Multi-Signal Informed Attention for Beat and Downbeat Detection

James Bolt, Johan Pauwels and György Fazekas
School of Electronic Engineering and
Computer Science
Queen Mary University of London
London, UK
Email: {j.g.bolt, j.pauwels, george.fazekas}@qmul.ac.uk

***Abstract*—Processing multiple signal sources presents a challenge but also an opportunity in music information processing, especially when the sources provide complementary information. The attention mechanism paves the way toward addressing this challenge. In this paper, a novel transformer for beat and downbeat detection is proposed, named Informed Beat Transformer. It is theorised to both improve upon previous beat and downbeat detection models and take advantage of auxiliary information. Two experiments are run to test these two hypotheses. In the first experiment, it is directly compared to the Beat Transformer and found to have an average improvement of 0.005 in the F1 score across 4 datasets. The second experiment compares the Beat Transformer and madmom's beat tracker to the Informed Beat Transformer in a situation where a beat coherent signal stream is available, in this case a drum track. It was found to have a significant improvement to the Beat Transformer with an increase in F1 score of 0.702 and an increase of 0.018 to the madmom beat tracker. These results show the efficacy of the Informed Beat Transformer in both experimental settings.**

I. Introduction

The detection of rhythmic components in music is crucial in several applications, including networked music performance and immersive experiences [1] [2], where parts of the performance may be controlled by or need to track the beat of the music. This has become more important with the proliferation of Virtual Reality and virtual concerts [3]. Another example are smart instruments [4], [5], which react to musical content and online delivery and retrieval of audio and music where features such as search, navigation or distributed audio editing may depend on beat related features.

In these contexts, and more generally within the application of Music Information Retrieval in Internet of Sounds [6] applications, there is an emerging need to process multiple information sources or multiple signals concurrently that are available in the environment. This could potentially reinforce decisions or inform algorithmic decision making where such signals are complementary.

So far, most beat tracking algorithms focus on processing a single audio signal. Some early signal processing techniques attempted to add contextual information [7], but this generally relied on handcrafted features, extracted from a single audio source. Modern neural networks have investigated de-mixing audio [8] and combined tempo-beat learning [9], but there has been no investigation into utilising external signals.

In this paper, a new beat transformer is proposed to address this gap. By informing the models' processing via available external sources, we hypothesise that the accuracy of the beat detection model can be improved. Two separate regimes are tested, using different external sources. The first one looks at using the output from a previous beat detection model. The second looks at using an accompanying drum track. Both of these signals are predicted to be coherent with the beats of the input audio and represent an external source which could be substituted for a signal obtained during a networked music event.

II. Background

Beat and downbeat tracking has seen improvements over the last 10 years due to the use of neural networks. Recurrent neural networks (RNNs) [10], convolutional neural networks (CNNs) [11] and deep Bayesian networks (DBNs) [12] all produced increased accuracy when compared to standard signal processing techniques. More recently, the use of temporal convolutional networks (TCNs) [13] and transformers [8], [14], [15], [16] have produced the highest quality outputs. The joint training of tempo and beat estimation has also improved accuracy [8], [9].

Within the transformer architecture, changes to the attention mechanism have been a focal point of research in the last two years. The use of dilated attention (DA), in the Beat Transformer [8] and dilated neighbourhood attention (DNA), in the All-in-One metrical analysis transformer [15] were both designed to reduce computational complexity of the attention mechanism, whilst also improving the accuracy of the results. Alongside this, the use of de-mixing audio, using modern source separation models, has been attempted to leverage the information contained within each instrument in a mix.

However, we hypothesise that modern transformer models could be improved by utilising auxiliary information to infer the location of beats and downbeats. Two major use-cases can be distinguished.

The output of a previous model on the same audio input can be utilised to inform the attention mechanism in transformers such that the increase in accuracy gained over the last few years can be harnessed. This would allow for continual improvement of results, via iteration, without relying

on increasing model size and complexity. This also allows for a reduction in computational costs as previous models have already produced a beat informed embedding. The reduction in computational complexity is a prerequisite for real-time applications, however, our research focuses on demonstrating the concept and was tested in offline contexts. Therefore real-time applications are currently not supported. The next use-case provides an example application for this system.

The attention mechanism can also be informed by other sources, for instance percussive tracks, which often have a strong correlation with the beat, or other beat-correlated sources. A typical regime would be a multi-track setup, potentially live, where the percussive track could aid in determining the beat of more challenging tracks, such as strings.

