A PROBABILISTIC APPROACH TO SIMULTANEOUS EXTRACTION OF BEATS AND DOWNBEATS

Maksim Khadkevich†, Thomas Fillon‡, Gaël Richard‡, Maurizio Omologo†

† Center of Information Technology, Fondazione Bruno Kessler - Irst
via Sommarive 18, 38123 Trento, Italy
‡ Institut Telecom, Telecom ParisTech, CNRS-LTCI, 46, rue Barrault, 75634 PARIS Cedex 13 – France

ABSTRACT

This paper focuses on the automatic extraction of beat structure from a musical piece. A novel statistical approach to modeling beat sequences based on the application of Hidden Markov Models (HMM) is introduced. The resulting beat labels are obtained by running the Viterbi decoder and subsequent lattice rescoring. For the observation vectors we propose a new feature set that is based on the impulsive and harmonic components of the reassigned spectrogram. Different components of observation vectors have been investigated for their efficiency. The main advantage of the proposed approach is the absence of imposed deterministic rules. All the parameters are learned from the training data, and the experimental results show the efficiency of the proposed schema.

1. INTRODUCTION

Extracting different types of semantic information from music data has become an emerging area of research in Music Information Retrieval (MIR) community. Beat/downbeat tracking is one of the most challenging tasks in MIR. Beats are commonly defined as the time instants at which human beings would tap their foot for rhythm of the music. From the musicological viewpoint, downbeat position is defined as the first beat in a bar. Classification of rhythmical events into beats and downbeats brings a portion of useful information about metrical structure, that can be used as high-level feature in many MIR tasks.

Numerous approaches exist for beat/downbeat extraction. Most of them are based on searching for periodicities in some kinds of onset detection function (ODF) [1], [2]. The most common periodicity detection methods are based on autocorrelation [3], [4], bank of comb filter resonators [5], or short-time Fourier transform of the ODF [3]. However, estimating beat structure for non-percussive sounds, especially with soft note onsets, becomes a more complex problem due to the noisy ODF. In order to circumvent this, more sophisticated methods that are based on pitch [6] and group delay [7] analysis were proposed.

Recently, several HMM-based approaches have been proposed. Peeters in [8] used a reverse Viterbi algorithm which decodes hidden states over beat-numbers, while beat-templates are used to derive observation probabilities. Y. Shiu et al. [9] used a periodic HMM structure to extract beat locations, based on the tempo information obtained on the previous step.

In this paper[1] we suggest an approach that performs a simultaneous estimation of beats and downbeats. It consists of two hierarchical layers, which include acoustic modeling and word sequence modeling, with a novel schema to represent periodic metrical structure. In section 2, a brief introduction to the feature vectors extraction techniques is provided. Section 3 describes the system architecture. Experimental results are given in section 4, and finally, section 5 suggests the conclusions.

[1]This work was partly realized as part of the Quaero Program, funded by OSEO, French State agency for innovation. The internship of Maksim Khadkevich at Telecom ParisTech was funded by the Province of Trento PAT.

2. FEATURE EXTRACTION

Feature extraction is an essential step towards effective and accurate beat/downbeat positions extraction. In this paper two different types of features are proposed to model onset events and harmonic changes. The first dimension is represented by an Onset Detection Function (ODF), based on the impulsive part of the reassigned spectrogram. The second and third dimensions are introduced to model the dynamics of harmonic changes. In order to model fast and slow changes a Chroma Variation Function (CVF) is calculated for short and large context windows. The choice of CVF as a feature vector component is based on the assumption that most harmonic changes that occur inside a piece of music are located at metric bars.

2.1. Time-frequency reassignment technique for harmonic and impulsive spectral components extraction

The proposed feature set is based on the computation of the time-frequency reassigned spectrogram and harmonic/impulsive component separation proposed in [10]. Time-frequency reassignment (TFR) is a well-known technique in spectral analysis and has been widely used in different tasks such as sinusoidal estimation, cover song identification, chord recognition [11], or beat detection [12]. The main idea behind TFR is to remap the spectral energy of each spectrogram cell into another cell that is the closest to the true region of support of the analyzed signal.

