# MusT3: Unified Multi-Task Model for Fine-Grained Music Understanding

**Martin Kukla, Minz Won, Yun-Ning Hung & Duc Le**
SAMI, ByteDance
{Martin.Kukla,Minz.Won,Yunning.Hung,Duc.Le}@bytedance.com

## Abstract

Recent advances in sequence-to-sequence modelling enabled new powerful multi-task models in text, vision, and speech domains. This work attempts to leverage these advances for music. We propose **MusT3** (**Mus**ic-**T**o-**T**ags **T**ransformer), a novel model for fine-grained music understanding. First, we design the unified music-to-tags form, which enables us to cast any music understanding task as sequence prediction problem. Second, we utilize Transformer-based model to predict that sequence given music representation. Third, we leverage multi-task learning framework to train a single model for many tasks. We validate our approach on four tasks: beat tracking, chord recognition, key detection, and vocal melody extraction. Our model performs significantly better than the current state-of-the-art models on two of these tasks, while staying competitive on the remaining two. Finally, in controlled experiment, we demonstrate that our model can reuse knowledge between tasks, leading to better performance on low-resource tasks with limited training data.

## 1 Introduction

Music understanding focuses on extracting meaningful features from music. For fine-grained music understanding tasks, the extracted features are specific to a certain time in a piece of music. Since these features can often be only understood by trained musician (e.g. chord triad), the data labelling is expensive. Consequently, there is limited training data available, which can also vary between different music understanding tasks. In order to utilize the available data well, we argue that one should train a single unified multi-task model capable of reusing knowledge between the tasks.

Sequence-to-sequence models have been successfully used for multi-task learning for text (Raffel et al. [2020], Xue et al. [2020]), vision (Hu and Singh [2021], Bhattacharjee et al. [2022], Seong et al. [2019]) and speech (Moritz et al. [2020]). Inspired by T5 (Raffel et al. [2020]), we also use the unified form for expressing any task. However, unlike their work, we do not fine-tune our model to task at hand. Instead, we train a single model simultaneously for multiple tasks. This is similar to MulT (Bhattacharjee et al. [2022]), but, unlike that work, we use a single transformer head for all the tasks.

There has been a significant amount of prior work for beat tracking, chord recognition, and key detection. Most of the work has been focused on single task (Böck and Davies [2020], Hung et al. [2022], Zhao et al. [2022], Park et al. [2019], Korzeniowski and Widmer [2017], Jiang et al. [2018]). There is some work trying more than one task (Lee and Slaney [2007], Wu et al. [2020], Fuentes et al. [2019], Kim and Nam [2023], Mauch et al. [2009b]), but, to our knowledge, there is no work trying more than two tasks. This work attempts to model many tasks by a single, unified architecture.

Automatic music transcription (AMT), which includes voice melody extraction, is often thought of as multi-task problem (Lu et al. [2023], Manilow et al. [2020] Lin et al. [2021], Cheuk et al. [2022],

37th Conference on Neural Information Processing Systems (NeurIPS 2023) - ML For Audio Workshop.

Hung et al. [2019] Bittner et al. [2018], Gardner et al. [2021]). However, we argue that to transcribe different instruments (including voice) is not to perform semantically different functions. Thus, in our opinion, AMT should not be viewed as multi-task model in its traditional meaning. Nevertheless, the closest work to ours is MT3 (Gardner et al. [2021]), which also uses a single representation for multiple transcription tasks, and also, leverages modern sequence-to-sequence architecture (T5). However, in MT3's MIDI-like representation, the length of the output sequence increases with the number of tasks, making it unsuitable for modelling many tasks. Our time-synchronous representation addresses this issue.

## 2 The Unified Model

**Music-to-Tags**: We formulate any music understanding task as sequence-to-sequence problem. We use pre-trained model to represent a piece of music as sequence of continuous embeddings. By prefixing the sequence with the task, we create the input to our model. The output of the model is the sequence of tags. Each output tag describes a single frame of the input music. Which frame a tag describes is determined by task temporal resolution, and the position of the tag in the output sequence. For example (below), in 2Hz key detection task for 3 seconds music, the output will be 6 tags long, and the second tag would describe the second half of the 1st second. Every task can have different temporal resolution. From computational reasons, our model can't be applied to a whole song at inference time. Instead, we split a song in smaller overlapping chunks, feed them into the model and combine tags probabilities from multiple outputs.


