A MUSIC STRUCTURE INFORMED DOWNBEAT TRACKING SYSTEM USING SKIP-CHAIN CONDITIONAL RANDOM FIELDS AND DEEP LEARNING

Magdalena Fuentes[1,2] *Brian McFee*[3,4] *Hélène C. Crayencour*[1] *Slim Essid*[2] *Juan Pablo Bello*[3,*]

[1] L2S, CNRS-Univ.Paris-Sud-CentraleSupélec
[2]LTCI, Télécom ParisTech, Univ. Paris-Saclay
[3]Music and Audio Research Laboratory, New York University
[4]Center for Data Science, New York University

ABSTRACT

In recent years the task of downbeat tracking has received increasing attention and the state of the art has been improved with the introduction of deep learning methods. Among proposed solutions, existing systems exploit short-term musical rules as part of their language modelling. In this work we show in an oracle scenario how including longer-term musical rules, in particular music structure, can enhance downbeat estimation. We introduce a skip-chain conditional random field language model for downbeat tracking designed to include section information in an unified and flexible framework. We combine this model with a state-of-the-art convolutional-recurrent network and we contrast the system's performance to the commonly used Bar Pointer model. Our experiments on the popular Beatles dataset show that incorporating structure information in the language model leads to more consistent and more robust downbeat estimations.

Index Terms— Downbeat Tracking, Music Structure, Deep Learning, Skip-Chain Conditional Random Fields, Convolutional-Recurrent Neural Networks.

1. INTRODUCTION

Downbeat tracking consists of retrieving the first beat of each bar in a music excerpt. It is an important task in Music Information Retrieval (MIR) which provides useful pre-processing tools for several applications such as automatic music transcription [1], computational musicology [2] or rhythm similarity [3].

In general, the pipeline of downbeat tracking systems based on deep learning consists of a first stage of low-level feature extraction, where representations such as mel-spectrograms or chromagrams are computed and synchronized to some temporal grid [4, 5]. Subsequently, the representations are input to a deep neural network (DNN) which is used either to extract more complex features to feed another machine learning model [6], or to produce a likelihood of possible downbeat candidates [7, 5, 8]. The likelihood is fed to a language model such as Hidden Markov Models (HMMs) or Dynamic Bayesian Networks (DBNs), which is used to smooth the estimation obtained by the DNNs. In most existing systems, the language models use local transition rules and observations, which usually model relations between neighbouring events in time, up to the bar scale [5, 9, 8], not exploiting longer-term musical dependencies.

In diverse music styles such as pop, rock or classical music the format of repeated sections is common practice. Typically, a song would feature an intro, verses and refrain, noted using symbols such as '*AAB*'. It is then likely that music objects such as chord progressions or metric structure are similar among repetitions. Considering information from several instances of the same section is likely to provide complementary information and thus improve models' estimations [10].

1.1. Our contributions

In this paper we test whether for a fixed system, the introduction of music structure information improves downbeat tracking performance. We propose a novel skip-chain CRF language model for downbeat tracking that incorporates structure information in a flexible way, and we assess its performance using a convolutional-recurrent network for the observations. We compare the performance of this language model with a popular Bayesian network model [11, 5], showing its advantages. We also contrast the model to simpler approaches for including structural information. We validate our claim by assessing the different methods using annotated beats and sections on the Beatles songs dataset, to isolate noise due to beat/section estimation. Our experiments show that including music structure in language models for downbeat tracking helps in challenging cases, obtaining more consistent downbeat estimations.

2. RELATED WORK

Downbeat tracking has received substantial attention in the MIR community in recent years. Several methods exploiting deep learning were proposed, covering different DNN architectures that have been shown to be adequate for downbeat tracking: convolutional neural networks (CNNs) [4, 8, 12], bi-directional long-short term memory networks [7], bi-directional gated recurrent units (Bi-GRUs) [5] and a combination of convolutional and recurrent networks (CRNNs) [13]. Different input representations such as mel-spectrograms, chromagrams, multi-band spectral flux, low-frequency spectrograms and combinations of them have been also explored [14, 15]. Regardless of the architecture, most systems rely on HMMs or DBNs for the final downbeat inference. In particular, the Bar Pointer model [11] has received considerable attention and it has been refined in different scenarios such as inferring tempo, time signature and downbeats [12] or long metrical cycles [16]. In these models the relation between latent-states are only modeled between time consecutive events, ignoring longer term dependencies, which is a considerable simplification. Music has rich and interrelated dependencies within different time scales, thus accounting for music attributes in both short and long term is a more realistic approach.

