DECONSTRUCT, ANALYSE, RECONSTRUCT: HOW TO IMPROVE TEMPO, BEAT, AND DOWNBEAT ESTIMATION

Sebastian Böck
enliteAI, Vienna, Austria
s.boeck@enlite.ai

Matthew E. P. Davies
University of Coimbra, CISUC, DEI
mepdavies@dei.uc.pt

ABSTRACT

In this paper, we undertake a critical assessment of a state-of-the-art deep neural network approach for computational rhythm analysis. Our methodology is to deconstruct this approach, analyse its constituent parts, and then reconstruct it. To this end, we devise a novel multi-task approach for the simultaneous estimation of tempo, beat, and downbeat. In particular, we seek to embed more explicit musical knowledge into the design decisions in building the network. We additionally reflect this outlook when training the network, and include a simple data augmentation strategy to increase the network's exposure to a wider range of tempi, and hence beat and downbeat information. Via an in-depth comparative evaluation, we present state-of-the-art results over all three tasks, with performance increases of up to 6% points over existing systems.

1. INTRODUCTION

A central concept in much of the work on audio beat tracking is the "tactus" – described as the most comfortable foot-tapping rate when unconsciously tapping to a piece of music. As stipulated by London [1, Ch.1] (and references therein), the tactus is essential for our perception of metre. The tactus by itself carries no information concerning the metrical organisation within a piece of music, but it is informative about both local and global tempo. To perceive metre, we require the hierarchical organisation between at least two levels, and ideally three: a level above the tactus which indicates the longer-term grouping of beats into bars (or measures), and a lower level to describe how the beats are sub-divided – whether in *simple* time (divided by two), or *compound* time (divided by three).

In this sense, we can expand the notion of (unmarked) foot-tapping towards "counting" in time to music. While numerous counting systems exist for the teaching of musical rhythm [2], the "traditional" American system is perhaps the most well-known. For two-level counting, we can mark the three beats of the bar of a waltz as follows: **1** 2 3 **1** 2 3 **1**..., where the **1** indicates the first beat of each bar, the downbeat. Moving to three-level counting, we can count the sub-divisions of a four beat bar into two as: **1** + 2 + 3 + 4 + **1**..., (*one - and - two - and - three - and - four - and*), and the sub-divisions of the same four beat bar into four as: **1** e + a 2 e + a 3 e + a 4 e + a **1**... (*one - "ee" - and - "ah" - two - "ee" - and -"ah"* and so on).

From the perspective of computational rhythm analysis, we can thus make a distinction between approaches which target one metrical level in isolation, as opposed to those which estimate more than one. Among the single-level approaches, the vast most majority fall within the domain of beat-tracking (e.g [3–6]). When the focus of the analysis moves towards downbeats, this almost exclusively relies on the implicit or explicit modelling of another metrical level, either the beat [7], tatum [8], or a contrast between both [9]. One notable outlier is the downbeat prediction approach of Jehan [10] which relies instead on onset-synchronous analysis.

Concerning the modelling of three simultaneous metrical levels, few published approaches exist. Goto [11] presents a real-time system for estimating the quarter-note, half-note, and measure levels, but doesn't address the sub-beat level. Klapuri et al. [12] on the other hand, address the estimation of tatum, beat, and downbeat, with explicit dependencies between the phase of the beat and the tatum, and the period of the beat and downbeat. For a recent review of beat and downbeat estimation, see [13].

Considering the topic of tempo estimation, which, in most instances, seeks to retrieve a single value to describe a global tempo, existing approaches can be split into two categories: those which make their estimate of the tempo based the post-processing of a sequence of beat times (e.g. for a discussion of techniques, see [14]) and those which treat the task as a classification or regression problem and do not require any prior estimate of the beats [15–18].

In this paper, we seek to work from the perspective of leveraging shared connections in musical structure, and address the simultaneous estimation of three highly interconnected properties of musical rhythm: tempo, beat, and downbeat within a single model. In line with much of the recent literature concerning the extraction of musical information from audio signals [19], we adopt a deep learning approach. We depart from our recent multi-task approach [20] for tempo and beat estimation using a temporal convolutional network (TCN), which was shown to provide state-of-the-art results. We undertake a critical assessment of its constituent parts, and on the basis of our

Attribution: Sebastian Böck, Matthew E. P. Davies, "Deconstruct, Analyse, Reconstruct: How to improve Tempo, Beat, and Downbeat Estimation", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.,* Montréal, Canada, 2020.

analysis, adapt it in several ways. At the broadest level, we wish to leverage the benefit of modelling a metrical hierarchy (as opposed to just the beat level) by the inclusion of an additional learning task, downbeat estimation. In terms of the structure of the network itself, we adapt the shallowest layers of the network (i.e. those closest to the musical surface) to provide a better model of harmonic musical sounds. In addition, we propose a novel formulation of the TCN architecture which incorporates an additional dilation rate to each layer as a means to embed understanding of integer ratios modelling the metrical structure.

