

MODELING BEATS AND DOWNBEATS WITH A TIME-FREQUENCY TRANSFORMER

Yun-Ning Hung^{2,*}, Ju-Chiang Wang¹, Xuchen Song¹, Wei-Tsung Lu¹, and Minz Won¹

¹ ByteDance, Mountain View, CA, USA

² Center for Music Technology, Georgia Institute of Technology, Atlanta, GA, USA

amy.hung@gatech.edu, {ju-chiang.wang, xuchen.song, weitsung.lu, minzwon}@bytedance.com

ABSTRACT

Transformer is a successful deep neural network (DNN) architecture that has shown its versatility not only in natural language processing but also in music information retrieval (MIR). In this paper, we present a novel Transformer-based approach to tackle beat and downbeat tracking. This approach employs SpecTNT (Spectral-Temporal Transformer in Transformer), a variant of Transformer that models both spectral and temporal dimensions of a time-frequency input of music audio. A SpecTNT model uses a stack of blocks, where each consists of two levels of Transformer encoders. The lower-level (or spectral) encoder handles the spectral features and enables the model to pay attention to harmonic components of each frame. Since downbeats indicate bar boundaries and are often accompanied by harmonic changes, this step may help downbeat modeling. The upper-level (or temporal) encoder aggregates useful local spectral information to pay attention to beat/downbeat positions. We also propose an architecture that combines SpecTNT with a state-of-the-art model, Temporal Convolutional Networks (TCN), to further improve the performance. Extensive experiments demonstrate that our approach can significantly outperform TCN in downbeat tracking while maintaining comparable result in beat tracking.

Index Terms— Beat, Downbeat, Transformer, SpecTNT

1. INTRODUCTION

Beat is regarded as one of the fundamental rhythmic units perceived by humans, and a downbeat is conceived to occur at the first beat of a bar (measure), which usually indicates the start of a chord. Beat and downbeat tracking is concerned with developing algorithms to detect the beats and downbeats as pulse signals in music audio. This is a well-defined problem in music information retrieval (MIR) and has been a regular task in MIREX challenge¹ for over a decade.

In recent years, researchers proposed to employ deep neural network (DNN)-based methods to model beats and downbeats, which are regularly repeating events in the audio sequence. Böck et al. [1] successfully used a recurrent neural network (RNN) with input mel spectrogram features to jointly model beats and downbeats. Such a system requires annotated data to train and then predict accordingly. This has set up a data-driven paradigm for the subsequent systems, where improvement can be expected by either involving more training data or developing a more advanced machine learning algorithm.

Compared to beats, predicting downbeats still remains a challenge, because it requires a more complicated awareness of harmonic, timbral, and structural context, which may be related to chords, instrumentation, and percussive arrangements [2]. To better

model these aspects, one need to explore more local spectral information, and at the same time allow the important local information to be exchangeable among different temporal positions far apart in the audio sequence, since a measure may require a longer audio context to be discovered.

Recently, Transformer architecture [3, 4] has brought lots of attention due to its remarkable performance in natural language processing. For song-level classification tasks in MIR, researchers treat it as a temporal encoder to aggregate the temporal features to represent a music audio sequence [5, 6]. As a replacement of RNN, it also works well for chord recognition [7, 8]. However, these prior works did not attempt to make full exploitation of Transformer to model the spectral information that interacts with the temporal encoder as a whole. On the other hand, Transformer is data-hungry in nature. For many MIR tasks, annotations are very expensive, and the size of annotated data is typically small. Therefore, training an effective Transformer-based model without modification and data augmentation is likely infeasible.

In this paper, we propose to use SpecTNT (Spectral-Temporal Transformer in Transformer) [9], a variant of Transformer that models the spectrogram along both time and frequency axes. To the best of our knowledge, this is the first successful attempt to apply Transformer-based model to beat and downbeat tracking. The basic principle of SpecTNT stems from the interaction between two levels of Transformer encoders, namely *spectral Transformer* and *temporal Transformer*. The former is responsible for extracting the spectral features for each frame. Whereas the latter is capable of exchanging the local information over the time axis. Thanks to the hierarchical design [9], SpecTNT permits a smaller number of parameters compared to the original Transformer [4], so it can maintain good generalization ability for MIR tasks with a smaller size of training data. To further improve the performance, we also propose a novel structure that combines SpecTNT with a state-of-the-art (SOTA) model, temporal convolutional networks (TCN). We evaluate our proposed systems on various public datasets. Our results demonstrate that SpecTNT can achieve SOTA performance in downbeat tracking on most of the datasets while maintaining comparable performance in beat tracking. The combined structure of SpecTNT with TCN can further boost the performance.