Both regimes are considered in this work. To achieve them, Informed Attention (IA) is proposed. A similar method has been proposed in machine translation tasks [17], where critical word dependencies are masked in the attention matrix. Another formulation, Mixed-Informed Attention, has also been proposed for few-shot medical image segmentation [18]. This technique utilised query and support features (masked images) to remove background information, thus allowing the network to focus on the relevant information.

The method proposed in this paper follows a similar logic. However, the implementation differs. By obtaining an auxiliary signal, whether via previous models or by other techniques, the importance of areas in the input signal can be determined. Areas of little importance are then masked to allow for the network to focus on the important information. To test this architecture, we compare with the Beat Transformer (BT) [8]. The latter was selected because of its state-of-the-art beat and downbeat detection results across a variety of datasets, and its use of multi-track data in the input data.

In addition to being a reference for comparison, the output of the BT is used as the auxiliary signal in Experiment I. The fact that the BT has varying accuracy across the chosen datasets allows for the validity of the method to be compared for a wide range of informed input accuracy. As the informed architecture relies on input from a previous model, the accuracy of the previous model could be a strong determining factor for the success of IA. Having a variety of accuracies across the dataset allows for this to be tested.

Two main components of the BT architecture will be briefly discussed, which contribute to its strong beat detection performance. Firstly, the use of dilated attention. This follows a similar structure to a TCN, but applied to an attention mechanism. This allows for the model to attend to longer input streams without quadratically increasing the computational complexity. Each level of the dilation architecture produces features for different metrical levels of the input audio. Secondly, the use of instrument-wise attention. This follows the same structure as self attention (which is discussed further in section III-A1), but across the instrument domain, rather than the time domain. As shown in [8], these methods increase the accuracy of the transformer and are therefore both utilised alongside IA in the model presented in this research.

To compare the BT and the Informed Beat Transformer (IBT) in the two regimes outlined above, two experiments were conducted. The first is a direct comparison between the BT and IBT. The second investigates the potential of using specific instrument tracks that are known to have strong beat coherence to explicitly inform the attention mechanism. The aim of these experiments is two-fold. Firstly, to show the potential of IA to improve upon previous beat detection models and secondly, to show the practical uses of IA in a situation where a beat coherent data stream is available, but not part of the target audio.

III. Method

A. Informed Attention

1) Self-Attention: The core of attention in transformers is self-attention, which was introduced in [19]. Self-attention is calculated via Equation 1.

$$\mathrm{Attn}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^{\mathsf{T}}}{\sqrt{d_h}}\right)V = \mathrm{softmax}\left(E\right)V, \tag{1}$$

with Q, K and V being the query, key and value matrices which are formed via a linear projection of the input X and d_h being the head dimensionality. X has a sequence length T and a feature dimension d. Q, K and V all have the dimensionality of (N, T, d/N), with N being the number of heads for multi-head attention and $d_h = d/N$.

If we focus on the attention weight matrix E, we notice that it has dimensions $T \times T$ and its elements $E_{i,j}$ are calculated as

$$E_{i,j} = \frac{Q_{i,:}K^{\mathsf{T}}_{j,:}}{\sqrt{d_h}}, \tag{2}$$

with i and j represent the row indices for Q and K respectively, ranging from 1 to T.

2) Informed Attention: IA follows the same attention mechanism as in Equation 2, but utilises auxiliary information to indicate locations which are more relevant to the problem. This modification is shown in Equation 3.

$$E_{i,j} = \frac{Q_{i,:}\ (K^{\mathsf{T}}_{j,:} + C_j)}{\sqrt{d_h}}. \tag{3}$$

In this case, C is a vector which defines the weight given to each of the T temporal locations. By setting C to $-\infty$ in locations which are not relevant to the problem, C acts as a mask and only T' elements of K contribute to the attention score. The superfluous positions can be removed from the attention matrix and V to achieve a computational speedup.

The mask is obtained by finding the areas of interest from an auxiliary signal. The areas outside of these regions are then set to $-\infty$ in the mask, whereas regions of interest can be weighted to express varying degrees of interest. An example of the resulting attention matrix can be seen in Figure 1.