[*] This work is partially supported by the NYU Abu-Dhabi Research Enhancement Fund.

The use of music structure information to inform other MIR tasks has been previously explored. Dannenberg [17] proposed a system that incorporates structure information to perform beat tracking. He considers beat tracking as an optimization problem where the goal is to infer the best beat sequence given the constraint of a chroma similarity matrix. Mauch et al. [10] addressed the use of musical structure to enhance automatic chord transcription. Their method consists in averaging the chroma feature inputs in repeated sections and replace the occurrences by the average of the chroma. The authors showed the suitability of this approach for chord recognition, since it helps in obtaining consistent and more readable chord progressions. Although both methods showed promising results, they use very simple features and have limited flexibility to include other musically relevant information. Papadopoulos et al. [18] proposed the use of Markov Logic Networks to include music structure information for chord transcription in a flexible way. The authors model the probability of a chord progression to occur in repeated occurrences of similar sections given the underlying chroma observations. This work showed that graphical models are capable of incorporating musical knowledge in different time scales in a flexible and unified manner. The main limitations of this work are the use of simple features and the very slow inference.

Among probabilistic graphical models, Conditional Random Fields (CRFs) are discriminative classifiers for structured data prediction, which allow for modeling complex and interrelated properties both in the observations and output labels at different time scales, thus making them appealing for music modelling. Linear-chain CRFs have been successfully applied in MIR tasks such as beat tracking [19] and downbeat tracking [6], but those models are still limited to relating time neighbouring output labels. In turn, skip-chain CRFs have been successfully applied in Natural Language Processing (NLP), Sutton et al. [20] used skip-chain CRFs in a simple speaker identification task on seminar announcements, showing that they outperform linear-chain CRFs while modelling more complex structure between words.

3. PROPOSED METHOD

Our model consists of two main stages: first we compute the downbeat likelihood using a CRNN, and then we obtain the downbeat positions with a structure-informed skip-chain CRF (SCCRF). The different components are explained in the following.

3.1. Convolutional-Recurrent Network

We use the CRNN proposed in our previous work [13] to obtain the observations of our model. It consists of an ensemble of two CRNNs, representing the harmonic and percussive content of the signal respectively. In particular, to simplify our analysis we use the CRNN of the *CUBd* configuration in [13] which has beat-synchronous features and no structured encoding.

Briefly explained, a set of beat-synchronous features describing percussive content, based on a multi-band spectral flux, and a harmonic content's representation based on the Chroma-Log-Pitch [21] are the inputs of each CRNN in the ensemble. Each beat is fed into a CRNN considering a context window of approximately one bar. First the CNN processes each window independently, and its output is then fed to the recurrent network. The CNN architecture consists of a cascade of convolutional and max-pooling layers, with dropout used during training to avoid over-fitting, and batch normalization to avoid small values within the DNN that could hurt performance.

The recurrent layer consists of a 512-dimension Bi-GRU. The last layer of the network is a fully connected layer with a sigmoid activation, resulting in a downbeat likelihood per time unit. The optimization of the model parameters is carried out by minimizing the binary-cross-entropy between the estimated and reference observations. We refer the interested reader to [13] for further information.

