parts of the similarity matrix to compare. Time instant of the maximum of these measures is stored when the threshold is exceeded. After initializing the time to zero, the method proceeds as follows:

- **Rough estimation :** We estimate an absolute time instant of a pattern change with low accuracy, using 0.5 second frames, on the whole remaining signal.
- **Refinement :** We estimate an absolute time instant of a pattern change with high accuracy, using 0.05 second frames, around the low accuracy time instant previously estimated.
- **Update :** We store this high accuracy time instant and we update the initializing time with the obtained value.

We can therefore have a good compromise between computation cost and accuracy. Each dimension of this feature is equal to the number of pattern changes instants that fit the associated downbeat sequence.

3.4.2. Relative pattern changes

The first downbeat sequence, in which the first annotated beat is a downbeat, is selected. We compute the similarity matrix $\underline{\underline{S}}$ of the signal with bar-length frames according to this downbeat sequence. We then compute its "smoothness score" and iterate this process for all potential downbeat sequences. Each dimension of the relative pattern change feature is equal to the smoothness score of the associated downbeat sequence.

The smoothness score is computed as follows:

- We find the pattern change time instant t with the same measures as those used in the absolute method above.
- The local smoothness score is the difference between the similarity matrix coefficients $S(t_{b_a}, t_{b_a-1})$ and $S(t_{b_a-2}, t_{b_a-1})$, with t_{b_a} the frame corresponding to the ${b_a}^{th}$ bar.
- The smoothness score is the sum of the local smoothness scores at all pattern change time instants.

We assume that if downbeats are correctly estimated, bars will be relatively more similar before a pattern change and relatively less similar after a pattern change. This way, $S(t_{b_a}, t_{b_a-1})$ will be high and $S(t_{b_a-2}, t_{b_a-1})$ will be low.

4. DOWNBEAT EXTRACTION

A measure of beat periodicity to estimate the number of beats per bar and two models to infer downbeat positions from the features and the time signature are presented in this section.

4.1. Time signature estimation

The measure $P(N_b)$ of the number N_b of beats per bar takes advantage of the nature of the similarity matrix. Beat similarity matrix diagonals represent the similarity of equally spaced beats. A periodicity measure of these diagonals can therefore give us the best beat periodicity and the time signature [22]. Our beat similarity matrix $\underline{\underline{S}}$ is computed with the same parameters as in 3.4 but with beat-length frames. It allows a fast computation time and a good accuracy. The measure d of the diagonals of $\underline{\underline{S}}$ is done as in [22] but the periodicity measure P takes into account additional boundary constraints:

$$P(N_b, i) = \left(\frac{\sum_{n=1}^{i} \text{Ш}_{N_b}(n) d(n)}{\sum_{n=1}^{i} \text{Ш}_{N_b}(n)} \right) \tag{4}$$

where Ш_{N_b} is a Dirac comb of periodicity N_b and $d(n)$ is the average of the n^{th} diagonal of $\underline{\underline{S}}$:

$$d(n) = \text{mean}(\text{diag}_n(\underline{\underline{S}})) \tag{5}$$

$$d(n) \leftarrow -d(n) + \max(d) \tag{6}$$

The opposite of d is calculated because the euclidean distance of similar beats is low while we want the highest d in that case. We don't consider the diagonals far from the main diagonal of $\underline{\underline{S}}$ which represent the similarity of beats far from each others. From a computational point of view, it leads to few values to average (and thus a weak robustness) and from a musical point of view, those beats are less related and able to predict a periodicity in the signal. On the other hand, we need as much beat similarity information as we can to infer the periodicity. The optimal number of diagonals needed differs from the excerpts and we thus define a boundary interval $I = [0.5B, 0.8B]$, considering the number B of beats in the excerpt, on which we average the periodicity measure P:

$$\overline{P(N_b)} = \text{mean}_{i \in I}(P(N_b, i)) \tag{7}$$

The time signature TS is the argument of the maximum of $\overline{P(N_b)}$.

$$TS = \arg\max_{N_b} \left(\overline{P(N_b)} \right) \tag{8}$$

This system has a 95.94 % time signature detection accuracy rate given the assumptions (2 and 4 beats per bar are considered equivalent, 3 and 6 beats per bar are considered equivalent as in [21] and only the prominent time signature of the excerpt is considered).

4.2. Features fusion or selection

We express the task of the downbeat detection given the musical features and the time signature as a classification problem. Given our assumption of constant number N_{bj} of beats per bar in an excerpt j, there are N_{bj} downbeat sequences possible. We present here two methods to determine the best downbeat sequence among N_{bj} given the features.