From a signal downsampled to 11025 Hz, the feature extraction scheme adopted here aims at separating the reassigned spectrogram into three components: impulsive, harmonic and noise by following the method proposed in [10]. For each time frame k and each frequency bin n, the impulsive and harmonic part of the reassigned spectrogram ($S_{imp}(k,n)$ and $S_{harm}(k,n)$ respectively) are derived. In the present study, the noise component is discarded. To enhance the time resolution, the spectrogram use to derived S_{imp} was computed with a Hanning window of 92 ms with 90% overlap between subsequent frames. Whereas, to enhance frequency resolution, the spectrogram use to derived S_{harm} was computed with a Hanning window of 184 ms with 95% overlap.

2.2. Onset detection function

The TFR technique has proved to be efficient to derive an onset detection function (see [12], for example). In our approach, the onset detection function is obtained by summing all spectral components of the impulsive reassigned spectrogram in a given frame:

$$ODF(t) = \sum_{k} S_{imp}(k, n) \tag{1}$$

The ODF based on the spectral energy sum of the impulsive part of the reassigned spectrogram acts as the first dimension in the feature vector space.

2.3. Chroma variation function

Discrimination between beats and downbeats is particularly challenging and often needs a richer feature set, rather than a single ODF. Papadopoulos and Peeters [13] exploited interrelationships between chords and downbeats to perform simultaneous estimation of these attributes. Davies [14] used spectral difference between band-limited beat synchronous analysis frames as a robust downbeat indicator. In this work we propose to use a so-called Chroma Variation Function. The main concept here on which we base our ideas is the fact that harmonic (chord) changes frequently occur on the downbeat positions. CVF reflects the discrepancies between mean chroma vectors of two adjacent segments. This technique was used in [15] and [16], where spectral variation function features were used for speech recognition and automatic segmentation purposes. It was shown that using variable context lengths along with mean subtraction leads to more robust features. In this paper we adopt a similar approach.

Let $c(k)$ be a chromagram derived from $S_{harm}(k, n)$ as proposed in [11]. Left $c_{l_L}(k)$ and right $c_{r_L}(k)$ contexts of length L correspond to the bins with indexes $[k - L, ..., k]$ and $[k, ..., k + L]$ respectively. The chromagram $c(k)$ is extracted from the harmonic part of the reassigned spectrogram introduced in [10].

Let ${c_{l_j}}'(k)$ and ${c_{r_j}}'(k)$ be the left and the right contexts with subtracted mean value over time $m(k)$ of the context that corresponds to the bins with indexes $[k - L, ..., k + L]$:

$${c_{l_j}}'(k) = c_{l_j}(k) - m(k) \tag{2}$$

$${c_{r_j}}'(k) = c_{r_j}(k) - m(k) \tag{3}$$

Let $\rho(c_l, c_r)$ be the normalized inner product between the two context means $\bar{c_l}$ and $\bar{c_r}$:

$$\rho(c_l, c_r) = \frac{< \bar{c_l}, \bar{c_r} >}{|\bar{c_l}||\bar{c_r}|} \tag{4}$$

The CVF is then defined as:

$$CVF(k) = \frac{1 - \min(M_{left}, M_{right})}{2} \tag{5}$$

$$M_{left} = \min_{1 \le j \le L} (\rho({c_{l_L}}'(k), {c_{r_j}}'(k))) \tag{6}$$

$$M_{right} = \min_{1 \le j \le L} (\rho({c_{l_j}}'(k), {c_{r_L}}'(k))) \tag{7}$$

The meaning of $CVF(k)$ can be interpreted as a cosine of an angle between the two mean chroma vectors with subtracted $m(k)$ value. In order to identify the highest (i.e., most significant) chroma variations, given the left and the right contexts, minimum values in Equations 6 and 7 are used. Varying context length L allows one to set up the ability to detect smooth or fast harmonic changes.

3. BEAT/DOWNBEAT DETECTION SYSTEM OVERVIEW

Our method for beat/downbeat estimation follows a statistical approach that consists of two hierarchical levels: acoustic modeling and word sequence modeling. These general ideas have already proved to be effective in speech recognition or chord recognition. Statistical approaches have also been used in beat/downbeat estimation but usually do not involve a word sequence modeling step. As opposed to some deterministic approaches, where beat locations are obtained by some kinds of periodicity analysis and subsequent beat detection using the tempo information extracted on the previous step, no prior information is needed here. The proposed approach relies on the concept that the rhythmic events can be described as a dictionary of words. Each word represents a time segment between two adjacent beat events. In other words, we introduce a specific dictionary and unit alphabet, which opens the way to the application of language model approaches to the beat/downbeat estimation problem. Then, similarly to speech recognition, a unit-based transcription of each word from the dictionary is provided. The alphabet includes 5 units (beat pre-attack/attack, downbeat pre-attack/attack, no-beat) and words are then defined by aggregating units. Each word is then characterized by a given duration.