**Multi-task learning**: Task loss is defined as an average Binary Cross Entropy loss per tag for output sequence. We define multitask loss as a weighted sum of task losses, and use GradNorm (Chen et al. [2018]) to set the tasks weights. During training, we partition a training batch equally among all the tasks. As an alternative to multi-task learning, we have also tried multi-task mixture approach (Xue et al. [2020], Gardner et al. [2021]). However, in our preliminary experiments, this yielded worse results.

**Music representation sequence**: We use the in-house representational learning model (Won et al. [2023]) to obtain music representation sequence for music. The model is based on BestRQ architecture(Chiu et al. [2022]), and it has been trained on 160 hours of in-house music data. The model's output temporal resolution is 25Hz: for 10 seconds-long music used in this work, the model generates 250 long representation sequence. We refer to Won et al. [2023] for more details on the model.

**Sequence-to-sequence model**: LLaMa (Touvron et al. [2023]) architecture is used as our sequence-to-sequence model. LLaMa is a model based on Transformer (Vaswani et al. [2017]) with minor architectural changes (we refer to Touvron et al. [2023] for details). In addition, our model consists of embedding layer for task and the output layer for tags (linear layer + GeLu (Hendrycks and Gimpel [2016]) + linear layer). The total number of model parameters is 108 millions.

## 3 Music Understanding Tasks

We focus on four fine-grained music understanding tasks: beat tracking, chord recognition, key detection and vocal melody extraction. For every task, we use the same input length of 10 seconds. To split a song into 10 seconds chunks, we apply sliding window with hop factor of 2 seconds. As a result, we get multiple output probabilities for a single tag. We combine probabilities by summation, followed by normalization. We provide the summary of the task in the Table 1, while we give some task-specific details below.

**Beat tracking**: The task combines beat and downbeat tracking with tempo prediction. The size of tag dictionary for all three subtasks is 303 (extra 300 tags for tempo). We add noise model for labels, and use DBN-based model (Böck et al. [2016]) for post-processing. Our training dataset consist of

Table 1: Summary of music understanding tasks. Task output lengths are computed for 10 seconds long music input. *Total output sequence length for beat tracking includes additional single tempo tag

| Task | Training dataset size | Temporal resolution | # Outputs per frame | Task output length | Size of tag dictionary |
|---|---|---|---|---|---|
| Beat tracking | 654.6 hours | 25 Hz | 1 | 251* | 303 |
| Chord recognition | 119 hours | 25 Hz | 2 | 500 | 20 |
| Key detection | 678.4 hours | 0.5 Hz | 1 | 10 | 25 |
| Vocal melody | 74.1 hours | 25 Hz | 2 | 500 | 121 |
| All tasks | 1526.1 hours | - | - | - | 469 |

Ballroom (Gouyon [2006]), Beatles (Davies et al. [2009]), Hainsworth (Gouyon et al. [2006]), HJDB (Fujinaga et al. [2012]), RWC (Goto et al. [2002]), Simac (Holzapfel et al. [2012]), Harmonix (Nieto et al. [2019]), SMC (Gouyon [2006]). We also use our in-house dataset, which contains 50k clips and a total size of 580 hours.

**Chord recognition**: The task is further split into recognizing chord root and chord triad recognition: the former is represented by 13 (extra tag for unknown), and the latter by 7 tags. We use pitch shifting as data augmentation technique. Our training dataset is combination of Isophonics (Mauch et al. [2009a]), Billboard (Smith et al. [2011]), HookTheory training set (Donahue et al. [2022]), RWC (Goto et al. [2002]), and USPOP (Berenzweig et al. [2004]).

**Key detection**: A key can be represented by 25 different tags (extra tag for unknown key). The training dataset includes Isophonics (Mauch et al. [2009a]), Billboard (Smith et al. [2011]), HookTheory training set (Donahue et al. [2022]). We also use our in-house dataset, which contains 50k clips and a total size of 580 hours.

**Vocal melody extraction**: The task is further split into predicting notes (represented by 61 tags) and notes onset (60 tags). For training dataset, we use MIR-ST500 (Wang and Jang [2021]), and our in-house dataset, which contains 860 songs with a total size of 50 hours.