The first method is a weighted sum of the features. Each feature is first normalized so the sum of all its coefficient on the entire dataset is 1.

$$f_k(i,j) = \frac{f_k(i,j)}{\sum_l \sum_m f_k(l,m)} \tag{9}$$

where i is the downbeat sequence number, j the excerpt number in the dataset and $k = \{CC, HB, NA, APC, RPC\}$ the type of feature with CC: chord change, HB: harmonic balance, NA: note accent, APC: absolute pattern changes and RPC: relative pattern

changes. A sum of the features is computed because the higher the coefficients of the features are the more accented a downbeat sequence is. This fits our assumption that a downbeat is usually more accented than another beat in the bar. The normalization on all the dataset takes into account that a feature has little decision-making power if it has lower values than usual, exhibiting a lack of robustness, given an excerpt. The most likely downbeat sequence ds_j is computed according to the following weighting sum:

$$ds_j = \arg\max_i \left(\sum_k c_k f_k(i,j) \right) \quad (10)$$

where c_k is the weight of the feature f_k. The numerical values ($c_{CC} = 1$, $c_{HB} = 0.85$, $c_{NA} = 0.65$, $c_{APC} = 0.46$, $c_{RPC} = 0.53$) can give an idea of the significance of each feature. One can note that no feature is insignificant.

The second method is a C-SVC Support Vector Machine with a radial basis Kernel type classification computed through the LibSVM toolbox [25]. We use the five aforementioned features for a total of $5N_{bj}$ hypotheses, where N_{bj} is the number of beats per bar for the excerpt j. The support vector machine algorithm selects the best hypothesis, the best regular downbeat sequence ds_j considering the time signature N_{bj}, for each excerpt. A 10-fold cross-validation is used. The ground truth and features were randomly permuted in the learning phase so the class repartition was uniformly distributed and no bias was induced (the first annotated beat is usually a downbeat).

5. EVALUATION AND RESULTS

Scoring methods used for testing our algorithm are taken from the Davies' evaluation toolbox [26]. The F-measure is calculated with a precision window whose length is equal to 10% of the minimum distance between two successive beats in the track. We also use the Correct Metrical Level with continuity required (cmlC) with a similar precision window. While the F-measure averages accuracy and recall rates, the ability of the method to successively detect downbeats is measured with the cmlC.

The parameters were learned on the "RWC Music Database: Music Genre" dataset, thus benefiting from a wide variety of music styles. Annotations of the RWC datasets are provided by the AIST [27].

Evaluations were carried on five different datasets. The first one is a subset of a database made and used in [8], containing 40 randomly chosen tracks among four music styles (Blues, Classical, Jazz and Electro/Dance) given two constraints : the time signature is fixed for each excerpt and there are 10 tracks per genre. This is the only dataset with constant meter. The second dataset is composed of 72 modern Popular songs and was used for the evaluation inside the European Quaero project [28][1]. The 3^{rd}, 4^{th} and 5^{th} ones are extracted from the "RWC Music Database: Popular, Classical and Jazz Music" [29]. They respectively include 100, 60 and 50 tracks from Popular, Classical and Jazz music styles. These datasets have already been used by researchers for evaluation purposes on downbeat detection [8, 11, 15].

Our system is more similar to that of [9] since it also takes beat positions and audio signal as input for downbeat detection. The comparison with the Sonic Annotator software[2] with the *Bar Tracker* QM Vamp plug-in[3] [5] is more difficult since our version of this algorithm does not use beat positions as input and has to estimate them beforehand. That's why we only consider cases were the F-measure of the beat positions was above 95% for a given excerpt (Subset "Sub"). Since this algorithm also needs time signature inputs, we fed it with our time signature estimations. To compete with the aforementioned systems, both versions of our own are tested: the one denoted by C5 SVM uses the SVM classification method and the C5 Sum corresponds to the weighted sum method.

Algorithm	Kla	Quaero	Jazz	Cla	Pop	**Mean**	Sub
F-measure							
[5]							77.6
[9]	61.6	78.6	65.9	**75.0**	87.6	73.7	81.4
C5 Sum	**71.0**	85.0	**84.0**	72.7	88.4	**80.3**	**84.1**
C5 SVM	66.0	**89.0**	74.7	71.0	**91.3**	78.4	83.9
cmlC							
[5]							88.3
[9]	66.3	80.4	56.2	48.4	90.9	68.5	83.6
C5 Sum	**79.8**	84.8	**74.7**	**56.7**	**92.8**	**77.8**	**88.6**
C5 SVM	74.7	**88.3**	68.7	55.1	91.3	75.6	86.0

Table 1. Downbeat detection results for several datasets. Configurations: C5 Sum: All features and a weighted sum classification method, C5 SVM: All features and a Support Vector Machine classification method. Datasets: Kla: Klapuri dataset subset, Quaero: Quaero dataset, Jazz: RWC Jazz dataset, Cla: RWC Classical dataset, Pop: RWC Popular dataset, Mean: Mean value of the results, Sub: Dataset subset where the beat F-measure of [5] is above 95 %.

Algorithm	C1	C2	C3	C4	C5
F-measure	59.99	61.07	63.49	66.03	71.03

Table 2. Downbeat detection results (F-measure) for several configurations on the Klapuri dataset subset. Configurations : C1: CC, C2: CC+ APC, C3: CC + APC + NA, C4: CC + APC + NA + HB, C5 : All features.