3.2. Skip-Chain Conditional Random Field model

3.2.1. Model

The SCCRF model consists of a linear-chain CRF with additional long-distance connections between nodes, so called skip-connections [20]. The evidence at one endpoint of the skip connection influences the label at the other distant endpoint, as illustrated in Figure 1. Formally, the conditional probability of a label sequence $\mathbf{y} = (y_1, ..., y_T)$ of length T given an input sequence of observations $\mathbf{x} = (x_1, ..., x_T)$ is given by:

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{t=1}^{T} \psi_t(y_t, y_{t-1}, \mathbf{x}) \phi(y_t, \mathbf{x}) \prod_{(u,v) \in \mathcal{I}} \psi_{uv}(y_u, y_v, \mathbf{x})$$

where ψ_t is the transition potential for the linear-chain (neighbouring nodes), ϕ is the observation potential and ψ_{uv} is the potential of the skip connections, which is defined for a pre-selected subset of nodes $\mathcal{I}$. Note that the transition and observation potentials play a similar role to transition and observation probabilities in DBNs or HMMs, with the difference that the potentials in a CRF are not proper probabilities and thus the need for the normalization factor $Z(\mathbf{x})$. The skip potential also differs by modeling interactions between distant nodes, which is possible given the flexibility of CRFs in terms of independence assumptions [22].

We consider a set of labels $\mathcal{Y}$ which represents the beat position inside a bar. Following [5, 7], we consider bar lengths of 3 and 4 beats, corresponding to 3/4 and 4/4 meters, and we model the beat position inside different time signatures as separate labels. The first beat in a 3/4 time signature will have a different label compared to the first beat of a 4/4 time signature. The output labels $\mathbf{y}$ are a function of two variables: the beat position $b \in \mathcal{B} = \{1, ..., b_{max}(r)\}$ and the number of beats per bar $r \in \mathcal{R} = \{r_1, ..., r_n\}$, which relates to the time signature of the piece. This results in seven possible labels $\mathcal{Y} = \{1, .., 7\}$, one for each position b in $\mathcal{R} = \{3, 4\}$, i.e. the second beat of a 4/4 bar would be $b = 2$, $r = 4$ and $y = 5$.

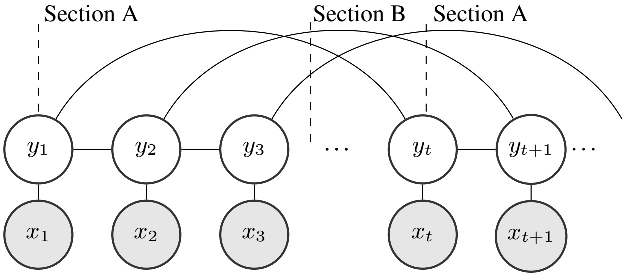

Fig. 1. SCCRF graph. Observations and labels are indicated as gray and white nodes respectively. Beats of repeated section occurrences are connected to each other.

Transition potential ψ_t: the transition potential depends only on consecutive labels $\psi_t(y_t, y_{t-1}, \mathbf{x}) = \psi_t(y_t, y_{t-1})$. Similar to [5, 6, 19], it forces the beat position b inside a bar to increase by one up to the maximum bar length considered, and to switch to one at the end of the bar. Time signature changes are unlikely and only allowed at the end of the bar. Formally:

$$\psi_t(b_t, b_{t-1}, r_t, r_{t-1}) = \begin{cases} 1 & \text{if } b_t = b_{t-1} + 1 \\ 1-p & \text{if } r_t = r_{t-1},\ b_t = 1,\ b_{t-1} = r_{t-1} \\ p & \text{if } r_t \neq r_{t-1},\ b_t = 1,\ b_{t-1} = r_{t-1} \\ 0 & \text{otherwise} \end{cases}$$

where $p = 10^{-6}$ is the probability of changing the time signature. We chose the value of p similarly to the DBN in [5, 13].

Observation Potential ϕ: the observation potential depends on the current observation x_t so that $\phi(y_t, \mathbf{x}) = \phi(y_t, x_t)$, and is given by:

$$\phi(b_t, x_t) = \begin{cases} a(t) & \text{if } b_t = 1 \\ 1 - a(t) & \text{otherwise} \end{cases}$$

where $a(t)$ is the downbeat likelihood estimated by the CRNN.

