BEAT-U: MULTI-TASK MUSIC UNDERSTANDING WITH HIERARCHICAL TIMESCALES

Jingwei Zhao[♯,♭] **Rei Nishiyama**[♯] **Kouhei Sumi**[♯]
Takuya Fujishima[♯] **Akira Maezawa**[♯]
[♯] Yamaha Corporation [♭] National University of Singapore
jzhao@u.nus.edu, first.last@music.yamaha.com

ABSTRACT

We present Beat-U, a multi-task U-shape Transformer for music understanding across multiple timescales. It jointly addresses four sequential MIR tasks—beat tracking, downbeat tracking, chord recognition, and structure analysis—assigning each to a proper temporal scale while benefiting from shared representations. Training and evaluation are conducted on public beat-tracking datasets and an internal J-Pop corpus annotated for all four tasks. Experiments show highly competitive results across all tasks on pop music, while a genre-breakdown analysis reveals underfitting on more diverse styles, likely due to the predominance of J-Pop in the training data. This highlights cross-genre generalisation as an important direction for our future work.

1. INTRODUCTION

Music foundation models have opened new possibilities for music information retrieval (MIR). Examples include BERT-style models for music audio [1,2], generative models [3–5], and cross-modal architectures aligned between audio and other modalities [6–8]. Through large-scale pre-training, these models produce intermediate representations that capture broader aspects of musical content, thereby enhancing performance across diverse downstream MIR tasks. Such a paradigm has proven particularly effective for time-invariant tasks, including genre/key classification [9] and music captioning [10], where a global label is derived from aggregated representations. However, extending pre-trained music representations to sequential, time-varying tasks remains challenging [11, 12].

In this pilot study, we investigate how multiple sequential MIR tasks can be integrated into a shared representation framework. We focus on four classical tasks: (1) beat tracking, (2) downbeat tracking, (3) chord recognition, and (4) structure analysis. Figure 1 illustrates that all four tasks share a common formulation as *sequence classification* problems, which involve boundary detection (i.e., binary 0/1 labels) and, optionally, interval classification (e.g., chord qualities). Despite the commonality, each task requires a distinct temporal resolution to localise the boundaries, which may not naturally align with that of pre-trained music representations. Moreover, the hierarchical dependencies in this multi-task setting are not explicitly encoded by existing music foundation models.

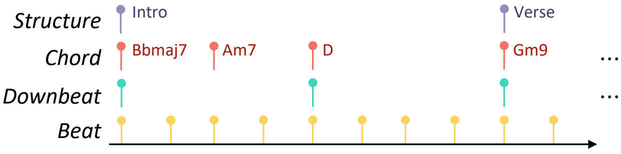

Figure 1. Beat, downbeat, chord progression, and phrase structure exhibit hierarchical dependencies within a piece of music, each unfolding at a distinct timescale.

To enable multi-task, multi-resolution sequence modelling, we propose *Beat-U*, a U-shape Transformer jointly designed for beat/downbeat tracking, chord recognition, and song structure analysis. Analogous to U-Net [13], the proposed model comprises stacks of Transformer encoder modules interleaved with downsampling and upsampling operators. Beat tracking, the most fine-grained boundary detection task, is handled at the top level without downsampling, while structure analysis, requiring the coarsest temporal scale, is modelled at the bottom, most downsampled level. Downbeat tracking and chord recognition are assigned to intermediate levels, each aligned with an appropriate temporal resolution. This design ensures that every task operates at a proper timescale and mitigates the class imbalance between positive and negative labels in sequence classification. Furthermore, by integrating all four tasks within a unified framework, we can leverage shared representations such that the joint hierarchical understanding benefits individual tasks.

In our preliminary experiments, we train and evaluate the proposed model on public beat/downbeat tracking datasets in addition to a larger internal J-Pop music dataset annotated for all four tasks. Experimental results on pop music demonstrate highly competitive performance across each task. A genre-specific breakdown, however, suggests possible underfitting to other genres, likely due to the predominance of J-Pop in our training data.

Attribution: J. Zhao, R. Nishiyama, K. Sumi, T. Fujishima, and A. Maezawa, "Beat-U: Multi-Task Music Understanding with Hierarchical Timescales", in *Extended Abstracts for the Late-Breaking Demo Session of the 26th Int. Society for Music Information Retrieval Conf.,* Daejeon, South Korea, 2025.