A peculiar aspect of the evaluation in [20] was the ability of the multi-task model to perfectly estimate the tempo of the HJDB dataset [21] when it was included in the training splits, with good, but noticeably lower performance when it was left as a hold-out test set. Given the characteristic fast tempo of *HJDB*, we speculate that the gap in performance arose due to the lack of any similarly fast-tempo music in the training sets. Following this argument, a secondary motivation of this work is to consider how data augmentation can be used in an efficient way to extrapolate information from regions of the training data which are well-covered in terms of tempo annotations to those which are more sparse.

Via a thorough evaluation across the three tasks of tempo, beat, and downbeat estimation, we demonstrate state-of-the-art performance, and draw attention to the ability of TCN-based approaches to leverage shared representations for multi-task analysis of musical audio signals.

The remainder of this paper is structured as follows. In Section 2 we describe our multi-task formulation and data augmentation strategy in detail. In Section 3 we present an ablation study and comparative evaluation against existing reference systems. Finally, in Section 4 we discuss the impact of the contribution and promising areas for future work.

2. APPROACH

Our earlier multi-task approach for tempo and beat estimation [20] was itself an extension of an earlier TCN-based approach for beat tracking [22]. The core component, which is common to both, is a deep neural network (DNN) architecture based on dilated convolutions, most well-known from *WaveNet* [23]. It is quite striking to consider that an architecture designed for the causal generation of raw audio (primarily for speech synthesis), and with its roots in an auto-regressive process, can find application in a problem cast as binary classification through time, i.e. the classification per frame of the presence or absence of a beat. From an alternative perspective, we may view the strength of the TCN in this problem domain as resulting from multiple connections (both forwards and backwards in time) at different time scales, and thus bearing similarity to much earlier work on the cognitively-inspired use of multi-resolution signal processing for beat tracking [24].

For a detailed description of the existing architecture, we refer to the reader to [20, 22]. In brief, the multi-task approach uses a log-magnitude spectrogram with 81 logarithmically spaced frequency bins and a frame rate of 100 frames per second as input. Overlapping spectrogram snippets are passed through 3 convolution and max pooling layers, followed by 11 dilated convolutional layers whose dilation rate increases by a factor of 2 per layer. The so-called “skip connections” between these layers are provided as an auxiliary output of the TCN and are used to generate a prediction of the tempo across a linear range from $0 - 300$ beats per minute (bpm). The main output of the TCN, a beat activation function, is then processed by a dynamic Bayesian network (DBN) [25] to obtain a final sequence of beat estimates.

In spite of the reported high performance of the multi-task approach on a wide range of musical material for both beat and tempo estimation, we believe it is valuable to question the design decisions of this network and consider the ways in which it could be modified to improve performance. Our focus in this paper is on the core of the network, namely the convolutional and max pooling layers together with the TCN. In a coarse sense, we can consider the convolutional and max pooling layers to relate to more surface-level properties of the music and hence local information, i.e. *what are the spectro-temporal properties of the beats?* with the deeper TCN layers oriented more towards their temporal dependency over longer time scales, i.e. *how is the beat and metrical structure organised over the duration of musical pieces?*

Concerning the first question, a common limitation of beat tracking systems is their ability to reliably detect the beat in music without the presence of drums, as typified, at least in part, by lower reported performance in classical music. Given the high prevalence of rock, pop, jazz, and electronic dance music among existing beat tracking datasets [26], we consider the modelling of harmonic sounds to be important when addressing under-represented musical styles and of crucial importance to reliably detect downbeats in Western music, where harmonic changes often occur at bar boundaries [27]. Regarding the second question, we directly enable the network to learn feature representations which are integer multiplies of each other by deploying multiple concurrent dilated convolutions at each TCN layer, and in so doing embed some implicit hierarchical structure into the model.

In what follows, we describe the specific modifications made to the network. Since the work in this paper explicitly targets the improvement of an existing approach, we allude to performance increases wherever relevant, with detailed results in Section 3.