## 4 Experiments

**Configuration**: Our LLaMa model uses 8 layers, 16 heads and embedding dimension of 1024. This accounts to the model with 108 million parameters in total. The batch size is set to 32, which is equally splitted between all four task (i.e. single task's batch size is 8). We use AdamW (Loshchilov and Hutter [2017]) as an optimizer (learning rate $= 3e-5$, weight_decay$= 0.1$, Betas $(0.9, 0.98)$ and eps$= 1e-8$). Number of warmup steps is set to 5000. We use GradNorm (with alpha$= 0.5$) to learn task loss weights.

**Evaluation & Baselines**: We use the mir_eval package (Raffel et al. [2014]) for task evaluation. We report *Beat and Downbeat F-Measure* for beat tracking, root accuracy and major/minor weighted accuracy for chord recognition, weighted accuracy (with error tolerance that gives partial credits to reasonable errors) for key detection, and *Onset F1*, *Onset Pitch F1* and *Onset Pitch Offset F1* for vocal melody extraction. As test dataset, we use GTZAN (Marchand et al. [2015]) for beat tracking, HookTheory test set (Donahue et al. [2022]) for chord recognition and key detection, and MIR-ST500 test set (Wang and Jang [2021]) for vocal melody extraction. We use SpecTNT-TCN (Hung et al. [2022]) as a baseline for beat tracking, BTC (Park et al. [2019]) for chord recognition, CNN (Korzeniowski and Widmer [2017]) for key detection and Perceiver Tf (Lu et al. [2023]) for vocal melody extraction.

**Results**: The Table 2 presents the comparison of MusT3 model with the baselines. We trained two version of the model: "Single-Task MusT3" for single task data only, and "Multi-Task MusT3" for all four tasks. Single-Task MusT3 performs much better than the baselines on two tasks: chord recognition and key detection. The results for beat tracking and vocal melody extraction are competitive (the model is better on two of the metrics, but worse on the remaining three). Next, we compare Single-Task MusT3 model with its multi-task equivalent: the results are almost equal with

Table 2: The comparison of MusT3 with the SOTA single-task models. **Bold** font marks the best result, while *italics* marks the second best result (the higher value, the better).

| Model | Beat tracking | Chord recognition | Key detection | Vocal melody extraction |
|---|---|---|---|---|
| SpecTNT-TCN | **0.887** / 0.756 | - | - | - |
| BTC | - | 0.803 / 0.769 | - | - |
| CNN | - | - | 0.732 | - |
| Perceiver TF | - | - | - | **0.798** / **0.777** / 0.490 |
| Single-Task MusT3 | 0.851 / **0.767** | - | - | - |
| Single-Task MusT3 | - | **0.839** / **0.798** | - | - |
| Single-Task MusT3 | - | - | *0.809* | - |
| Single-Task MusT3 | - | - | - | 0.779 / 0.752 / **0.518** |
| Multi-Task MusT3 | *0.852* / *0.764* | *0.829* / *0.789* | **0.810** | *0.787* / *0.761* / *0.513* |

the exception of chord recognition (drop of around 1%). Multi-Task MusT3 still performs much better than the baselines on chord recognition and key detection, and the results for remaining two tasks remain competitive.

**Knowledge sharing between tasks**: In order to investigate whether the model can share knowledge between tasks, we design the following experiment. We train Single-Task MusT3 models on key detection task only, and Multi-Task MusT3 model on key detection and chord recognition tasks. The Figure 1 shows how the performance of both models changes depending on the size of key detection dataset. For four different sizes, chord recognition data improves the results of key detection task. However, the improvement becomes less significant once more data is available.

## 5 Conclusions

In this work, we have developed a novel multi-task model for fine-grained music understanding. Firstly, the model performs better than single-task SOTA baselines on two tasks (chord recognition and key detection), while remaining competitive on another two. Secondly, in controlled experiment, we have shown that the model is capable of reusing knowledge between tasks. Thanks to the unified music-to-tags form and multi-task learning framework, the model can be easily expanded by new tasks, which we aim to do next (in particular, adding musical structure should be beneficial). Furthermore, we plan to scale up our model: the current size of the model is relatively small for the standard of modern deep learning. Finally, we think that the model can yield better results if it's jointly trained with representational learning model.