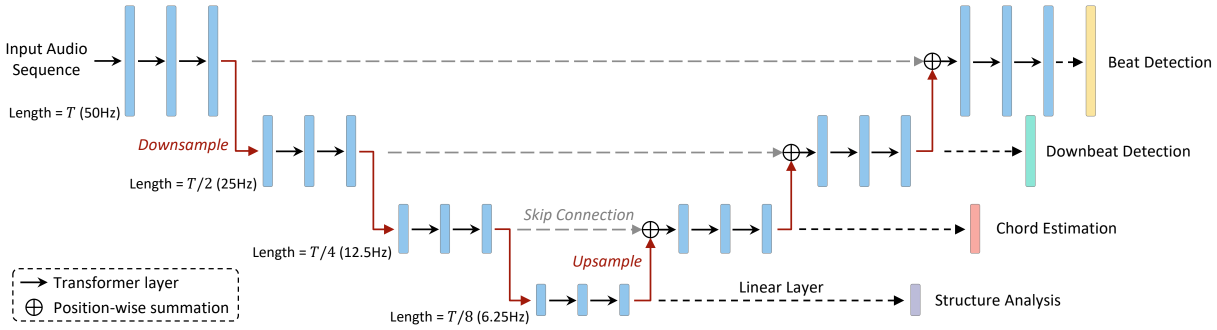

Figure 2. Beat-U is a U-shape Transformer with 7 encoder blocks across 4 levels of temporal resolution. Each level models a designated MIR task (beats, downbeats, chords, and structure), enabling joint modelling across hierarchical timescales.

2. METHOD

We propose a Transformer-based U-shape architecture with 4 levels of timescale abstraction, each tailored to a specific MIR task. Section 2.1 introduces our music representation, and Section 2.2 details the model architecture.

2.1 Music Representation

We use the EnCodec-32k model [14] to preprocess music data, encoding each song as a discrete code sequence of 50Hz. EnCodec representations retain rich musical content and have been adopted as the input space for several music foundation models. Compared with spectrograms or deeper continuous embeddings, their discrete codes are also more efficient to handle. For model input, we recover the continuous representations from the pre-trained vector-quantisation codebook, which, in our preliminary experiments, outperformed learning new embeddings directly from the discrete codes.

2.2 Model Architecture

As shown in Figure 2, our model comprises 7 stacked Transformer encoder blocks. Connected through downsampling and upsampling operators, these blocks span 4 temporal-resolution levels and form a U-shape architecture. The input sequence is gradually downsampled by a factor of 2 at each block and then upsampled back to the original resolution. At every resolution level, the output is projected through a linear layer to the boundary (and quality) detection logits for an MIR task at that very timescale.

Each encoder block consists of 8 Transformer layers with *dilated self-attention*, an efficient sparse attention implementation optimised for beat/downbeat tracking [15]. The attention window has a base length of 9 and expands exponentially as the layer goes deeper. The layers in one block yield a receptive field of 2,048 frames (i.e., each frame can reach up to 1,024 frames away on both sides). At the top-level block without downsampling, this corresponds to 40s audio at a 50Hz frame rate. At the bottom-level with 8x downsampling, the receptive field is stretched up to 5min at 6.25Hz. Across all 4 levels, our model captures hierarchical timescales from local segments to the entire song. Different MIR tasks are supervised at selected timescales: beats at 50 Hz, downbeats at 25 Hz, chord progression at 12.5 Hz, and structure boundaries at 6.25 Hz. Blocks at the same timescale are linked with skip connections, enabling the integration of both coarse, global information and fine-grained, local detail.

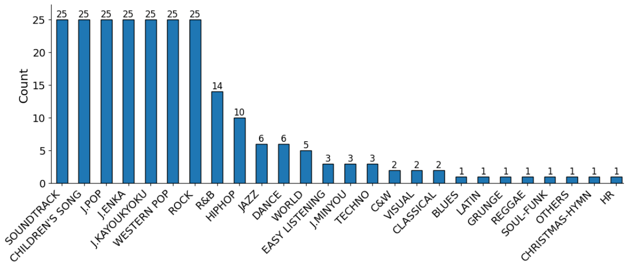

Figure 3. Genre distribution for the Balanced test set.

3. PRELIMINARY EXPERIMENT

3.1 Datasets

We leverage an internal dataset of 10k music pieces, primarily in J-Pop, annotated with genre, beat/downbeat, chord progression, and structural boundaries. For evaluation, we hold out two test sets: (1) JPop: 250 pieces explicitly labelled as J-Pop, and (2) Balanced: 239 pieces spanning a broader range of genres. While the latter set remains biased toward J-Pop, it provides a wider genre coverage as shown in Figure 3. The remaining pieces are split 90%–10% for training and validation. We apply pitch augmentation by transposing each training sample to ±3 semitones. As this preliminary study places more emphasis on beat tracking, we also incorporate public beat/downbeat tracking datasets: Ballroom [16], Hainsworth [17], RWC-Pop [18], Harmonix [19], and SMC [20]. Each of these datasets is split 90%–10% for training and validation. GTZAN [21] is held out exclusively for testing.