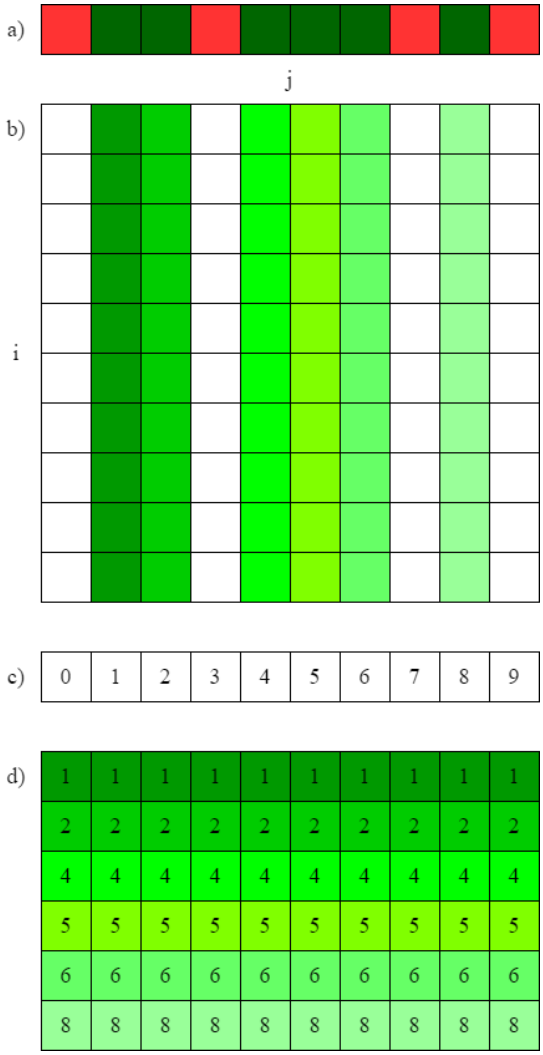

Fig. 1. An example attention matrix for informed attention. a) shows the areas to attend to in green and the areas to mask in red of a sequence c). b) shows the masked attention matrix. The colours match with the reduced matrix shown in d). The reduced matrix removes the $-\infty$ found in the mask for computational efficiency.

B. Obtaining the Attention Locations

1) Procedure I: To utilise the output from a previous model, we found that the best way to create an auxiliary signal starts with processing the raw output from the final layer of the model. For the BT, this is two sequences of length T, one for the beats and one for the downbeats. After applying sigmoid to this output, the locations to attend to will be close or equal to 1. The locations to ignore will be close or equal to 0. Further processing of the signal is explained below. An example output can be seen at the top of Figure 2.

Once the sigmoid function is applied, its output is scaled such that it ranges from m to 0 instead of from 0 to 1. m is negative and its value determines the sensitivity of the attention mechanism to the auxiliary signal. Two adaptive thresholds are then used. Any values above the first are set to 0. Any values below the second are set to m. The results of this can be

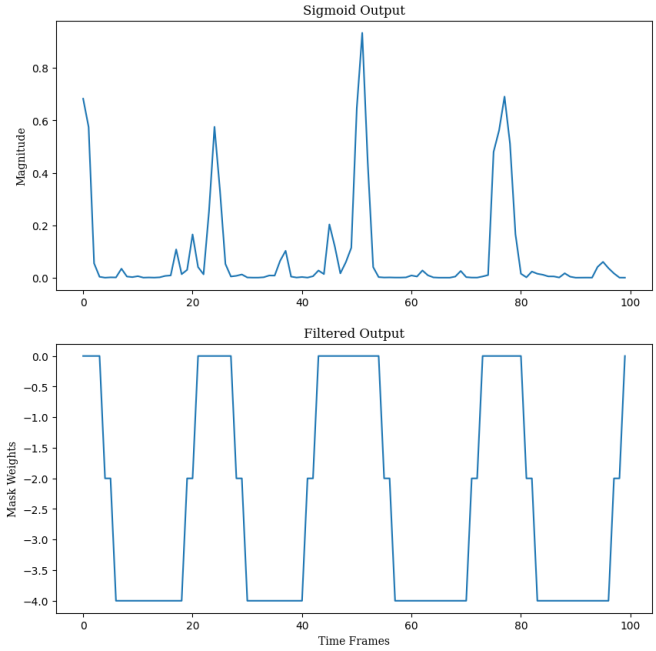

Fig. 2. Output of the BT after being passed through the sigmoid function (Top), and the same output after being passed through the processing outlined in section III-B1 (bottom). For visualisation, the values that would be $-\infty$ after processing have been left at $m = -4$ on the Filtered Output graph.