2.1 Multi-task formulation

Based on the model described above, we add the additional task of downbeat tracking. This can be accomplished in various ways. One option is to model the downbeats and beats jointly as a multi-class problem, i.e. by classifying each input frame to be a beat, a downbeat, or neither. This approach was successfully deployed in [28], but has the downside that it cannot fully leverage the information if a dataset contains only beat or downbeat annotations. Thus we treat the problem as a multi-label classification problem instead, with the downbeat task treated as a separate binary classification problem with its own output. We model the downbeat output similarly to the beat output as a single *sigmoid* unit which is fed directly from the main TCN output. Whereas the approach in [20] used 16 filters per layer for the multi-task estimation of tempo and beat, with the addition of the downbeat task, we expand the network to include more filters and increase this to 20.

This additional output is then also post-processed with a DBN. Since the beat and downbeat outputs do not define a joint probability density function (i.e. their sum is not guaranteed to be 1 as for multi-class problems), the DBN post-processing used in [28] cannot be applied directly to the combined beat and downbeat activations. Thus the difference of the beat and downbeat activations (limited to positive values) and the downbeat activations are used as state-conditional observations for beats and downbeats, respectively. In Section 3 we refer to this approach as *joint downbeat tracking*.

An alternative approach is to first detect the beats and then in a second inference step to find the downbeats given the set of beat predictions. This approach was chosen in [29] and has the most notable advantage that the large joint state space which is required to model multiple bar lengths and tempi at a frame level resolution can be split into two smaller ones. The first one (tracking the beats) only requires multiple tempi to be modelled at the frame level resolution, whereas the second one operates at beat resolution and is completely tempo invariant, thus requiring only very few states. The downside of this approach is that errors made in beat tracking directly propagate to downbeat tracking. In Section 3 we refer to this approach as *sequential downbeat tracking*.

2.2 Conv layers

Both the original beat tracking paper [22] and the multi-task extension tackling global tempo estimation presented in [20] use the same convolutional block to reduce multiple consecutive STFT frames to a one-dimensional feature vector which is then processed by the TCN. Two groups of alternating 3×3 convolution and 1×3 max-pooling layers were used to reduce overlapping spectrogram windows of size 5×81 (time $\times$ frequency) down to 3×26 and 1×8 before these eight bands (roughly representing one octave each) were combined into a singular value with a 1×8 convolution. This feature representation is closely related to the one used in [12] and was shown to work well.

However, a musically motivated reordering of these layers can have a positive effect on the performance of the model. Convolutional filters covering multiple frequency bins but only a single time step have been shown to concentrate on harmonic and timbral features [30] and proven to work well for multiple tasks, including key estimation [31] and automatic music transcription [32]. Moving the "frequency only" convolution in between the two 3×3 convolutions as shown in Figure 1, enables the network to better capture harmonic content across a wider frequency range instead of detecting local changes in smaller regions of the spectrogram only and then later aggregating them.

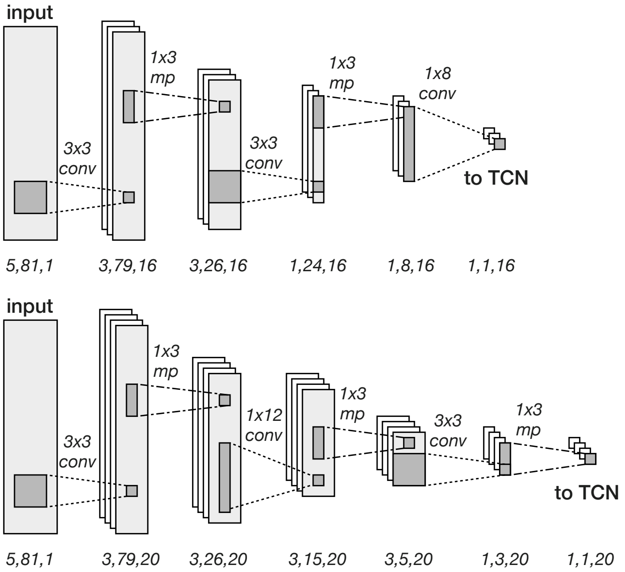

Figure 1: Comparison of the convolution (*conv.*) and max pooling (*mp*) layers. The architecture from [20, 22] (top). Our proposed architecture (bottom). The dimensions of the tensors are shown below each layer.