3.1. Acoustic modeling

As opposed to many other tasks, such as speech/chord recognition, the use of HMMs for decoding highly periodic events, e. g. beat/downbeat positions, has some distinctive features. One of the most serious problems one can come across, when trying to use HMM for decoding highly periodical events, is the problem of keeping periodicity in the output labels. Self-transitions in the states of an HMM allow the model to remain in the same state for quite a long period of time. At the same time, time intervals containing numerous note onsets can force the model to produce beat labels with very short durations. Y. Shiu et al. [9] proposed a periodic left-to-right model that produces periodic output. However, a prior information on the tempo is required. The solution proposed here is to discard all self-transitions and to add an additional *word sequence modeling layer* to the system architecture.

The acoustic unit dictionary was build is such a way that different units model the following events: beat pre-attack (BTp), beat attack (BTa), downbeat pre-attack (DBTp), downbeat attack (DBTa) and "no beat" (NB). We draw an analogy between a unit in the beat extraction task and a phoneme in speech recognition. All the units, apart from NB, are represented by a Bakis HMM with a number N_{st} of hidden states and no self-transitions. The NB unit has only one emitting state. The number of states N_{st} imposes a duration constraint and corresponds to the necessary number of time frames to output the unit.

The model parameter estimation is done using training material with ground-truth markers, manually labeled. The extracted feature vectors are segmented according to the ground-truth labels so that each segment contains N_{st} frames corresponding to a specific unit. All emission probabilities are learned from the training data using the Baum-Welch algorithm.

In such a way, different units model different phases of a beat/downbeat event, following at the same time the duration constraint. The proposed training schema rules out the possibility to

stay in any state for more than one frame. Figure 1 depicts an example of the acoustic training when $N_{st} = 4$ and $n(i)$ is the number of frames used to train the NB unit in i-th ground-truth beat segment.

3.2. Language modeling

The language modeling layer is an essential part in the proposed beat detection system. Its main target is to provide statistical information about beat sequences and beat periodicity. The dictionary for the beat/downbeat tracking task consists of two word classes: beat and downbeat words. Each word from the dictionary is characterized by the duration information.

For each word, a unit-level transcription is provided. It consists of a pre-attack unit, followed by an attack unit and a number D of NB units that define the duration factor. The first 7 words of the dictionary are provided in Table 1.

Table 1: Dictionary for the beat/downbeat tracking task

Word	Unit transcription
beat20	BTp BTa 20NB
downbeat20	DBTp DBTa 20NB
beat21	BTp BTa 21NB
downbeat21	DBTp DBTa 21NB
beat22	BTp BTa 22NB
downbeat22	DBTp DBTa 22NB
beat23	BTp BTa 23NB

Having ground-truth annotations for both beat and downbeat, one can collect training text from it and learn the statistics on possible word sequences. Training language models starts with the extraction of input text from the ground-truth labels. Each sentence is composed of word tokens defined as described above. The duration information for each word is extracted from the time instants corresponding to the boundaries of the segment. In order to take into account all possible tempo variations, scaling factors in the range $0.2 - 2.0$ are applied. As a consequence, a number of sentences is extracted from each ground-truth song. An example of the training text extracted from a very short song is given in Table 2. Symbols $< s >$ and $< /s >$ denote the beginning and the end of a musical piece respectively. The extracted text is given as an input to train N-gram language models.

Table 2: Text extracted from the ground-truth labels

$< s >$ downbeat52 beat52 beat52 beat52 downbeat52 ... beat52 $< /s >$
$< s >$ downbeat54 beat54 beat54 beat54 downbeat54 ... beat53 $< /s >$
$< s >$ downbeat56 beat55 beat56 beat55 downbeat55 ... beat55 $< /s >$
$< s >$ downbeat57 beat57 beat57 beat57 downbeat57 ... beat57 $< /s >$
...
$< s >$ downbeat94 beat94 beat94 beat94 downbeat94 ... beat94 $< /s >$
$< s >$ downbeat96 beat96 beat96 beat96 downbeat95 ... beat95 $< /s >$
$< s >$ downbeat98 beat97 beat98 beat97 downbeat97 ... beat97 $< /s >$