3.2 Configuration and Training Details

Across the Transformer layers, we apply 8 attention heads, 0.1 dropout ratio [22], and layer norm [23] before attention. The hidden dimensions are set to $d_{\text{model}} = 128$ for the attention layers and $d_{\text{ff}} = 512$ for the feed-forward layers. In addition to sinusoidal positional encoding [24], we

Test Set	Model	Beat Accuracy			Downbear Accuracy			Chord Accuracy		Structure Boundary	
		F-Measure	CMLt	AMLt	F-Measure	CMLt	AMLt	MeanSeg	MajMin	HR.5F	HR3F
JPop	Beat-U (Ours)	**0.966**	**0.945**	**0.971**	**0.957**	**0.931**	**0.951**	0.886	**0.807**	**0.739**	**0.767**
	Beat-This	0.940	0.896	0.866	0.916	0.864	0.832				
	Jiang et al.							**0.901**	0.800		
	All-In-One	0.920	0.813	0.930	0.900	0.804	0.927			0.576	0.740
Balanced	Beat-U (Ours)	**0.926**	**0.902**	**0.931**	**0.895**	**0.872**	**0.907**	0.874	**0.819**	**0.673**	**0.711**
	Beat-This	0.906	0.845	0.877	0.863	0.767	0.811				
	Jiang et al.							**0.890**	0.812		
	All-In-One	0.876	0.786	0.900	0.827	0.765	0.880			0.534	0.684

Table 1. Comprehensive evaluation for four MIR tasks on the JPop and Balanced test sets.

incorporate relative music timing condition [3] to indicate the progressive structure of a song. The complete model comprises 10.3M learnable parameters.

For model training, each boundary detection task is formulated as a binary sequence classification problem, where steps in the sequences are labelled as either 1 or 0. We adopt the shift-tolerant BCE loss [25] to encourage confident, sharp peaks for each task. The tolerance window is set to 3, 2, 2, and 2 frames for beat, downbeat, chord boundary, and structure boundary, respectively. Positive examples are further weighted to address class imbalance.

The model is trained on entire music pieces with a batch size of 16 for 150 epochs (100k iterations), using two H100 GPUs under FP16 precision. Optimisation is performed with AdamW [26] at an initial learning rate of 1e-4, scheduled by a 250-step linear warm-up followed by cosine decay to a final rate of 1e-6. At test time, peak picking [25] is applied to extract boundaries from the raw activations.

3.3 Evaluation Results

We select three baseline models: Beat-This [25] for beat and downbeat tracking, Jiang et al. [27] for chord recognition, and All-In-One [28] for song structure analysis. We report evaluation results using standard metrics: F-Measure and continuity metrics for beat/downbeat tracking, MeanSeg and MajMin for chord recognition, and hit rates at 0.5s and 3s for structure boundary detection.

Table 1 presents the evaluation results on J-Pop and Balanced across the four MIR tasks. For beat and downbeat tracking, our model outperforms the baseline by a clear margin, particularly on continuity metrics. For structure boundary detection, we also observe improvements, most notably on the finer HR.5 metric. These improvements suggest that our model can more consistently capture the metrical structure of a song, presumably benefiting from our multi-task formulation, architectural design, and the ability to process entire songs at a time. Chord recognition accuracy is comparable to the baseline, with a slight improvement in major/minor quality classification. Overall, both our model and the baselines perform better on the J-Pop dataset than on the Balanced set, which is expected since pop music generally exhibits a more straightforward metrical structure. This also highlights the challenge of accurately modelling more diverse music genres.

With this in mind, we further evaluate our model on the GTZAN dataset for beat/downbeat tracking and analyse the results by genre. GTZAN contains 1k music segments spanning 10 diverse genres, including pop, jazz, and classical. As shown in Table 2, our model remains highly competitive with the baseline on Disco, Reggae, and Pop, but performs notably worse on Blues, Jazz, and Classical. This suggests underfitting to non-pop genres, likely due to the predominance of J-Pop in our training data. These genres often feature more diverse acoustic instrumentation and more dynamic tempo variations, and we aim to address these challenges in our future work.

Genres	Beat F-Measure		Downbeat F-Measure	
	Ours	Beat-This	Ours	Beat-This
Overall	0.850	**0.891**	0.736	**0.787**
Country	0.930	**0.944**	0.903	**0.915**
Disco	0.965	**0.966**	0.931	**0.937**
Hiphop	0.945	**0.975**	0.851	**0.898**
Pop	0.944	**0.953**	0.923	**0.939**
Reggae	**0.908**	0.907	**0.745**	0.688
Rock	0.907	**0.930**	0.793	**0.821**
Metal	0.847	**0.877**	0.757	**0.769**
Blues	0.787	**0.831**	0.552	**0.639**
Jazz	0.757	**0.868**	0.534	**0.749**
Classical	0.508	**0.659**	0.353	**0.514**

Table 2. Genre-breakdown evaluation on the test-only GTZAN dataset for beat and downbeat tracking.