seen at the bottom of Figure 2. Finally all values at m are set to $-\infty$ and the results are saved. This is performed for every song in the training, validation and test datasets. The use of the adaptive thresholding has two major advantages. Firstly, it allows the output to be correctly scaled regardless of the input range. Therefore, the number of locations to attend to remains relatively linear when compared to the input sequence length regardless of the accuracy of the original model. Secondly, it keeps the information which lies in the middle of the output range. This provides extra information to the attention mechanism by suggesting areas that could be less important but not removing them entirely, allowing the model to determine there relevance. Finally, the width of the peaks in the signal are increased to account for temporal errors from the original model. In the ideal case, these peaks would cover the important regions of the audio allowing for the irrelevant locations to be masked. These values are then used as the mask in Equation 3.

2) Procedure II: In Experiment II the informed input is produced from the drum track of a multi-track dataset. This has to be processed so it follows the structure outlined in section III-B1. To achieve this, the drum audio is processed via the madmom beat tracker [20]. The resulting beats and downbeats are then set as peaks in a vector length T. This is then filtered using a maximum filter to produce wider peaks to allow for errors in the madmom beat detection. The result can be seen in Figure 3.

As in Experiment I, the resulting vector is scaled between 0 and n. Values at n are then set to $-\infty$. These are then used

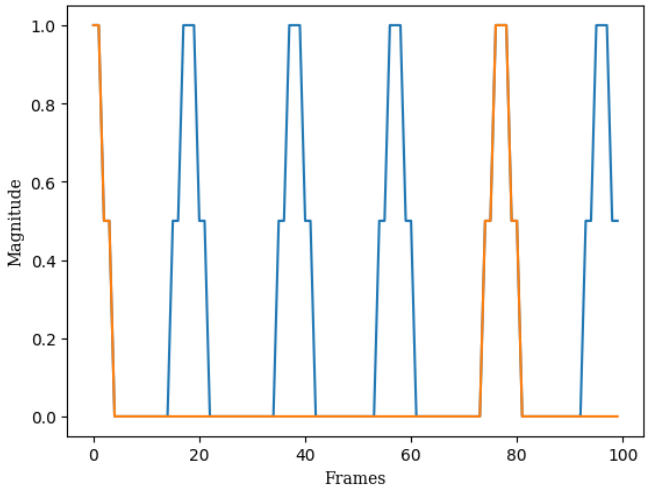

Fig. 3. An example of the filtered peaks used for the informed input in Experiment II. The orange and blue represent the downbeat and beat informed input respectively.

for the informed input when testing the model.

C. Reducing Computational Complexity

One issue with using a masked attention mechanism, as described in Section III-A2, is the computational complexity. Self-attention does not scale well with longer sequence lengths. This is one of the main motivations for the BT [8] and the All-in-One metrical analysis transformer [15]. This is due to the size of the attention matrix created via $\frac{QK^{\mathsf{T}}}{\sqrt{d_h}}$, as the resulting matrix has a dimensionality of $T \times T$. In comparison, the dilation layers in the BT have a dimensionality of $T \times L$, with L being a pre-determined attention length (5 in this case). In IA, entire rows of the attention matrix are being set to $-\infty$. This means that the attention mechanism cannot attend to these locations. Therefore those rows can be entirely removed from the attention matrix. In practice, this involves removing the locations from K, allowing for a precise implementation of equation 3. An example of this can be seen in Figure 1. The reduced size of the attention matrix is $T \times T'$. It should be noted that T' is generally larger than L. For the data used in this research it was usually 10% of T. This reduction is crucial for reducing computational complexity and thus the memory restrictions and speed of training.

D. Model Architecture

An overview of the model architecture can be seen in Figure 4. This architecture can be split into three distinct sections. The encoder, which is taken from the BT, the self-attention, dilation and time attention blocks and the informed attention blocks. The self, dilation and instrument attention blocks are designed to capture information which is not found in the IA blocks. The instrument attention block, pioneered in [8] performs self-attention in the instrument dimension, rather than the time dimension as is standard with music transformers. This is designed to more explicitly utilise the information from each different instrument in the multi-track dataset. Three dilation blocks with an attention length of 5 and dilation levels of 1, 3 and 5 were also utilised. These are designed to capture dependencies at different metrical levels. The standard self-attention block is also used to allow the network to gain an understanding of the entire input sequence without any masking or dilation. The output of all attention layers are combined and then pooled to jointly train beat, downbeat and tempo. Finally the output of the final linear layer is passed through madmoms DBN to calculate the final beats and downbeats.