3.3. Beat/downbeat detection

The process of beat structure extraction starts with feature vector extraction for a given test song as described in section 2. The extracted feature vectors are then passed to the decoder. Similarly to the approach of multiple-pass decoding, which has been successfully used in speech recognition [17], the decoding procedure consists of two steps. In the first step, time-and-space efficient bigram language model is applied in the stage of Viterbi decoding, producing a lattice. The different lattice nodes denote time instants and lattice arks denote different hypotheses about beat and downbeat events. In the second step, the obtained lattice is rescored applying more sophisticated 4-gram language models on the reduced search space. A set of important parameters here includes acoustic model weight, language model weight and insertion penalty. Finally, the obtained transcription labels are matched against ground-truth.

4. EXPERIMENTAL RESULTS

Following MIREX evaluations, the scoring methods were taken from the beat evaluation toolbox and are described in [18]. **F-Meas.** and **DBt F-Meas.** are F-measures calculated using 70 ms precision window for beats and downbeats respectively. **Cemgil** is calculated using a Gaussian error function with 40ms standard deviation. **Goto** score measures correct or incorrect tracking based on statistical properties of a beat error sequence. The system was tested using 2-fold cross-validation on a dataset that is composed of 72 modern pop songs and used for the evaluations inside the Quaero project. An additional evaluation was performed using the Hainsworth database. Since this database does not contain downbeat information, **DBt F-Meas.** is excluded from the evaluation results. The results for different feature vector configurations are given in Table 4, where algorithm *Davies* corresponds to the results obtained with the Sonic Annotator software[2] with *Bar and Beat Tracker* QM Vamp plug-in[3]. The feature vector configuration is given in Table 3.

Table 3: Feature vector configurations

1dims	ODF	-	-	-
2dims	ODF	CVF 0.4s window	-	-
3dims	ODF	CVF 0.4s window	CVF 2s window	-

Table 4: MIREX-based system performance.

Algorithm	**F-Meas.**	**DBt F-Meas.**	**Cemgil**	**Goto**	**McK. P-score**
Quaero dataset					
1dims	81.88	30.52	77.20	79.17	79.96
2dims	86.53	56.64	81.35	86.11	84.81
3dims	85.32	60.09	80.22	84.72	83.84
DAVIES	87.23	64.36	77.45	80.56	84.15
Hainsworth dataset					
1dims	73.23	-	61.61	62.43	73.80
2dims	76.03	-	63.98	65.19	75.92
3dims	74.01	-	62.06	62.43	74.15
DAVIES	75.93	-	61.73	66.85	76.87

The results for beat estimation indicate that the proposed method with *2dims* feature vector configuration performed nearly as well as the reference method by davies et al. in terms of **F-meas**. Better results are obtained with our approach for other metrics (**Cemgil** and **Goto**) for the Quaero dataset. Adding CVF feature component with the context length of 0.4 sec to a single ODF shows a significant increase in performance for both beats and downbeats estimation. However, downbeat F-measure is further improved by adding the third feature vector dimension, which is CVF with the context length of 2 sec. Needless to mention a slight decrease in the beats F-measure estimate in comparison with the *2dims* configuration. Nevertheless, these two features do improve the downbeat estimation results for the proposed method. Tests on the Hainsworth dataset proved the advantage of *2dims* feature configuration over *3dims* one for beat detection. **F-meas.** and **Cemgil** for *2dims* configuration outperformed the reference method.

Participation in MIREX 2011 Audio Beat Tracking[4] contest was an excellent opportunity to compare the performance of the proposed