2. RELATED WORK

Early approaches in beat tracking mostly focused on incorporating heuristic information into the systems. For example, Dixon et al. [10] proposed to estimate the tempo and beats by detecting salient rhythmic events at various metrical levels. Klapuri et al. [11] proposed a probabilistic model to process the music accents and estimate beat patterns. On the other hand, musical events such as chord

*The author conducted this work as an intern at ByteDance.

¹https://www.music-ir.org/mirex/wiki/MIREX_HOME

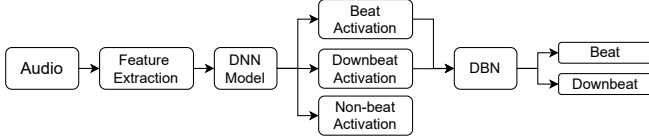


Fig. 1. Pipeline for the beat and downbeat tracking system.

changes and drum patterns were explored to facilitate the analysis of beat structures [12]. A two-stage approach was also proposed by [13] to estimate tempo and beat sequentially, where the process attempted to replicate human ability to tap along music.

More recently, data-driven approaches have been widely used [1], since they enable a system to automatically learn the relevant information to detect beats from annotated audio. Specifically, DNN architectures such as RNN [14] and TCN [15] were employed to effectively model the long sequence of beat events from audio. Multi-task approaches were also introduced to include more types of annotations in the learning system, such as tempo and downbeat [16, 17, 18, 19], owing to their high relevance to beats, and each task was improved by training multiple tasks altogether. These endeavors have improved beat tracking significantly, and the evaluation scores can easily achieve over 90% on many benchmarks.

As noted in the previous section, downbeat tracking still remains a challenge. Several prior instances tackled this problem by using musical features such as time-signature [20], onset positions [21], chords [22], and spectral difference in beats [23]. With machine learning approaches, we believe it can be improved by introducing a novel model architecture with more training data.

3. PROPOSED METHOD

3.1. System Overview

The system pipeline follows the standard setting [14, 17] as depicted in Figure 1. In this work, our contribution is to introduce two novel DNN models in the pipeline, and the rest of the modules remain as conventional. For feature extraction, we apply trainable harmonic filters to convert the input audio into a harmonic representation [24]. This step helps capture useful harmonic information and has proven beneficial in music auto-tagging [24]. Then, we define the DNN model to contain three modules: front-end, main model, and a linear layer. The DNN model predicts the likelihoods of beat, downbeat, and non-beat at each time step and forms three corresponding activation functions. Finally, a dynamic Bayesian network (DBN) [25] decodes the activation functions to output the timestamps of beats and downbeats. In what follows, we will detail three options of DNN models (two proposed and one baseline models) used in this work.

3.2. Spectral-Temporal Transformer in Transformer (SpecTNT)

An input harmonic representation is first processed by the front-end convolutional layers (called *ResNet* [26]). Then, we stack 3 residual units, each uses 256 feature maps and a kernel size of 3. This design is similar to that proposed in [6].

The main SpecTNT module (see Fig. 2(a)) is formed by stacking multiple SpecTNT blocks. A SpecTNT block consists of a *spectral encoder* and a *temporal encoder*. We use 64 feature maps with 4 attention heads for the spectral encoder to extract the spectral features into a set of *frequency class tokens* (FCTs) at each time step. Through the attention mechanism, a FCT can characterize useful

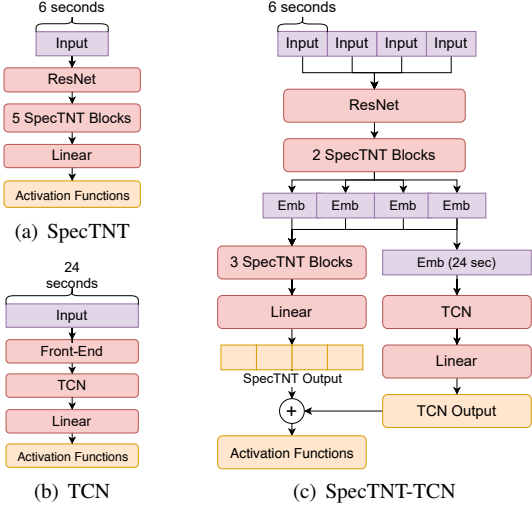


Fig. 2. Model architecture overview.