Skip potential ψ_{uv}: the skip potential depends on two labels y_u, y_v which are not neighbours in the time axis, and is independent from the observations: $\psi_{uv}(y_u, y_v, \mathbf{x}) = \psi_{uv}(y_u, y_v)$. It is given by:

$$\psi_{uv}(y_u, y_v) = \psi(b_u, b_v, r_u, r_v) = \begin{cases} \alpha & \text{if } b_u = b_v,\ r_u = r_v \\ \frac{1-\alpha}{|\mathcal{Y}|-1} & \text{otherwise.} \end{cases}$$

When connected, two distant nodes y_u and y_v are constrained to have the same label by a factor α, and to have different labels by $\frac{1-\alpha}{|\mathcal{Y}|-1}$ (where the different labels are equally possible). We found the best value $\alpha = 0.3$ on a grid search between 0 and 1.

Graph structure: the subset $\mathcal{I}$ that determines the skip connections is obtained using the musical section labels, given as input to the model. If a section s has multiple occurrences, we connect the first beats of each occurrence of s to each other, the second ones to each other, and so on, as shown in Figure 1. If the section repetitions are of different lengths, we connect the beats until the shortest section length is reached. We connect the beats of each occurrence of s, i.e. the more repetitions, the more connections.

3.2.2. *Inference*

Because the loops in the SCCRF can be long and overlapping, exact inference is intractable [20]. For this reason, we perform loopy-belief propagation (LBP) for inference [23]. LBP is an algorithm based on message updates between nodes. Although the algorithm is not exact and it is not guaranteed to converge if the model is not a tree, it has been shown to be empirically successful in a wide variety of domains such as text processing, vision, and error-correcting codes [24]. In the LBP algorithm, each node i sends a message to its neighbours, where neighbours are directly connected nodes in the graph no matter how distant those nodes are in time. The message from node i to node j is given by $\mu_{ij}(y_j) = \max_{y_i} \phi_i(y_i, x_i)\Psi_{ij}(y_i, y_j)\prod_{k \in N(i)\backslash j} \mu_{ki}(y_i)$, where $N(i)\backslash j$ indicates the neighbours of node i except node j, and $\Psi_{ij} = \psi_t$ if $|i - j| = 1$ and $\Psi_{ij} = \psi_{uv}$ otherwise. This message exchange continues until the messages converge. Once convergence is achieved the belief of each node is computed as $b_i(y_i) = \phi_i(y_i, x_i)\prod_{j \in N(i)} \mu_{ij}(y_i)$ and final inference is performed by $\mathbf{y}^* = \arg\max_{\mathbf{y}} b(\mathbf{y})$.

4. EXPERIMENTS

4.1. Experimental Setup

We investigate whether adding long-range interaction terms to language models can improve downbeat tracking in a fixed setup. To that purpose we use our state-of-the-art CRNN-DBN system proposed in [13] as baseline, and we compare it to different music structure-informed systems. We use manually annotated beats and sections to avoid noise due to bad beat/segment estimations, thus focusing this study on the analysis of the usefulness of music structure information for the downbeat tracking task. We leave the issue of exploiting instead automatically detected (hence imperfect) music sections for future work.

4.1.1. *Datasets*

We use the Beatles dataset, since it has beat, downbeat and music structure annotations. It consists of 179 Beatles songs up to 8h 01m of music. We follow the leave-one-dataset-out evaluation scheme of [5, 8, 13] and we train the CRNN network with 6 Western music datasets leaving the Beatles dataset out. Those datasets are: *Klapuri*, *R. Williams*, *Rock*, *RWC Pop*, *Ballroom* and *Hainsworth*, to a total of 35h 03m of music.

4.1.2. *Implementation, training and evaluation metrics*

The deep learning models were implemented with Keras 2.0.6 and TensorFlow 1.2.0 [1, 9]. We use the ADAM optimizer [15] with default parameters. We stop training after 10 epochs without improvement on validation accuracy, up to a maximum of 100 epochs. The low-level representations were extracted using the madmom library in Python [25] and mapped to the beat grid. The SCCRF was implemented using the factorgraph library.[1] We report the F-measure score following previous works. To determine statistical significance, we perform a Friedman test followed by post-hoc Conover tests for pairwise differences using Bonferroni-Holm correction for multiple testing [26]. Message convergence in the LBP algorithm is given by $|\mu_{ij}^m - \mu_{ij}^{m-1}| < \tau\ \forall i, j$ where m is the current iteration and τ is a tolerance; or a maximum amount of iterations is reached. We set $\tau = 10^{-8}$ and consider a maximum of 3000 iterations. Messages are normalized at each iteration to avoid values going to zero easily in practice. Inference takes a median time of 3.6s on 2m 30s of music using an Intel Xeon CPU E5-2643 v4 @ 3.40GHz.