IV. Experiments

A. Experiment I

The first experiment was designed to directly compare the BT and the IBT. The results would indicate whether IA can improve upon previous models by utilising their outputs. To this end, the experiment was set up in the same way as in [8]. The output of the BT was used as the auxiliary signal source. Due to GPU memory constraints, the Harmonix [21] and Carnatic [22] datasets were removed from the training regime. Therefore, the data consisted of the Hainsworth [23], Ballroom [24], [25], SMC [26], and GTZAN [27] datasets. The input to the models consisted of a multi-track mel spectrogram with 128 mel bins. They were created with a frame size of 2048, a hop size of 1024 and a sample rate of 44100 Hz. The GTZAN dataset was left out for testing only. For processing the auxiliary signal, the value of m in Section III-B1 was set to -4. This value was chosen as it represents a middle ground between providing information to the attention mechanism, without completely dominating the attention matrix. An 8-fold cross-validation was used for both the BT and IBT. The split for this cross-validation was the same in [8]. In order to get the informed input for the IBT, the BT was trained first. The output of this was then processed as described in Section III-B. The processed outputs were then fed into the Informed Transformer during training, validation and testing. Both models were trained for 30 epochs using RAdam as the optimizer and a learning rate of 0.001. Dropout was set to 0.1. The IBT had a size of 9.67M trainable parameters. The BT has a size of 9.29M.

B. Experiment II

Experiment II was designed to test the hypothesized advantages of IA when used in a setting where a beat correlated signal different from the audio signal to process is available. An example of this can be seen in Figure 5.

The pre-trained BT and IBT (from the previous experiment) were tested and compared on a small subsection from the tinyAAM dataset [28]. Specifically, songs were chosen which had drums running through the whole track. This left 5 songs for testing. The audio was synthesized from the tinyAAM MIDI data. In this experiment, a processed version of the drum track is used as informed input. The same value of m was chosen as in Experiment I. The drums are an example of an auxiliary signal stream which can provide information about

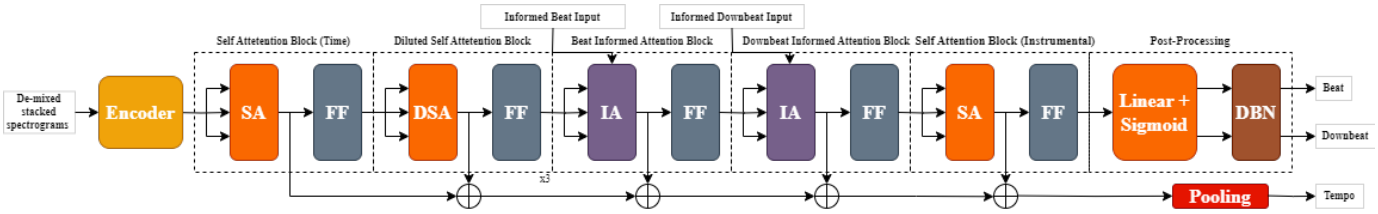

Fig. 4. Overview of the model architecture used in the experiments. The encoder is a combination of convolutional and pooling layers taken from the Beat transformer [8].

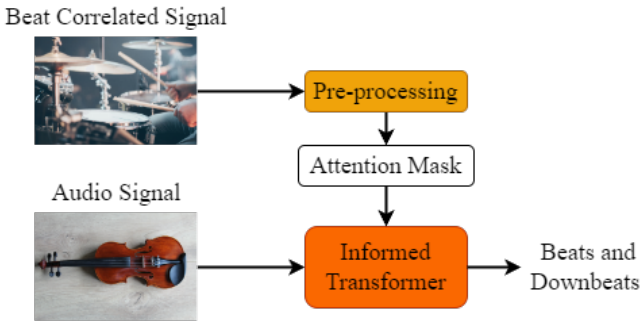

Fig. 5. An example of performing beat and downbeat detection on an audio signal, with a secondary, beat correlated signal present. In this case the violin is the input audio and the drum signal provides the informed input.

TABLE I
COMPARISON OF BEAT AND DOWNBEAT F1-SCORES ON FOUR DATASETS FOR BEAT TRANSFORMER AND INFORMED BEAT TRANSFORMER. THE 95% CONFIDENCE INTERVAL IS ALSO SHOWN.