harmonic and timbral components, which may better represent a chord or instrumentation. For the temporal encoder, we use 256 feature maps and 8 attention heads. It enables the important local spectral information to be exchangeable via FCTs to pay attention to the beat/downbeat positions. More details can be found in [9]. Finally, we stack 5 SpecTNT blocks, followed by a linear layer to output three activation functions for beat, downbeat, and non-beat.

3.3. Baseline TCN

We adopt the TCN architecture [17, 15] with slight modifications as our baseline model (see Fig. 2(b)), since it is regarded as the current SOTA. We re-implemented the model following [17]. Readers can refer to the codes recently released by the authors in [27].

The TCN front-end is composed of three convolutional layers, each having a filter size of 20 and kernels with various sizes as described in [17]. Each convolutional layer also applies a max-pooling to downsample the frequency dimension. After the front-end, the main model contains 11 dilated convolutional layers. Each convolutional layer has two dilated convolution blocks, and each block has a twice the dilation rate of the another. This design has resulted in better performance according to [17]. Finally, a linear layer is used to output the three activation functions.

3.4. Combining SpecTNT and TCN

SpecTNT and TCN are fundamentally different in many aspects, so we believe they can be complementary to each other if they are combined. From our pilot study, we also found that modeling beats and downbeats jointly can result in a sub-optimal model for either beat or downbeat tracking. Since the amount of downbeat annotations is less than that of beat annotations, downbeat optimum will require more epochs to attain. However, longer training may lead to worse performance for beat, because the model over-fits the beat annotations. We will discuss this later in Section 4.4. To this end, we propose a fused architecture, called *SpecTNT-TCN* (see Fig. 2(c)). We hope it can gain more flexibility during training so as to facilitate a joint optimum for beat and downbeat tracking simultaneously.

Specifically, SpecTNT-TCN first uses 2 SpecTNT blocks at the front, and then arranges a two-stream structure that allocates a

TCN branch and 3 succeeding SpecTNT blocks in parallel. Because SpecTNT requires shorter input, as discussed in Section 4.1, we divide the input into smaller chunks at the SpecTNT front-end. The two branches have their own independent linear layers, and the final output is the average of the two branch outputs. The TCN branch follows the same TCN module described in Section 3.3. To sum up, during training, parameters of each branch are updated according to its own loss and the same SpecTNT parameters at the front are updated according to the added losses. Final outputs are the average of the outputs from each branch. From our observations, such design is better than directly merging the outputs of two full models (i.e., averaging the two outputs of Fig. 2(a) and Fig. 2(b)).

4. EXPERIMENTS

4.1. Implementation Details

For data augmentation, we adopt a similar method to that proposed in [17] for training. Specifically, the hop size to construct the STFT representation is randomly changed during training using a normal distribution with the estimated tempo from the training data as the mean and 5% standard deviation. This can result in more samples of various tempi. Then, multiple triangular filters are applied to STFT to obtain the harmonic representation. Following [17], we apply the same target widening strategy for the frame-wise beat and downbeat labels, where the neighbouring frames of an annotated frame are set to be positive, but with a lower weight of 0.5. We use a weighted binary cross-entropy to compare between each target and prediction.

Based on our pilot study, we determine the input length to be 6 seconds for SpecTNT and 24 seconds for TCN. Although several counterpart systems (e.g., [18] and [17]) use 6-second, we found 24-second works better for TCN. In SpecTNT-TCN, the input is 24-second, where it is later divided into four 6-second chunks for the SpecTNT front module (see Fig. 2(c)). For STFT, we use sampling rate of 16000 Hz and a window length of 1024.

We apply random sampling to load a random chunk into a batch, instead of sequentially loading a (6- or 24-second) chunk from the beginning of each training song. That is, we first enumerate all the valid training chunks into a list by using a sliding window (6- or 24-second) with 1-second hop size on every song, and the last chunk of a song shall not exceed the length of the song. Then, the data loader draws a sample uniformly from the list during training. Therefore, a batch can include chunks from different locations of different songs. This technique leads to faster and better convergence.