4.2. Results and discussion

We compare the SCCRF performance to the DBN in [5, 13], which is our state-of-the-art baseline. We employ the downbeat likelihood estimated by the CRNN as observations for both language models. To compare the SCCRF to simpler approaches aware of structure information, we enhance the CRNN estimation before performing inference with the DBN by: averaging the input representations of section repetitions, replacing the occurrences by the average and feeding the network with the averaged features; and applying the same idea but averaging the downbeat likelihood estimation of repeated sections instead. We name these two approaches *DBN_AVF* and *DBN_AVA* respectively. We include a non structure-informed version of our model, without the skip potentials, in order to assess possible differences over the DBN due to the inference method. This model is a linear-chain CRF, and is denoted as *LCCRF*. Figure 3 summarizes the results of the different configurations.

The standard DBN approach benefits from adding structural information in terms of reducing variance in the performance. The SCCRF model brings the most benefit out of the compared models, due to improvement in difficult cases. The LCCRF performance is equivalent to the DBN, showing that the LBP algorithm is not the

[1] https://github.com/mbforbes/py-factorgraph.

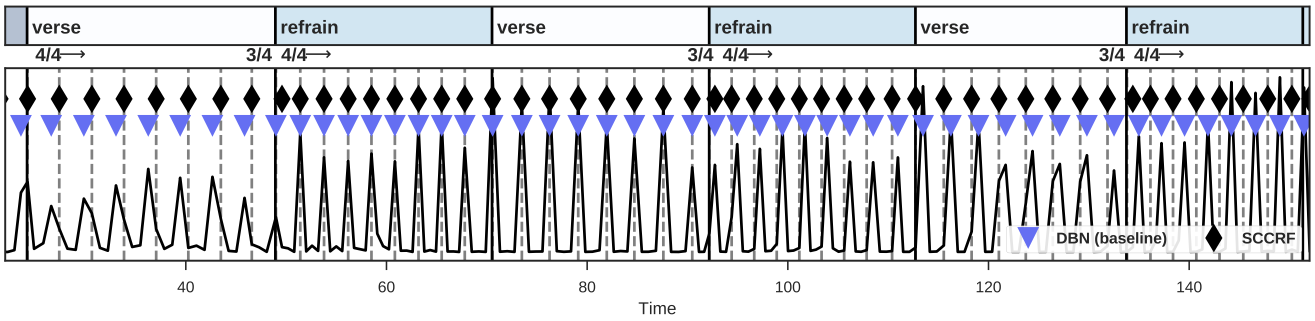

Fig. 2. Excerpt of '*Blue Jay Way*'. Upper figure shows sections and bottom figure shows model's estimations. Dashed lines denote ground-truth downbeat positions, the continuous curve is the downbeat likelihood estimated by the CRNN (without any structure information). The SCCRF improves downbeat tracking performance from 0.35 to 0.72 F-measure with respect to the non-structured DBN of [13].

source of improvement but the addition of structure. Mean and median performances of the SCCRF and DBN models are similar, with their difference not being statistically significant.

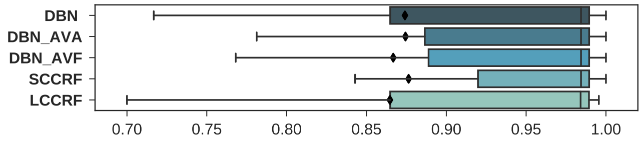

Fig. 3. F-measure scores. Boxes show median value and quartiles, whiskers the rest of the distribution. Black dots denote mean values.