Dataset	Model	F1 Beat	F1 Downbeat
Ballroom	BT	0.914 ± 0.0105	0.866 ± 0.0169
	IBT	**0.937 ± 0.00667**	**0.872 ± 0.0170**
Hainsworth	BT	0.875 ± 0.0592	**0.699 ± 0.0779**
	IBT	**0.885 ± 0.0530**	0.679 ± 0.0655
SMC	BT	0.544 ± 0.0263	-
	IBT	**0.554 ± 0.0270**	-
GTZAN	BT	0.981 ± 0.00172	0.966 ± 0.00370
	IBT	**0.984 ± 0.000698**	**0.968 ± 0.00177**

the beats and downbeats. The rest of the instruments range from piano and guitar, to cello and flute. The Beat Transformer was tested in three different regimes. The first did not include the drum track in the input audio. The second included the drum track and the final was tested only on the drum track. The Informed Beat Transformer was tested using the drum track as the informed input, once with the drums in the input audio, and once without. As both the Beat Transformer and Informed Transformer are designed to work with tracks without tempo changes, the songs were split, in time, depending on the tempo. For songs with few, or no, tempo changes, further splits were needed to reduce the sequence length to accommodate the GPU constraints.

V. RESULTS AND ANALYSIS

A. Experiment I

The results for Experiment I can be seen in Table I.

The BT results in Table I are lower than the reported results in [8], but the original code was used to obtain these results. This is due to the reduced training dataset as discussed in section IV-A. The results show that IBT improves the results of the BT across the large majority of the datasets. Specifically, the IBT outperforms the Beat Transformer in beat detection across the entirety of the dataset, with large improvements seen in the Ballroom beat detection. The smaller increases seen for the GTZAN dataset is to be expected as the F1 scores for these results are very high already. The small increase to the SMC results is less significant due to the low F1 scores for both models. This could be due to the lack of clear peaks in the BT output. This would then produce a less consistent informed input for the IBT. The main outlier in the results is the decrease in the downbeat F1 score for the Hainsworth dataset. The Hainsworth dataset had the most varied scores across the 8 folds for both the BT and IBT. This implies that the output for this dataset is varied and this would affect the informed input and therefore the IBT. The rest of the results show small improvements for both beat and downbeat detection. This consistent increase in F1 scores show the validity of IBT for improving upon previous beat detection models. The 95% confidence measures, found using the standard deviation of the scores across the 8 folds, suggests the IBT produced less variance for the majority of the datasets. This suggests that the IBT is more consistent than the BT.

Another point of comparison is the model size and training times. The IBT has 9.67M parameters compared to the 9.29M of the BT. However, its training time was significantly reduced. This was due to both the quicker convergence and the attention matrix reduction outlined in Section III-C. The matrix reduction reduced the attention matrix size by as much as 90% from the majority of the data. Whilst the IBT and BT where both trained for 30 epochs, the IBT produced strong results very early in the process. Often getting close to convergence after only 5-10 epochs. The faster convergence time is something that transformer architectures often struggle with [19] and these results demonstrate the advantages of IA. It also had a far lower epoch time, with the BT taking approximately 35 minutes per epoch compared to the 14 minutes of the IBT. It should be noted that the IBT does require an input from a previous model/signal stream. In this use-case this adds to the training time cost, as the BT needs to be trained and the results processed. However, in a regime where the auxiliary input comes from a separate data-stream, such as in Experiment II,

TABLE II
RESULTS OF EXPERIMENT II. "NO DRUMS" MEANS THAT DRUMS WERE NOT PRESENT IN THE INPUT AUDIO. "DRUMS ONLY" MEANS ONLY THE DRUMS WERE USED IN THE INPUT AUDIO.

Model	F1 Beat
BT No Drums	0.239
BT Drums	0.256
BT Drums Only	0.3041
IBT No Drums	0.958
IBT Drums	0.895

this would not be a sizeable factor. It should also be noted that, in iterative refinement, there is always the added cost of previous iterations, but if we can achieve a better result with the refinement architecture being "cheaper" to compute than the previous step (original algorithm) that's still an advantage.