We use PyTorch 1.8 and the Adam optimizer [28] with 0.001 learning rate and 80% weight decay. The numbers of parameters for TCN, SpecTNT, and SpecTNT-TCN are 86,679, 4,637,392, and 4,692,896, respectively. We train and test the models using 4 Tesla-V100-SXM2-32GB GPUs with batch sizes of 128, 128, and 32 for SpecTNT, TCN, and SpecTNT-TCN, respectively. Each epoch runs 500 mini-batches of training before a validation. Using the aforementioned specifications, TCN, SpecTNT, and SpecTNT-TCN take about 5.8 minutes, 10 minutes, and 6.5 minutes, respectively, to train 500 batches, and 6 seconds, 12.6 seconds, and 15.1 seconds, respectively, to inference 100 30-second audio clips. The best model is selected based on the validation set for testing.

4.2. Experiment Setting

We use nine public datasets to evaluate the proposed methods. Since the beat/downbeat tracking datasets are commonly small in size, we tried to collect as many as possible and combined them to conduct

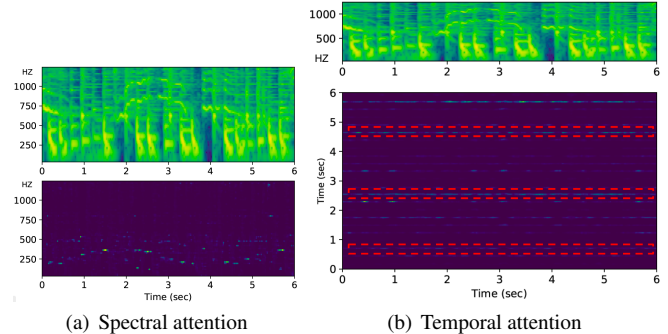


Fig. 3. Visualization of the attention maps for a *GTZAN* example, “pop.00098.wav” from 6 to 12 seconds. Upper subplot displays one of the time-synchronized harmonic representations.

training and evaluation in an 8-fold cross-validation manner following [17]. These datasets include *Beatles* [29], *Ballroom* [30], *SMC* [30], *Hainsworth* [31], *Simac* [32], *HJDB* [33], *RWC-Popular* [34], and *Harmonix Set* [35]. An additional dataset, *GTZAN* [36], is used for testing only.

We note that the *Harmonix Set* is comparably larger than other datasets, containing 912 pop/rock songs with more than 56 hours of beat- and downbeat-annotated audio content. Our copy is slightly different from the original one in [35]. The original audio files corresponding to the annotations are not available, but reference YouTube links are provided. Our investigation indicated there are many errors when aligning the annotations with the YouTube audio, so we searched for the audio and manually adjusted the annotations to ensure the labels and timestamps are sensible and aligned to the audio.

For evaluation, we report the standard metrics: *F1*, *CMLt*, and *AMLt*, with a tolerance window of 70 ms as defined in [29]. The resulting scores are computed using the *mir_eval* package [37].

4.3. Attention Map Visualization

To better understand the attention mechanism in SpecTNT when modeling beats and downbeats, we visualize one of the attention heads of the last SpecTNT block during the inference. For the spectral attention (see Fig. 3(a)), the frequency attention scores that generate the FCT of each frame are obtained. It can be seen that the attention map successfully characterizes the harmonic components. The spectral attention map emphasizes on the bass notes between 200 and 400 Hz which matches the bass line contour in the harmonic representation. For the temporal attention (see Fig. 3(b)), we display the time-by-time self-attention map from one of the 8 attention heads that we consider mostly relevant to the downbeat modeling. The red dashed rectangles highlight the downbeat locations. We can see that almost every frame attends to the downbeat locations, i.e., each column vector (a frame) has higher scores at the downbeat times. This will help generate the output embeddings that emphasize the downbeats, and predict higher probabilities for them accordingly.