Figure 2 illustrates a typical example where the inclusion of structure information and the flexibility of the SCCRF model help in the downbeat estimation. In this excerpt the CRNN likelihood estimation is inconsistent in different instances of the same section, and in particular, it is correct in some instances and wrong in others. For instance, the downbeat likelihood has peaks in the right positions in one verse and the estimation is partially correct or incorrect in the two others. The SCCRF downbeat estimation is consistent over all section occurrences despite the discordant likelihood estimations, and more accurate in the overall performance. In turn, the DBN is not able to overcome the likelihood estimation errors, which is expected given the limited information it handles and the hard transition constraints. The time signature of this song is mostly 4/4, with the exception of one bar in 3/4 at the end of each verse. The SCCRF finds there is a 3/4 transition bar between the verse and the refrain, but it estimates that the 3/4 bar is in the refrain instead of the end of the verse. We hypothesize that this is due to the observation values which give more evidence of having a 3/4 bar in the position where the model finds it. The combination of the information in different time scales and the inference carried out globally make the model capable of identifying rare music variations and to fit the global time signature consistently.

Figure 4 shows an example of the downbeat estimation with the DBN with structure-enhanced CRNN observations. The three likelihood estimations correspond to the three models DBN, DBN_AVF and DBN_AVA, and the downbeat positions found by the DBN using each set of observations are shown with dots of the respective color. We noticed that the inclusion of structure information through averaging features (DBN_AVF) has limited impact on the performance. Averaging the likelihood of different section occurrences presents the advantage that the likelihood has higher values where the CRNN finds strong evidence of downbeat occurrences and smaller values when it is unclear, so the average compensates with the correct estimation in many cases. Nevertheless, in examples like the one of Figure 4 which have shifted likelihood estimations and transition bars, the downbeat estimation of DBN_AVF and DBN_AVA do not achieve the consistency of the SCCRF on the different occurrences of the same section, indicating that it is necessary to have a flexible and robust language model to account for this information. Finally, we noticed examples where the DBN performance is better than the SCCRF. Those are mainly examples where the annotations have contradictions such as two occurrences of the same section beginning in different parts of a bar, so the skip connections of the SCCRF model are misaligned and the information they provide is inaccurate.

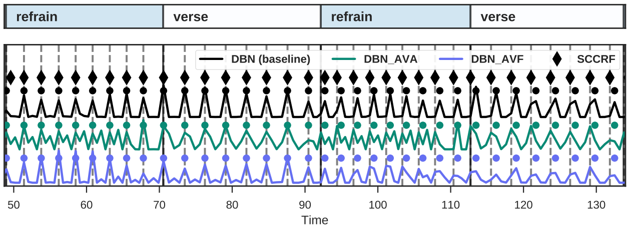

Fig. 4. Excerpt of '*Blue Jay Way*'. Sections are shown on top and DBN estimations with enhanced CRNN observations in the bottom. Dots denote the downbeat positions estimated by the DBN in each case. Dash lines denote the ground-truth positions.

5. CONCLUSIONS AND FUTURE WORK

We have presented a Skip-Chain Conditional Random Field language model for downbeat tracking which exploits music structure information in a unified and flexible manner. We have shown that using knowledge of repeating structure in the language model improves the downbeat estimation over state-of-the-art approaches by providing consistency among occurrences of the same section, being able to handle rare music variations. The proposed method can be directly applied to beat tracking, and easily extended to the joint tracking of beats and downbeats by incorporating suitable potentials. The structure of the skip-chain graph could be obtained by estimating boundaries and labels of sections with an external algorithm, in a fully automatic fashion. We will address this as an extension of our method to automatically estimate the graph structure. Considering information about rhythmic patterns as an intermediate temporal scale between bars and sections is a promising idea and will be also explored in future work.

6. REFERENCES

[1] M. Mauch and S. Dixon, "Simultaneous estimation of chords and musical context from audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1280–1289, aug 2010.

[2] M. Hamanaka, K. Hirata, and S. Tojo, "Musical structural analysis database based on GTTM," in *15th International Society for Music Information Retrieval Conference*, 2014, ISMIR.