Results are reported in table 1 and show that C5 SVM and C5 Sum both overcome the other systems in general. They are close for Popular music likely because downbeat detection is rather successful in this case for all systems. Conversely, our system proves more robust to different musical styles (*e.g.* Jazz and Blues), except for Classical music in the case of the F-measure where time signature switches often occur during long tracks while in our system a constant one is assumed. Coherently, all methods exhibit good results on the subset "Sub" since it includes mostly tracks from Popular music (76.8%). Another interesting result is shown in table 2 where the performance improves when the number of features increases.

6. CONCLUSION

Evaluation results show that using complementary high level musically inspired features is efficient for downbeat detection when facing different music styles.

It could be interesting to develop a more generic classification method or include our features in an integrated system that can detect the best downbeat sequence possible for songs with multiple time signatures. Our system could also be extended to a broader transcription algorithm. It would jointly use any of the cues from rhythm patterns, melody, harmony and musical structure to obtain the rhythmic and melodic parts of the transcription.

[1]This work was supported in part by the Quaero Program funded by Oseo French State Agency for Innovation.

[2]http://omras2.org/SonicAnnotator

[3]http://www.vamp-plugins.org

7. REFERENCES

[1] G. Cooper, *The rhythmic structure of music*, University of Chicago Press, 1963.

[2] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," 2010, vol. 18, pp. 1280–1289, IEEE.

[3] E. Tsunoo, G. Tzanetakis, N. Ono, and S. Sagayama, "Beyond timbral statistics: Improving music classification using percussive patterns and bass lines," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 1003–1014, 2011.

[4] M. Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[5] M. E. P Davies and M. D. Plumbley, "A spectral difference approach to extracting downbeats in musical audio," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2006.

[6] T. Jehan, "Downbeat prediction by listening and learning," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005, pp. 267–270.

[7] D. Ellis and J. Arroyo, "Eigenrhythms: Drum pattern basis sets for classification and generation," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2004, pp. 101–106.

[8] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 342–355, 2006.

[9] H. Papadopoulos and G. Peeters, "Joint estimation of chords and downbeats from an audio signal," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 1, pp. 138–152, 2011.

[10] N. Whiteley, A. T. Cemgil, and S. J. Godsill, "Bayesian modelling of temporal structure in musical audio," in *Proceedings of International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 29–34.

[11] M. Khadkevich, T. Fillon, G. Richard, and M. Omologo, "A probabilistic approach to simultaneous extraction of beats and downbeats," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 445–448.

[12] F. Lerdahl and R. Jackendoff, *A generative theory of tonal music*, Cambridge, MA: The MIT Press, 1983.

[13] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 9, pp. 2539–2548, 2012.

[14] M. Goto and Y. Muraoka, "Real-time rhythm tracking for drumless audio signals–chord change detection for musical decisions," in *Proceedings of the International Conference in Artificial Intelligence: Workshop on Computational Auditory Scene Analysis*, 1997.

[15] G. Peeters and H. Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: Theory and large-scale evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 6, 2011.

[16] P. Pfordresher, "The role of melodic and rhythmic accents in musical structure," *Music Perception*, vol. 20, no. 4, pp. 431–464, 2003.

[17] E. Hannon, J. Snyder, T. Eerola, and C. Krumhansl, "The role of melodic and temporal cues in perceiving musical meter," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 5, pp. 956, 2004.

[18] J. Thomassen, "Melodic accent: Experiments and a tentative model," *Journal of the Acoustical Society of America*, vol. 71, pp. 1596, 1982.

[19] R. Ellis and M. Jones, "The role of accent salience and joint accent structure in meter perception," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, no. 1, pp. 264, 2009.

[20] J. Foote, "Visualizing music and audio using self-similarity," in *Proceedings of the seventh ACM international conference on Multimedia (Part 1)*. ACM, 1999, pp. 77–80.

[21] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Proceedings of the Convention of the Audio Engineering Society*, 2003.

[22] M. Gainza, "Automatic musical meter detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 329–332.

[23] L. Oudre, Y. Grenier, and C. Févotte, "Chord recognition by fitting rescaled chroma vectors to chord templates," *IEEE Transactions on Audio, Speech and Language Processing,*, vol. 19, no. 7, pp. 2222–2233, 2011.

[24] A. Klapuri, "Multiple fundamental frequency estimation based on harmonicity and spectral smoothness," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 804–816, 2003.

[25] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, pp. 27, 2011.

[26] M. E. P Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," *Queen Mary University, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.

[27] M. Goto, "Aist annotation for the rwc music database," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2006, pp. 359–360.

[28] "Quaero programme website http://www.quaero.org/," .

[29] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "Rwc music database: Popular, classical and jazz music databases.," in *Proceedings of the International Conference on Music Information Retrieval (ISMIR)*, 2002, vol. 2, pp. 287–288.