4.4. Results and Discussion

The beat and downbeat tracking results are presented in Table 1 and Table 2, respectively. We include the results from prior literature for reference, although they are not necessarily comparable to ours, as the training data and experimental settings are different. The performance scores of *Harmonix Set* for Böck et al. [18] were evaluated

	F1	CMLt	AMLt	F1	CMLt	AMLt
	<i>RWC-POP</i>			<i>Harmonix Set</i>		
Böck et al. [18]	.943	-	-	.933 [†]	.841 [†]	.938 [†]
TCN (baseline)	.947	.922	.952	.946	.895	.942
SpecTNT	.953	.925	.957	.947	.896	.943
SpecTNT-TCN	.950	.925	.958	.953	.939	.959
	<i>SMC</i>			<i>Beatles</i>		
Böck et al. [18]	.516	.406	.575	.918	-	-
Böck et al. [17]	.544	.443	.635	-	-	-
TCN (baseline)	.560	.474	.621	.933	.870	.933
SpecTNT	.602*	.515*	.661	.940	.898	.929
SpecTNT-TCN	.605*	.514*	.663	.943	.896	.938
	<i>Ballroom</i>			<i>Hainsworth</i>		
Davies et al. [15]	.933	.881	.929	.874	.795	.930
Böck et al. [17]	.962	.947	.961	.902	.848	.930
TCN (baseline)	.940	.870	.957	.860	.849	.915
SpecTNT	.927	.856	.939	.866	.865	.914
SpecTNT-TCN	.962*	.939*	.967	.877	.862	.915

Table 1. Beat tracking results using an 8-fold cross validation, where * denotes statistical significance compared to TCN (baseline), and [†] denotes the predictions made by madmom [14].

on the predictions made by the madmom package [14], which uses the method proposed in [18]. *Harmonix Set* was not included in the training set of madmom.

Comparing the results among SpecTNT, TCN, and SpecTNT-TCN in Table 1, we observe that SpecTNT is generally better than TCN, and that SpecTNT-TCN outperforms SpecTNT in most cases. This demonstrates that SpecTNT is a promising DNN architecture for this task. In particular, SpecTNT and SpecTNT-TCN perform significantly better on *SMC*, one of the most challenging datasets in beat tracking. Our qualitative inspections indicate that SpecTNT is superior in handling expressive timing. For instance, it performs better in cases such as “SMC.00111” (with a pause) and “SMC.00165” (with varying tempo). However, it may suffer from non-trivial percussive patterns that yield phase errors (e.g., “SMC.00213”). Other than that, the results also show that SpecTNT seems less successful on *Ballroom* and *Hainsworth*. This could be attributed to the sub-optimal issue mentioned in Section 3.4. Further investigations reveal that SpecTNT tends to over-fit beat annotations of these two datasets early during training, and then the performance drops slightly in the later epochs. This is expected, because *Hainsworth* covers more diverse genres (e.g., classical, folk, and jazz), and *Ballroom* is compiled with old-school dance music. These two datasets provide relatively less examples for their specific genres, as compared to a much larger dataset, *Harmonix set*, which offers fairly more examples of pop-style in the training set. Since our model selection criterion considers the joint performance of beat and downbeat, a model that performs much better in downbeat than in beat is picked eventually (and it requires more epochs). Nevertheless, SpecTNT-TCN is more robust to this issue and shows improvements over SpecTNT and TCN.

From the downbeat tracking results in Table 2, similar observations can also be made. The performance difference between TCN and SpecTNT is more obvious. Such results are in line with our research motivation that SpecTNT can help capture harmonic and timbral information and improve downbeat tracking. The improvement varies across different datasets. Our proposed models perform better than the baseline by a wider margin for non-pop-centric datasets: *Ballroom* and *Hainsworth*. In particular, the performance difference between SpecTNT (or SpecTNT-TCN) and TCN are statistically sig-

	F1	CMLt	AMLt	F1	CMLt	AMLt
	<i>RWC-POP</i>			<i>Harmonix Set</i>		
Böck et al. [18]	.861	-	-	.804 [†]	.747 [†]	.873 [†]
TCN (baseline)	.930	.928	.938	.873	.839	.908
SpecTNT	.941	.929	.957	.897*	.862	.924*
SpecTNT-TCN	.945	.939	.959	.908*	.872*	.928*
	<i>Ballroom</i>			<i>Beatles</i>		
Fuentes et al. [38]	.830	-	-	.860	-	-
Böck et al. [17]	.916	.913	.960	.837	.742	.862
TCN (baseline)	.841	.788	.937	.843	.767	.851
SpecTNT	.884*	.835*	.937	.867	.810*	.860
SpecTNT-TCN	.937*	.927*	.968	.870	.812*	.865
	<i>Hainsworth</i>					
Böck et al. [17]	.722	.696	.872			
TCN (baseline)	.682	.683	.852			
SpecTNT	.729*	.734*	.879			
SpecTNT-TCN	.748*	.738*	.870			

Table 2. Downbeat tracking results using an 8-fold cross validation, where * denotes statistical significance compared to TCN (baseline), and [†] denotes the predictions made by madmom [14].