2.3 TCN layers

From a musical perspective, it is undeniable that discovering downbeats requires more knowledge about the signal than locating beat positions only. Independently of whether this additional knowledge is harmonic or rhythmic in nature, it always requires a longer temporal context. Increasing the temporal context of the TCN by either using larger kernel sizes or adding more layers (with exponentially increasing dilation rates), did not improve any of the tasks under investigation. This observation is not necessarily surprising since the temporal context modelled by the TCN is already about 40 seconds – which should be sufficient to tackle the task of tracking the locations of the downbeats and estimating the length of the bars. Instead, adding a second dilated convolution (with a doubled dilation rate) to each of the TCN layers enables the network to simultaneously model musical properties at various levels which are integer multiples of each other. We discovered that adding a third dilation rate did not further improve performance, but we believe this is very likely an artefact of the data utilised for training, since none of the datasets used have a noticeable number of musical pieces with compound time signatures. The feature maps of the two dilated convolutions are concatenated before spatial dropout [33] and an exponential linear unit (ELU) activation function [34] is applied. In order to keep the output dimensionality of the TCN layer constant, these feature maps are then combined by a 1×1 convolution, which increases the total number of parameters linearly with each TCN layer instead of exponentially.

2.4 Data augmentation

Our approach to data augmentation is both simple and straightforward and similar to the scaling approach applied in [17]. Contrary to other data augmentation strategies, which pre-process the audio signal and manipulate it in various ways (e.g. time stretching, pitch shifting, sample rate conversion to simulate speeding up or slowing down the signal [35, 36]), we do not change the audio signal itself, but instead only change the parameters of the STFT when obtaining a time-frequency representation. To be more precise, we only change the rate at which the overlapping frames of the STFT are obtained from the audio signal by sampling from a normal distribution with 5% standard deviation from the annotated tempo. By changing only the hop size, we obtain spectrograms with varying overlap factors and only the targets have to be adjusted accordingly. Using this data augmentation strategy leads to many more training examples for tempi which are otherwise underrepresented in the data, as can be seen in Figure 2.

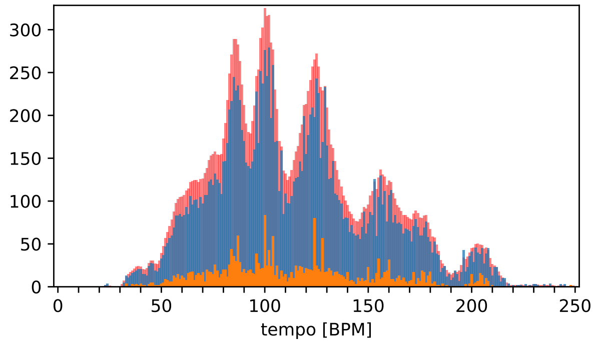

Figure 2: Tempo distribution of original tempo annotations (orange, foreground), after data augmentation (blue) and target widening (red, background).

2.5 Network training

Using data augmentation increases the amount of data the network can learn from. However, this also leads to increased training times when using conventional training procedures. Furthermore, the additional downbeat classification layer, the inclusion of a second dilated convolution and the usage of more filters in each of the TCN layers has a notable impact on the size of the model, which now has $116,302$ trainable parameters compared to $29,901$ of [20].

To this end, we make use of the latest training optimisation strategies, namely *RAdam* [37] and *Lookahead Optimization* [38]. The combination of these two drastically reduces the training time (even accounting for the larger number of weights) simultaneously leading to models being less sensitive to different random initialisations. All remaining hyper-parameters were left unchanged. We found the used learning rate of $2e^{-3}$ and clipping the gradients at a norm of 0.5 a sensible choice, as is training on full sequences with a batch size of 1.

We derive the tempo targets in the same way by computing a smoothed and interpolated histogram on the inter beat intervals. We apply the same target widening strategy to present the network not only the annotated frame and tempo, but also their direct neighbouring frames and ±2 BPM values as positive targets, albeit with lower weights of 0.5 and 0.25, respectively.

3. EXPERIMENTS AND RESULTS

We use the same datasets as in [20] with the most recent annotations available. *Beatles* [39], *Cuidado* [40], *Hainsworth* [20, 41], *Simac* [42], *SMC* [26], and *HJDB* [21, 28] are used for training and evaluated in an 8-fold cross validation manner. *ACM Mirum* [43, 44], *GiantSteps* [45, 46], and *GTZAN* [47, 48] are used as test datasets. Predictions for the test datasets are obtained by averaging the predictions of the networks trained for cross validation. To enable future comparisons, we make all annotations as well as the beat, downbeat, and tempo estimates available at the accompanying website. [1]

For evaluation, we use the standard metrics used in the literature. For tempo estimation, we report *Accuracy 1* and *Accuracy 2* scores with a tolerance of $\pm 4\%$ as used in [49]. For beat and downbeat tracking, we use *F-measure* and the continuity based metrics *CMLt* (requires beats being tracked at the annotated metrical level) and *AMLt* (allowing alternative metrical levels, such as double/half and triple/third tempo as well as off-beat) with the tolerances as defined as in [39].