B. Experiment II

In Table II it can be seen that the Informed Transformer outperforms the Beat Transformer to a significant degree, both when trained with and without the drums in the input mix. The large difference in results implies two main conclusions. Firstly, that the BT does not perform well on data that is out-of-domain. As the BT was trained using the data in Experiment I, but then tested on the tinyAAM dataset for Experiment II, the poor F1 score for the beat tracking suggests this. Secondly, when given a strong informed input, the IBT performs well on out-of-domain data. This is a positive result and shows the validity of the IBT in this setting. It should be noted that, while these results are significant, the results for the madmom beat tracker, which pre-processed the drums for the informed input, should be taken into account. When ran on the input mix, without the drums, madmoms beat tracker produced a beat F1 score of 0.940. Whilst the IBT still outperforms this, the difference is small. This suggests a major limitation of the BT, alongside the efficacy of IBT. Another interesting point to note is that the informed input signals for testing were created differently from those at training time. Whilst the post-processing was similar, the BT was used for training in this experiment (since the model trained in Experiment I was reused) but madmom's beat tracker for testing. However, the IBT still managed to perform at high levels without further training, regardless of these differences.

VI. CONCLUSION

The results from Experiment I show improvements across almost every dataset. This alone is a strong result, as in general, increases for beat detection models have not been so consistent in recent years. When comparing the results of the Hung Transformer [14], the Bock transformer [9] and the BT, as done in [8], the model that produces the best result varies across each dataset. This difference makes the consistent improvement found in this research worth noting. To look further into this, the IBT could be trained using the outputs of different beat detection models. If consistent improvement was found to all then it would suggest that IA could be used to continually increase accuracy. This could be done by repeated training and processing of the output from the IBT, creating a feedback loop between the IBT output and its informed input. This is something that needs further investigation.

The significantly reduced epoch time of the IBT suggests, with further refinement, a potential application to real-time processing, something that is difficult to achieve with modern transformer architectures due to their large size. This makes the IBT more suited to the potential live performance applications tested in Experiment II.

The outcome of Experiment II suggests that the use of IA could greatly benefit situations where external information regarding beats and downbeats is available. This could include live performance, or online tracking where extra information may not be in the form of an audio stream. Due to the flexibility of the IBT in regards to the informed input, it is suggested that the applications could go beyond the ones suggested in this paper. Furthermore, the application of IBT is not limited to beat tracking. By changing the task and the informed input, it is suggested that IBT could be easily adapted to other MIR tasks such as chord recognition. In this case, utilising pitch information from a harmonically rich instrument, such as a piano, chord help inform chord recognition models, when trained on more percussive sounds. This is an area of research which has great potential.

There are a number of other different avenues for further research. Experiment II indicates that using the IA architecture with multi-track data more explicitly, having multiple informed inputs for each instrument, for example, could prove fruitful. As discussed above, it is also worth investigating the potential of continually improving results by creating a feedback loop between the IBT output and the informed input across multiple training sessions.

Overall, there are advantages and disadvantages to using IA over other transformer architectures for beat detection. The reduced training and processing time of IA is noticeable and provides unique practical opportunities. The out-of-domain accuracy of the IBT also suggests this. However, the IBT is reliant on the informed input. If this is not obtainable, or unreliable, the accuracy suffers. Therefore, a major factor when training and applying this model is the pre-processing and feature crafting of these inputs. This can be time consuming and depending on the data stream used, be difficult to carry out. The IA architecture has also been shown to have practical applications in networked, or live music contexts. For example, using an auxiliary data stream to determine the beats of an audio signal to control the lights of a concert. Whilst the real-time validity of this needs to be explored, it is a step in the right direction for informed MIR.

REFERENCES

[1] L. Turchet, M. Lagrange, C. Rottondi, G. Fazekas, N. Peters, J. Ostergaard, F. Font, T. Backstrom, and C. Fischione, "The Internet of Sounds: Convergent Trends, Insights, and Future Directions," *IEEE Internet of Things Journal*, vol. 10, no. 13, pp. 11 264–11 292, July 2023.

[2] C. Rottondi, C. Chafe, C. Allocchio, and A. Sarti, "An Overview on Networked Music Performance Technologies," *IEEE Access*, vol. 4, pp. 8823–8843, 2016.

[3] B. Loveridge, "Networked Music Performance in Virtual Reality: Current Perspectives," *Journal of Network Music and Arts*, vol. 2, no. 1, pp. 1–19, 2020.