	Beat			Downbeat		
	F1	CMLt	AMLt	F1	CMLt	AMLt
Böck et al. [17]	.885	.813	.931	.672	.640	.832
Transformer	.853	.741	.887	.667	.617	.843
TCN (baseline)	.879	.802	.911	.702	.660	.859
SpecTNT	.883	.809	.906	.745*	.696*	.875*
SpecTNT-TCN	.887	.812	.920	.756*	.715*	.881*

Table 3. Beat and downbeat tracking result on (test-only) *GTZAN*, where * denotes statistical significance compared to TCN (baseline).

nificant in terms of *F1* and *CMLt*. As mentioned before, SpecTNT may sacrifice its beat tracking performance for downbeat, as the performance gain of downbeat is larger. Qualitatively, we also found SpecTNT can avoid quite a few phase errors as compared to TCN.

Lastly, we present the results of *GTZAN* in Table 3. In addition to the models presented above, we also include the results from a regular Transformer [6], which contains only the temporal encoder [4]. As it is non-hierarchical and without the spectral encoder, we treat the comparison as the ablation study. It is clear that the regular Transformer does not work well in this case, most likely because the training data is insufficient. Whereas SpecTNT’s design can handle this well by leveraging spectral encoders in a stacked architecture. Moreover, we once again see similar results to previous experiments when comparing among SpecTNT, SpecTNT-TCN, and TCN.

5. CONCLUSION

In this paper, we present a novel Transformer-based model, SpecTNT, to model beats and downbeats from audio signals. The design of spectral and temporal encoders enables a promising DNN approach, showing SOTA results in downbeats tracking.

For future work, we will try to mitigate the sub-optimal issue (see Section 4.4) with better solutions, as we expect it is currently the limitation for SpecTNT to improve beat tracking. We also plan to study the possibility to include additionally relevant MIR tasks (e.g., music structure [39], chords, and melody) into a unified multi-task learning framework and model them jointly with SpecTNT.