3.1 Ablation study

Before reporting comparative evaluation to other methods, we aim to understand how each of the proposed measures outlined in Section 2 contribute to the final performance of the system.

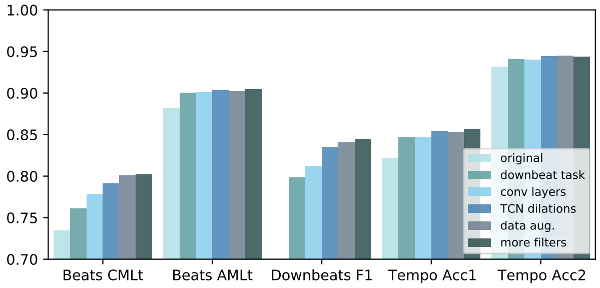

Figure 3: Impact of the improvements proposed for selected evaluation metrics. Mean values over the complete validation set are given.

From Figure 3 it can be seen that the measures undertaken to improve the original system do not contribute the same to the different tasks and the given evaluation metrics. For example, beat tracking *AMLt* and tempo *Accuracy 2* scores increase only marginally, which is best explained by the fact that the baseline system is already performing at a high level on these tasks. However, since these metrics allow metrical ambiguities, it is impossible

[1] https://github.com/superbock/ISMIR2020

to determine if the system is considering the correct metrical level in the case of beat tracking or the correct tempo octave. Both *CMLt* and *Accuracy 1* require the reported beat locations and tempo to exactly match the annotations (within the allowed tolerance). These metrics therefore better catch the ability of an algorithm to correctly predict the annotated information.

Concentrating on these metrics, it can be seen that additionally modelling and predicting downbeats has a positive effect on beat tracking and tempo estimation. This effect is then strengthened by the modifications made to the convolutional and TCN layers. Using data augmentation and more filters gives a small additional boost. It should be noted that the positive effect of data augmentation on the generalisation capabilities of the network are mostly visible for the task of tempo estimation if "out of tempo distribution" datasets are used for evaluation. Since the validation set is a randomly chosen subset of the training set (and hence has a very similar tempo distribution), the impact is not fully reflected in Figure 3.

3.2 Tempo estimation

Tempo estimation is the task with the most noticeable overall impact of the proposed refinements. While *Accuracy 2* values have been quite high for many systems among all datasets under consideration, the new system is the only one consistently achieving high *Accuracy 1* values as well (Table 1). The system's ability to model several tasks simultaneously and exploit mutual information relevant to all tasks leads to an increased performance of more than 6% points in *Accuracy 1* over the best results reported so far on certain datasets.

	Accuracy 1	*Accuracy 2*
ACM Mirum		
Gkiokas et al. [50]	0.725	0.979
Percival and Tzanetakis [44]	0.733	0.972
Schreiber and Müller [17]	0.781	0.976
Böck et al. [20]	0.749	0.974
Foroughmand & Peeters [18]	0.733	0.965
Ours	0.841	0.990
GiantSteps		
Gkiokas et al. [50]	0.721	0.922
Percival and Tzanetakis [44]	0.506	0.956
Schreiber and Müller [17] *	0.821	0.971
Böck et al. [20]	0.764	0.958
Foroughmand & Peeters [18] *	0.836	0.979
Ours	0.870	0.965
GTZAN		
Gkiokas et al. [50]	0.651	0.931
Percival and Tzanetakis [44]	0.658	0.924
Schreiber and Müller [17]	0.769	0.926
Böck et al. [20]	0.673	0.938
Foroughmand & Peeters [18]	0.697	0.891
Ours	0.830	0.950

Table 1: Tempo estimation results on unseen test data. Asterisks denote systems which have been trained on a disjoint set of the same source.

3.3 Beat tracking

Although beat tracking performance of existing systems is already very high, the new system sets new high scores in *CMLt* and even exceeds the very high performance values above 0.9 (on *Ballroom*) by more than 4% points. Other systems achieve such high scores only under the less strict *AMLt* metric, which also permits metrical errors, including double/half, triple/third tempo, and off-beat. This highlights the capability of the system to track beats exactly at the annotated metrical level.