[3] J. Paulus and A. Klapuri, "Measuring the similarity of rhythmic patterns," in *Proc. of the 3rd International Society for Music Information Retrieval Conference*, 2002, ISMIR.

[4] S. Durand, J. P. Bello, B. David, and G. Richard, "Downbeat tracking with multiple features and deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015, ICASSP.

[5] F. Krebs, S. Böck, M. Dorfer, and G. Widmer, "Downbeat tracking using beat synchronous features with recurrent neural networks.," in *17th International Society for Music Information Retrieval Conference*, 2016, ISMIR.

[6] S. Durand and S. Essid, "Downbeat detection with conditional random fields and deep learned features," in *17th International Society for Music Information Retrieval Conference*, 2016, ISMIR.

[7] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *17th International Society for Music Information Retrieval Conference*, 2016, ISMIR.

[8] S. Durand, J. P. Bello, B. David, and G. Richard, "Robust downbeat tracking using an ensemble of convolutional networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 76–89, Jan 2017.

[9] A. Srinivasamurthy, A. Holzapfel, A. T. Cemgil, and X. Serra, "Particle filters for efficient meter tracking with dynamic bayesian networks," in *16th International Society for Music Information Retrieval Conference*, 2015, ISMIR.

[10] M. Mauch, K. C Noland, and S. Dixon, "Using musical structure to enhance automatic chord transcription.," in *10th International Society for Music Information Retrieval Conference*, 2009, ISMIR.

[11] N. Whiteley, A. T. Cemgil, and S. J. Godsill, "Bayesian modelling of temporal structure in musical audio.," in *7th International Society for Music Information Retrieval Conference*. Citeseer, 2006, ISMIR.

[12] A. Holzapfel, F. Krebs, and A. Srinivasamurthy, "Tracking the "odd": Meter inference in a culturally diverse music corpus," in *15th International Society for Music Information Retrieval Conference*, 2014, ISMIR.

[13] M. Fuentes, B. McFee, H. Crayencour, S. Essid, and J.P. Bello, "Analysis of common design choices in deep learning systems for downbeat tracking," in *19th International Society for Music Information Retrieval Conference*, 2018, ISMIR.

[14] S. Durand, J. P. Bello, B. David, and G. Richard, "Feature adapted convolutional neural networks for downbeat tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, ICASSP.

[15] S. Böck, F. Krebs, and G. Widmer, "A multi-model approach to beat tracking considering heterogeneous music styles.," in *15th International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[16] A. Srinivasamurthy, A. Holzapfel, A. T. Cemgil, and X. Serra, "A generalized bayesian model for tracking long metrical cycles in acoustic music signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, ICASSP.

[17] R. B. Dannenberg, "Toward automated holistic beat tracking, music analysis and understanding.," in *6th International Society for Music Information Retrieval Conference*, 2005, ISMIR.

[18] H. Papadopoulos and G. Tzanetakis, "Exploiting structural relationships in audio music signals using markov logic networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2013, ICASSP.

[19] T. Fillon, C. Joder, S. Durand, and S. Essid, "A conditional random field system for beat tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, ICASSP.

[20] C. Sutton and A. McCallum, *An introduction to conditional random fields for relational learning*, vol. 2, Introduction to statistical relational learning. MIT Press, 2006.

[21] M. Müller and S. Ewert, "Chroma toolbox: Matlab implementations for extracting variants of chroma-based audio features," in *12th International Society for Music Information Retrieval Conference*, 2011, ISMIR.

[22] H. M. Wallach, "Conditional random fields: An introduction," *Technical Reports (CIS)*, p. 22, 2004.

[23] James Coughlan, "A tutorial introduction to belief propagation," *The Smith-Kettlewell Eye Research Institute*, 2009.

[24] J. S Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," *Exploring artificial intelligence in the new millennium*, vol. 8, pp. 236–239, 2003.

[25] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "madmom: a new python audio and music signal processing library," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, ACMMM.

[26] S. Garcia and F. Herrera, "An extension on"statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, Dec 2008.