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

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

[4] http://www.music-ir.org/mirex/wiki/2011:Audio_Beat_Tracking

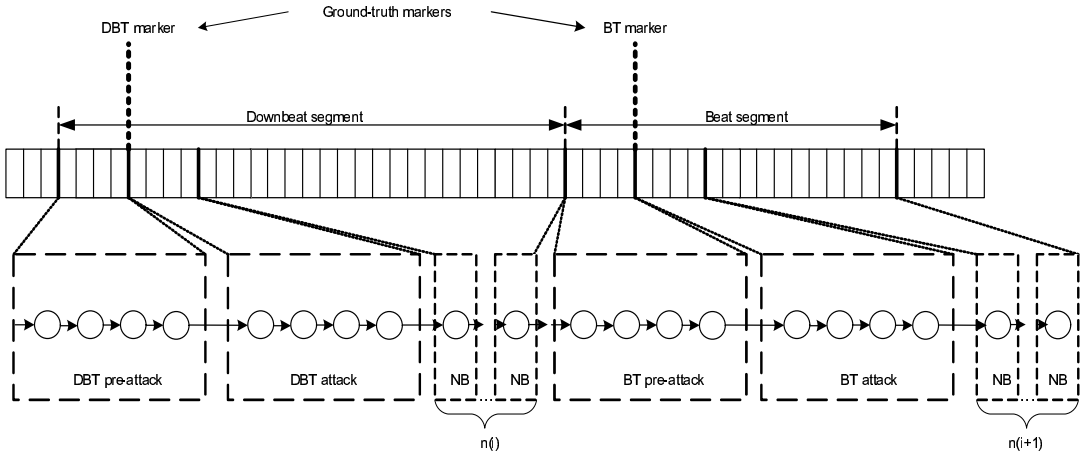

Fig. 1: Unit-based HMM training

system with many other systems. Two different systems, which are KFRO1 and KFRO2 were submitted. KFRO1 corresponds to the 2dims feature vector configuration, while KFRO2 correponds to the 3dims one. Model parameters were estimated using the Quaero dataset. Experimental results showed that both systems performed well on the "MCK" dataset, showing F-measure value very close to the top result.

5. CONCLUSION

The experimental results have shown that the proposed probabilistic approach to simultaneous estimation of beat/downbeat positions from audio is effective. The introduced explicit modeling of beat segment duration in the language modeling layer proved to be effective for solving the output labels periodicity problem. In addition, the system is also flexible and can be trained on other musical styles. Further improvement could be gained by incorporating tempo estimation into the model and by utilizing high-level features to enhance downbeat estimation. Another interesting investigation can be conducted in the area of application of multi-stream HMMs. Splitting feature vector into a number of separate streams and assigning different stream weights could be effective.

The potential of the proposed approach is clearly shown. Indeed, although the training material is sufficient to validate the approach, it is still small and a longer training would undoubtedly allow to improve those first results. Furthermore, another potential improvement could be brought by building genre specific language models.

6. REFERENCES

[1] M. Alonso, G. Richard, and B. David, "Accurate tempo estimation based on harmonic + noise decomposition," *EURASIP J. Appl. Signal Process.*, vol. 2007, pp. 161–175, January 2007.

[2] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings," in *Proceedings of the 2009 ISMIR Conference*, Kobe, Japan, 2009.

[3] G. Peeters, "Template-based estimation of time-varying tempo," *EURASIP J. Appl. Signal Process.*, vol. 2007, no. 1, pp. 158–158, 2007.

[4] D. P. W. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[5] A. P. Klapuri, A. J. Eronen, and J. T. 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.

[6] M. Mattavelli, G. Zoia, and R. Zhou, "Music onset detection based on resonator time frequency image," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1685–1695, 2008.

[7] A. Holzapfel and Y. Stylianou, "Beat tracking using group delay based onset detection," in *ISMIR*, 2008, pp. 653–658.

[8] G. Peeters, "Beat-tracking using a probabilistic framework and linear discriminant analysis," in *Proc. DAFX*, 2009.

[9] Y. Shiu and C.-C. J. Kuo, "A hidden markov model approach to musical beat tracking," in *Proc. ICASSP*, 2008.

[10] K. R. Fitz and S. A. Fulop, "A unified theory of time-frequency reassignment," *CoRR*, vol. abs/0903.3080, 2009.

[11] M. Khadkevich and M. Omologo, "Time-frequency reassigned features for automatic chord recognition," in *Proc. ICASSP*, may 2011, pp. 181 –184.

[12] G. Peeters, "Time variable tempo detection and beat marking," in *International Computer Music Conference (ICMC'05)*, Barcelona, Spain, 2005.

[13] 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.

[14] M. E. P. Davies and M. D. Plumbley, "A spectral difference approach to downbeat extraction in musical audio," in *Proc. EUSIPCO*, 2006.

[15] F. Brugnara, R. De Mori, D. Giuliani, and M. Omologo, "Improved connected digit recognition using spectral variation functions," in *Proc. ICSLP*, 1992.

[16] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on hidden markov models," *Speech Communication*, vol. 12, no. 4, pp. 357–370, 1993.

[17] D. Jurafsky and J. H. Martin, Eds., *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2000.

[18] M. E. P. Davies, N. Degara, and M. D. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Tech. Rep., Queen Mary University of London, Centre for Digital Music, 2009.