	F-measure	*CMLt*	*AMLt*
Ballroom			
Böck et al. [28]	0.938	0.892	0.953
Elowsson [51] ‡	0.925	0.903	0.932
Davies and Böck [22]	0.933	0.881	0.929
Ours (beat tracking)	0.956	0.935	0.958
Ours (joint tracking)	0.962	0.947	0.961
Hainsworth			
Böck et al. [5]	0.884	0.808	0.916
Elowsson [51] ‡	0.742	0.676	0.792
Davies and Böck [22]	0.874	0.795	0.930
Ours (beat tracking)	0.904	0.851	0.937
Ours (joint tracking)	0.902	0.848	0.930
SMC			
Böck et al. [5]	0.529	0.428	0.567
Elowsson [51] ‡	0.375	0.225	0.332
Davies and Böck [22]	0.543	0.432	0.632
Ours (beat tracking)	0.552	0.465	0.643
Ours (joint tracking)	0.544	0.443	0.635
GTZAN			
Böck et al. [5]	0.864	0.768	0.927
Davies and Böck [22]	0.843	0.715	0.914
Ours (beat tracking)	0.883	0.808	0.930
Ours (joint tracking)	0.885	0.813	0.931

Table 2: Beat tracking results on datasets used for training with 8-fold cross validation (top), and on unseen test data (bottom). ‡ was trained on *Ballroom* data only.

In Table 2 it can also be seen that joint modelling of beats and downbeats (in the DBN) can be beneficial for music with constant meter and steady tempo (e.g. *Ballroom*), whereas it negatively impacts performance for expressive music as contained in *Hainsworth* and *SMC*.

3.4 Downbeat tracking

For the task of downbeat tracking the systems, performance can be clearly separated into two main categories: i) the systems of Durand et al. [8] and Fuentes et al. [9], which explicitly model harmonic features (using chroma features as input for the neural network) and ii) the ones of Böck et al. [28] and ours which learn harmonic features implicitly. Whereas the former show better performance on pop music (e.g. the *Beatles* dataset) where downbeats often coincide with harmonic changes, they perform less well on data where bars are mostly defined based on rhythm.

	F-measure	*CMLt*	*AMLt*
Ballroom			
Böck et al. [28]	0.863	0.834	0.931
Durand et al. [8]	0.797	0.616	0.916
Fuentes et al. [9]	0.83	-	-
Ours (sequential tracking)	0.900	0.894	0.953
Ours (joint tracking)	0.916	0.913	0.960
Hainsworth			
Böck et al. [28]	0.684	0.628	0.832
Durand et al. [8]	0.664	0.500	0.804
Fuentes et al. [9]	0.67	-	-
Ours (sequential tracking)	0.713	0.686	0.855
Ours (joint tracking)	0.722	0.696	0.872
Beatles			
Böck et al. [28]	0.831	0.730	0.858
Durand et al. [8]	0.847	0.722	0.875
Fuentes et al. [9]	0.86	-	-
Ours (sequential tracking)	0.829	0.748	0.860
Ours (joint tracking)	0.837	0.742	0.862
GTZAN			
Böck et al. [28]	0.640	0.577	0.824
Durand et al. [8]	0.607	0.480	0.774
Ours (sequential tracking)	0.654	0.619	0.817
Ours (joint tracking)	0.672	0.640	0.832

Table 3: Downbeat tracking results on datasets used for training with 8-fold cross validation (top), and on unseen test data (bottom).

Regarding the question of whether *joint downbeat tracking* or *sequential downbeat tracking* is superior, Table 3 shows a consistent advantage for processing beats and downbeats simultaneously. The only exception is the *Beatles* dataset, which contains some music with changing metre. Due to memory constraints, joint downbeat tracking cannot model these metre changes. Modelling them is computationally only feasible with sequential downbeat tracking, which may further benefit from sub-beat modelling, as used in [9].

4. DISCUSSION AND CONCLUSIONS

In this paper we address the multi-task estimation of three inter-related properties of musical metre: tempo, beat, and downbeat. Our approach is somewhat unconventional as we do not propose a new method from scratch, but instead we deconstruct, analyse, and then reconstruct an existing approach as a means to further the state of the art. By pairing our methodology with an ablation study, we are able to directly observe the impact of the implemented changes, and in turn, to observe the cumulative gains in performance. Via our evaluation, it is clear that there is no "magic bullet" among our proposed modifications, yet their combination is clearly effective. Furthermore, we must accept that when the baseline performance is already high, the margin for improvement is somewhat limited.