4. CONCLUSION

In this pilot study, we introduced Beat-U, a U-shape Transformer architecture for multi-task, multi-scale MIR. By assigning beat, downbeat, chord, and structure analysis to different temporal resolutions within a unified framework, our approach balances label sparsity and captures both fine-grained and global musical context. Preliminary experiments on public datasets and internal J-Pop corpus demonstrate highly competitive performance, particularly on pop music, with improvements in beat/downbeat tracking, chord recognition, and structure boundary detection.

Despite these improvements, the current model is limited to simple chord qualities and structure boundaries. In future work, we plan to incorporate chord tensions and structural functions to enhance its capability and versatility. We also aim to investigate augmentation strategies and leverage larger, more diverse datasets to improve the model's generalisation across musical genres and styles.

5. REFERENCES

[1] Y. Li, R. Yuan, G. Zhang, Y. Ma, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. B. Dannenberg, R. Liu, W. Chen, G. Xia, Y. Shi, W. Huang, Z. Wang, Y. Guo, and J. Fu, "MERT: acoustic music understanding model with large-scale self-supervised training," in *ICLR 2024*.

[2] M. Won, Y. Hung, and D. Le, "A foundation model for music informatics," in *ICASSP 2024*.

[3] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[4] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," in *NeurIPS 2023*.

[5] W. Liao, Y. Takida, Y. Ikemiya, Z. Zhong, C. Lai, G. Fabbro, K. Shimada, K. Toyama, K. W. Cheuk, M. A. M. Ramírez, S. Takahashi, S. Uhlich, T. Akama, W. Choi, Y. Koyama, and Y. Mitsufuji, "Music foundation model as generic booster for music downstream tasks," *Trans. Mach. Learn. Res.*, 2025.

[6] S. Wu, Z. Guo, R. Yuan, J. Jiang, S. Doh, G. Xia, J. Nam, X. Li, F. Yu, and M. Sun, "Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages," in *ACL 2025*.

[7] Z. Deng, Y. Ma, Y. Liu, R. Guo, G. Zhang, W. Chen, W. Huang, and E. Benetos, "Musilingo: Bridging music and text with pre-trained language models for music captioning and query response," in *NAACL 2024*.

[8] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *ICASSP 2024*.

[9] R. Castellon, C. Donahue, and P. Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *ISMIR 2021*.

[10] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," in *ICML 2024*.

[11] Y. Ma, S. Li, J. Yu, E. Benetos, and A. Maezawa, "Cmi-bench: A comprehensive benchmark for evaluating music instruction following," *arXiv preprint arXiv:2506.12285*, 2025.

[12] Y. Zhang, H. Chen, J.-C. Wang, and J. Chen, "Temporal adaptation of pre-trained foundation models for music structure analysis," *arXiv preprint arXiv:2507.13572*, 2025.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI 2015*.

[14] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *Trans. Mach. Learn. Res.*, 2023.

[15] J. Zhao, G. Xia, and Y. Wang, "Beat transformer: Demixed beat and downbeat tracking with dilated self-attention," in *ISMIR 2022*.

[16] F. Krebs, S. Böck, and G. Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio," in *ISMIR 2013*.

[17] S. W. Hainsworth and M. D. Macleod, "Particle filtering applied to musical tempo tracking," *EURASIP J. Adv. Signal Process.*, 2004.

[18] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical and jazz music databases," in *ISMIR 2002*.

[19] O. Nieto, M. McCallum, M. E. P. Davies, A. Robertson, A. M. Stark, and E. Egozy, "The harmonix set: Beats, downbeats, and functional segment annotations of western popular music," in *ISMIR 2019*.

[20] A. Holzapfel, M. E. P. Davies, J. R. Zapata, J. L. Oliveira, and F. Gouyon, "Selective sampling for beat tracking evaluation," *IEEE Trans. Speech Audio Process.*, 2012.

[21] C. Dittmar, M. Pfleiderer, and M. Müller, "Automated estimation of ride cymbal swing ratios in jazz recordings," in *ISMIR 2015*.

[22] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, 2014.

[23] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NISP 2017*.

[25] F. Foscarin, J. Schlüter, and G. Widmer, "Beat this! accurate beat tracking without DBN postprocessing," in *ISMIR 2024*.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR 2018*.

[27] J. Jiang, K. Chen, W. Li, and G. Xia, "Large-vocabulary chord transcription via chord structure decomposition," in *ISMIR 2019*.

[28] T. Kim and J. Nam, "All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio," in *WASPAA 2023*.