[4] L. Tuchet, A. Mcpherson, and C. Fischione, "Smart instruments: Towards an ecosystem of interoperable devices connecting performers and audiences," in *Proceedings of Sound and Music Computing Conference*, 2016, pp. 498–505. [Online]. Available: http://instrum.ircam.fr/smartinstruments/

[5] L. Turchet, D. Stefani, and J. Pauwels, "Musician-AI Partnership Mediated by Emotionally-Aware Smart Musical Instruments," *International Journal of Human-Computer Studies*, vol. 191, p. 103340, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S107158192400123X

[6] L. Turchet, J. Pauwels, C. Fischione, and G. Fazekas, "Cloud-smart Musical Instrument Interactions: Querying a Large Music Collection with a Smart Guitar," *ACM Transactions on Internet of Things*, vol. 1, no. 3, pp. 1–29, July 2020.

[7] M. E. Davies and M. D. Plumbley, "Context-dependent Beat Tracking of Musical Audio," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1009–1020, March 2007.

[8] J. Zhao, G. Xia, and Y. Wang, "Beat Transformer: Demixed Beat and Downbeat Tracking with Dilated Self-Attention," in *arXiv:2209.07140*, 2022.

[9] S. Böck, M. E. P. Davies, and P. Knees, "Multi-Task Learning of Tempo and Beat: Learning One to Improve the Other," in *International Society for Music Information Retrieval*, 2019, pp. 486–493.

[10] S. Böck and M. Schedl, "Enhanced Beat Tracking with Context-aware Neural Networks," in *International Conference on Digital Audio Effects*, 2011, pp. 135–139.

[11] S. Durand, J. P. Bello, B. David, and G. Richard, "Feature Adapted Convolutional Neural Networks for Downbeat Tracking," in *International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 296–300.

[12] S. Böck, F. Krebs, and G. Widmer, "A Multi-model Approach to Beat Tracking Considering Heterogeneous Music Styles," in *International Society for Music Information Retrieval*, 2014, pp. 603–608.

[13] M. E. P. Davies and S. Böck, "Temporal Convolutional Networks for Musical Audio Beat Tracking," in *European Signal Processing Conference*, 2019, pp. 1–5.

[14] Y. N. Hung, J. C. Wang, X. Song, W. T. Lu, and M. Won, "Modeling Beats and Downbeats with a Time-Frequency Transformer," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 2022, pp. 401–405.

[15] T. Kim and J. Nam, "All-in-One Metrical and Functional Structure Analysis with Neighborhood Attentions on Demixed Audio," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1–5.

[16] T. Cheng and M. Goto, "Transformer-Based Beat Tracking with Low-Resolution Encoder and High-Resolution Decoder," in *ISMIR*, 2023, pp. 466–473.

[17] H. Deguchi, A. Tamura, and T. Ninomiya, "Dependency-based self-attention for transformer NMT," in *International Conference Recent Advances in Natural Language Processing, RANLP*. Incoma Ltd, 2019, pp. 239–246.

[18] J. Li, Z. Wang, and S. Zhu, "Mixed Informed Transformer for Few-Shot Medical Image Segmentation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Institute of Electrical and Electronics Engineers (IEEE), March 2024, pp. 1501–1505.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[20] 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 2016 ACM Multimedia Conference*, 2016, pp. 1174–1178.

[21] O. Nieto, M. Mccallum, M. E. P. Davies, A. Robertson, A. Stark, and E. Egozy, "The Harmonix Set: Beats, Downbeats, and Functional Segment Annotations of Western Popular Music." in *Proc. of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 565–572. [Online]. Available: https://musicbrainz.org/

[22] A. Srinivasamurthy and X. Serra, "A Supervised Approach to Hierarchical Metrical Cycle Tracking from Audio Music Recordings," in *International Conference on Acoustics, Speech and Signal Processing*, 2014, pp. 5237–5241.

[23] S. W. Hainsworth and M. D. Macleod, "Particle filtering Applied to Musical Tempo Tracking," in *EURASIP Journal on Advances in Signal Processing*, 2004, pp. 1–11.

[24] 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, pp. 1832–1844, 2006.

[25] K. Florian, S. Böck, and G. Widmer, "Rhythmic Pattern Modeling for Beat and Downbeat Tracking in Musical Audio," in *International Society for Music Information Retrieval*, 2013, pp. 227–232.

[26] H. Andre, M. Davies, J. Zapata, J. Lobato Oliveria, 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] 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.

[28] F. Ostermann and I. Vatolkin, "Tiny AAM: Sample from the Artificial Audio Multitracks Dataset (v1.1.0) [Data set]," 2022.