6. REFERENCES

- [1] S. Böck and M. Schedl, “Enhanced beat tracking with context-aware neural networks,” in *Proc. DAFx*, 2011.
- [2] S. Durand, J. P. Bello, B. David, and G. Richard, “Robust downbeat tracking using an ensemble of convolutional networks,” *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 25, no. 1, 2016.
- [3] Ashish Vaswani et al., “Attention is all you need,” in *Proc. NIPS*, 2017, pp. 5998–6008.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL*, 2019.
- [5] M. Won, S. Chun, and X. Serra, “Toward interpretable music tagging with self-attention,” *arXiv preprint arXiv:1906.04972*, 2019.
- [6] M. Won, K. Choi, and X. Serra, “Semi-supervised music tagging transformer,” *Proc. ISMIR*, 2021.
- [7] T.-P. Chen and L. Su, “Harmony Transformer: Incorporating chord segmentation into harmony recognition,” in *Proc. ISMIR*, 2019, pp. 259–267.
- [8] J. Park, K. Choi, S. Jeon, D. Kim, and J. Park, “A bi-directional transformer for musical chord recognition,” in *Proc. ISMIR*, 2019.
- [9] W.-T. Lu, J.-C. Wang, M. Won, K. Choi, and X. Song, “SpecTNT: A time-frequency transformer for music audio,” in *Proc. ISMIR*, 2021.
- [10] S. Dixon, “Automatic extraction of tempo and beat from expressive performances,” *J. NMR*, vol. 30, no. 1, 2001.
- [11] A. P. Klapuri, A. J. Eronen, and J. T. Astola, “Analysis of the meter of acoustic musical signals,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 14, no. 1, 2005.
- [12] M. Goto, “An audio-based real-time beat tracking system for music with or without drum-sounds,” *J. New Music Research*, vol. 30, no. 2, 2001.
- [13] M. E. Davies and M. D. Plumbley, “Context-dependent beat tracking of musical audio,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, 2007.
- [14] S. Böck et al., “Madmom: A new python audio and music signal processing library,” in *Proc. ACM Multimedia*, 2016, pp. 1174–1178.
- [15] E. M. Davies and S. Böck, “Temporal convolutional networks for musical audio beat tracking,” in *Proc. EUSIPCO*, 2019.
- [16] S. Böck, M. EP Davies, and P. Knees, “Multi-task learning of tempo and beat: Learning one to improve the other,” in *Proc. ISMIR*, 2019.
- [17] S. Böck and M. EP Davies, “Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation,” in *Proc. ISMIR*, 2020.
- [18] S. Böck, F. Krebs, and G. Widmer, “Joint beat and downbeat tracking with recurrent neural networks,” in *Proc. ISMIR*, 2016.
- [19] M. Fuentes, B. Mcfee, H. C. Crayencour, S. Essid, and J. P. Bello, “A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning,” in *Proc. ICASSP*, 2019.
- [20] F. Krebs, S. Böck, and G. Widmer, “Rhythmic pattern modeling for beat and downbeat tracking in musical audio,” in *Proc. ISMIR*, 2013.
- [21] T. Jehan, “Downbeat prediction by listening and learning,” in *Proc. IEEE WASPAA*, 2005.
- [22] H. Papadopoulos and G. Peeters, “Joint estimation of chords and downbeats from an audio signal,” *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 1, 2010.
- [23] M. E. Davies and M. D. Plumbley, “A spectral difference approach to extracting downbeats in musical audio,” in *Proc. EUSIPCO*, 2006.
- [24] M. Won, S. Chun, O. Nieto, and X. Serra, “Data-driven harmonic filters for audio representation learning,” in *Proc. ICASSP*, 2020, pp. 536–540.
- [25] S. Böck, F. Krebs, and G. Widmer, “A multi-model approach to beat tracking considering heterogeneous music styles,” in *Proc. ISMIR*, 2014.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, “Identity mappings in deep residual networks,” in *Proc. ECCV*. Springer, 2016.
- [27] M. E. P. Davies, S. Böck, and M. Fuentes, *Tempo, Beat and Downbeat Estimation*, <https://tempobeatdownbeat.github.io/tutorial/intro.html>, 2021.
- [28] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *Proc. ICLR*, 2015.
- [29] M. E. Davies, N. Degara, and M. D. Plumbley, “Evaluation methods for musical audio beat tracking algorithms,” *Queen Mary University of London, Centre for Digital Music*, 2009.
- [30] F. Gouyon et al., *A computational approach to rhythm description-Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*, Ph.D. dissertation, Universitat Pompeu Fabra, 2005.
- [31] S. W. Hainsworth and M. Macleod, “Particle filtering applied to musical tempo tracking,” *EURASIP J. Advances in Signal Process.*, vol. 2004, no. 15, 2004.
- [32] A. Holzapfel, M. EP Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, “Selective sampling for beat tracking evaluation,” *IEEE Trans. ASLP*, vol. 20, no. 9, pp. 2539–2548, 2012.
- [33] J. Hockman, M. EP Davies, and I. Fujinaga, “One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass,” in *Proc. ISMIR*, 2012, pp. 169–174.
- [34] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, “RWC music database: Popular, classical and jazz music databases,” in *Proc. ISMIR*, 2002, vol. 2.
- [35] O. Nieto et al., “The Harmonix Set: Beats, downbeats, and functional segment annotations of western popular music,” in *Proc. ISMIR*, 2019, pp. 565–572.
- [36] U. Marchand and G. Peeters, “Swing ratio estimation,” in *Proc. DAFx*, Trondheim, Norway, 2015.
- [37] C. Raffel, B. McFee, E. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. PW Ellis, “mir_eval: A transparent implementation of common mir metrics,” in *Proc. ISMIR*, 2014.
- [38] 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 *Proc. ISMIR*, 2018.
- [39] J.-C. Wang, Y.-N. Hung, and B. L. J. Smith, “To catch a chorus, verse, intro, or anything else: Analyzing a song with structural functions,” in *Proc. ICASSP*, 2022.