By close inspection of the performance of our approach in comparison both to the baseline and other existing systems, we consider the main impact of our approach as constituting a "closing of the gap" between stricter and more lenient evaluation metrics across each of the tasks. For tempo estimation, our approach is the first to exceed 0.83 for *Accuracy 1* across three large reference datasets, which are completely unseen to our training scheme. Likewise, when considering the positive impact for beat tracking, we find the clearest improvements in the evaluation metric which enforces tracking at the annotated metrical level. Since the relative improvements under the more lenient metrics are much smaller, we do not believe that our approach has unlocked the means to accurately infer the tempo, beat, or downbeat in extremely challenging musical examples. Reference to the incremental improvements for the *SMC* dataset for beat tracking can immediately attest to this. Indeed, the lack of improvement for this kind of musical material may require the reformulation of the inference techniques used to recover the final outputs, rather than intervention at the point of training the networks. Alternatively, they may require a fundamentally different way in which to present targets to the network which is better able to model temporal uncertainty in the annotations. We consider both of these to be promising areas for future work in order to address more challenging data in a robust way.

Ultimately, we believe the main contribution of our work rests in the increased reliability of the good predictions made by the model across these three tasks. It is well-established within music cognition that the perception of tempo, beat, and metre is ambiguous and varies among listeners; therefore within the MIR community, it is easy to justify the use of "multiple-choice" evaluation methodologies. However, this evaluation practice explicitly masks the fact that for almost any piece of music, at least some of these allowed options will be much less reasonable than others. Thus, in the absence of a multi-level annotation methodology in which the set of allowed annotations are specific to individual pieces of music, the only way to guarantee a high-quality prediction (in an unsupervised way) is to aim to maximise performance under stricter evaluation metrics. The alternative is to perform a subjective assessment of beat and downbeat performance via listening to clicks mixed with the audio signals. Given the large amount of musical material in existing datasets, this remains a daunting prospect. However, by restricting this kind of supervised analysis to the subset of excerpts which are accurate only when allowing for alternative interpretations of the annotations, we may move towards a closer estimate of the true performance of these systems. In addition, this kind of partial subjective evaluation could act as a means to "bootstrap" the specification of alternative hypotheses on a per-excerpt basis.

5. ACKNOWLEDGMENTS

This work is funded by national funds through the FCT - Foundation for Science and Technology, I.P., within the scope of the project CISUC - UID/CEC/00326/2020 and by European Social Fund, through the Regional Operational Program Centro 2020, as well as by Portuguese National Funds through the FCT - Foundation for Science and Technology, I.P., under the project IF/01566/2015.

We wish to thank Kazuyoshi Yoshii and Leigh M. Smith for inspiring discussions which helped shape this paper.

6. REFERENCES

[1] J. London, *Hearing in Time: Psychological Aspects of Musical Meter.* Oxford University Press, 2004.

[2] S. L. Gage, "An analysis and comparison of rhythm instructional materials and techniques for beginning instrumental music students," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1994.

[3] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[4] B. McFee and D. P. W. Ellis, "Better beat tracking through robust onset aggregation," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2014, pp. 2154–2158.

[5] S. Böck, F. Krebs, and G. Widmer, "A multi-model approach to beat tracking considering heterogeneous music styles." in *Proc. of the 15th Intl. Society for Music Information Retrieval Conf.*, 2014, pp. 603–608.

[6] T. Cheng, S. Fukayama, and M. Goto, "Convolving Gaussian Kernels for RNN-Based Beat Tracking," in *Proc. of the 26th European Signal Processing Conf.*, 2018, pp. 1919–1923.

[7] 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, pp. 1754–1769, 2011.

[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, 2017.

[9] M. Fuentes, B. McFee, H. C. Crayencour, S. Essid, and J. P. Bello, "Analysis of common design choices in deep learning systems for downbeat tracking," in *Proc. of the 19th Intl. Society for Music Information Retrieval Conf.*, 2018, pp. 106–112.

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

[11] 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.

[12] 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.

[13] M. Fuentes, "Multi-Scale Computational Rhythm Analysis: A Framework for Sections, Downbeats, Beats, and Microtiming," Ph.D. dissertation, Université Paris-Saclay, 2019.

[14] H. Schreiber, "Data-driven approaches for tempo and key estimation of music recordings," Ph.D. dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), 2020.

[15] A. J. Eronen and A. P. Klapuri, "Music tempo estimation with k-nn regression," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 1, pp. 50–57, 2009.

[16] H. Schreiber and M. Müller, "A post-processing procedure for improving music tempo estimates using supervised learning." in *Proc. of the 18th Intl. Society for Music Information Retrieval Conf.*, 2017, pp. 235–242.

[17] ——, "A single-step approach to musical tempo estimation using a convolutional neural network," in *Proc. of the 19th Intl. Society for Music Information Retrieval Conf.*, 2018, pp. 100–105.

[18] H. Foroughmand and G. Peeters, "Deep-rhythm for global tempo estimation in music," in *Proc. of the 20th Intl. Society for Music Information Retrieval Conf.*, 2019, pp. 636–643.

[19] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 206–219, 2019.

[20] S. Böck, M. E. P. Davies, and P. Knees, "Multi-task learning of tempo and beat: Learning one to improve the other," in *Proc. of the 20th Intl. Society for Music Information Retrieval Conf.*, 2019, pp. 486–493.

[21] J. Hockman, M. E. P. Davies, and I. Fujinaga, "One in the Jungle: Downbeat detection in Hardcore, Jungle, and Drum and Bass," in *Proc. of the 13th Intl. Society for Music Information Retrieval Conf.*, 2012, pp. 169–174.

[22] M. E. P. Davies and S. Böck, "Temporal convolutional networks for musical audio beat tracking," in *Proc. of the 27th European Signal Processing Conf.*, 2019.

[23] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR abs/1609.03499*, 2016.

[24] L. M. Smith, "A multiresolution time-frequency analysis and interpretation of musical rhythm," Ph.D. dissertation, University of Western Australia, 2000.

[25] F. Krebs, S. Böck, and G. Widmer, "An efficient state space model for joint tempo and meter tracking," in *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, 2015, pp. 72–78.

[26] 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.

[27] 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, 2010.

[28] S. Böck, F. Krebs, and G. Widmer, "Joint beat and downbeat tracking with recurrent neural networks," in *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, 2016, pp. 255–261.

[29] F. Krebs, S. Böck, M. Dorfer, and G. Widmer, "Downbeat tracking using beat-synchronous features and recurrent neural networks," in *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, 2016, pp. 129–135.

[30] J. Pons, T. Lidy, and X. Serra, "Experimenting with musically motivated convolutional neural networks," in *Proc. of 14th Intl. Workshop on Content-Based Multimedia Indexing*, 2016.

[31] H. Schreiber and M. Müller, "Musical tempo and key estimation using convolutional neural networks with directional filters," in *Proc. of the Sound and Music Computing Conf.*, 2019, pp. 47–54.

[32] R. Kelz, S. Böck, and G. Widmer, "Deep polyphonic ADSR piano note transcription," in *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 129–135.

[33] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, 2015, pp. 648–656.

[34] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," in *Proc. of the 4th Intl. Conf. on Learning Representations*, 2016.

[35] B. McFee, E. Humphrey, and J. Bello, "A software framework for musical data augmentation," in *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, 2015, pp. 248 – 254.

[36] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks." in *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, 2015, pp. 121–126.

[37] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the variance of the adaptive learning rate and beyond," in *Proc. of the 8th Intl. Conf. on Learning Representations*, 2020.

[38] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba, "Lookahead optimizer: k steps forward, 1 step back," in *Proc. of the 33rd Conf. on Neural Information Processing Systems*, 2019.

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

[40] F. Gouyon and P. Herrera, "Determination of the meter of musical audio signals: Seeking recurrences in beat segment descriptors," in *Audio Engineering Society Convention 114*, 2003.

[41] S. Hainsworth and M. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP Journal on Applied Signal Processing*, vol. 15, pp. 2385–2395, 2004.

[42] F. Gouyon, "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.

[43] G. Peeters and J. Flocon-Cholet, "Perceptual tempo estimation using GMM-regression," in *Proc. of the 2nd ACM workshop on music information retrieval with user-centered and multimodal strategies (MIRUM)*, 2012, pp. 45–50.

[44] G. Percival and G. Tzanetakis, "Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1765–1776, 2014.

[45] P. Knees, A. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. Le Goff, "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections." in *Proc. of the 16th Intl. Society for Music Information Retrieval Conf.*, 2015, pp. 364–370.

[46] H. Schreiber and M. Müller, "A crowdsourced experiment for tempo estimation of electronic dance music," in *Proc. of the 19th Intl. Society for Music Information Retrieval Conf.*, 2018, pp. 409–415.

[47] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[48] U. Marchand and G. Peeters, "Swing ratio estimation," in *Proc. of the 18th Intl. Conf. on Digital Audio Effects*, 2015, pp. 423–428.

[49] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, “An experimental comparison of audio tempo induction algorithms,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, p. 1832–1844, 2006.

[50] A. Gkiokas, V. Katsouros, G. Carayannis, and T. Stafylakis, “Music tempo estimation and beat tracking by applying source separation and metrical relations,” in *Proc. of the 37th IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, 2012, pp. 421–424.

[51] A. Elowsson, “Beat tracking with a cepstroid invariant neural network,” in *Proc. of the 17th Intl. Society for Music Information Retrieval Conf.*, 2016, pp. 351–357.