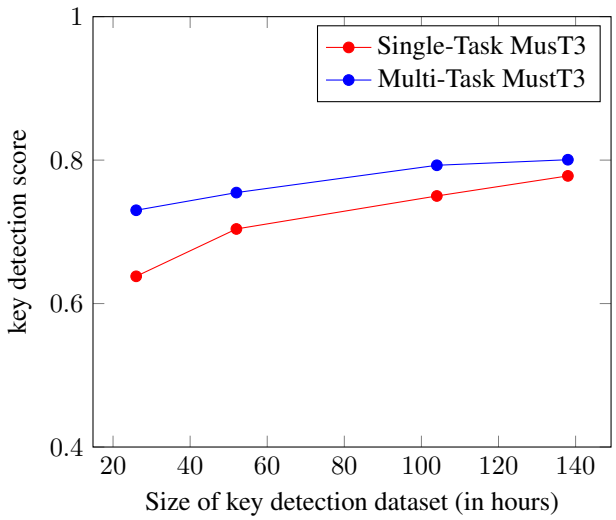


Figure 1: The effect of changing the size of key detection dataset on the performance of Single-Task and Multi-Task MusT3 models.

## References


Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Computer Music Journal*, pages 63–76, 2004.

Deblina Bhattacharjee, Tong Zhang, Sabine Süsstrunk, and Mathieu Salzmann. Mult: An end-to-end multitask learning transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12031–12041, 2022.

Rachel M Bittner, Brian McFee, and Juan P Bello. Multitask learning for fundamental frequency estimation in music. *arXiv preprint arXiv:1809.00381*, 2018.

Sebastian Böck and Matthew EP Davies. Deconstruct, analyse, reconstruct: How to improve tempo, beat, and downbeat estimation. In *ISMIR*, pages 574–582, 2020.

Sebastian Böck, Florian Krebs, and Gerhard Widmer. Joint beat and downbeat tracking with recurrent neural networks. In *ISMIR*, pages 255–261. New York City, 2016.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International conference on machine learning*, pages 794–803. PMLR, 2018.

Kin Wai Cheuk, Keunwoo Choi, Qiuqiang Kong, Bochen Li, Minz Won, Amy Hung, Ju-Chiang Wang, and Dorien Herremans. Jointist: Joint learning for multi-instrument transcription and its applications. *arXiv preprint arXiv:2206.10805*, 2022.

Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *International Conference on Machine Learning*, pages 3915–3924. PMLR, 2022.

Matthew EP Davies, Norberto Degara, and Mark D Plumbley. Evaluation methods for musical audio beat tracking algorithms. *Queen Mary University of London, Centre for Digital Music, Tech. Rep. C4DM-TR-09-06*, 2009.

Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. *arXiv preprint arXiv:2212.01884*, 2022.

Magdalena Fuentes, Brian McFee, Hélène C Crayencour, Slim Essid, and Juan Pablo Bello. A music structure informed downbeat tracking system using skip-chain conditional random fields and deep learning. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 481–485. IEEE, 2019.

Ichiro Fujinaga, Jason Hockman, and Matthew Davies. One in the jungle: Downbeat detection in hardcore, jungle, and drum and bass. 2012.

Josh Gardner, Ian Simon, Ethan Manilow, Curtis Hawthorne, and Jesse Engel. Mt3: Multi-task multitrack music transcription. *arXiv preprint arXiv:2111.03017*, 2021.

Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. Rwc music database: Popular, classical and jazz music databases. In *Ismir*, volume 2, pages 287–288, 2002.

Fabien 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*. Universitat Pompeu Fabra, 2006.

Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Andre Holzapfel, Matthew EP Davies, José R Zapata, João Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9): 2539–2548, 2012.

Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1439–1449, 2021.

Yun-Ning Hung, Yi-An Chen, and Yi-Hsuan Yang. Multitask learning for frame-level instrument recognition. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 381–385. IEEE, 2019.

Yun-Ning Hung, Ju-Chiang Wang, Xuchen Song, Wei-Tsung Lu, and Minz Won. Modeling beats and downbeats with a time-frequency transformer. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 401–405. IEEE, 2022.

Junyan Jiang, Ke Chen, Wei Li, and Guangyu Xia. Mirex 2018 submission: A structural chord representation for automatic large-vocabulary chord transcription. *Proceedings of the Music Information Retrieval Evaluation eXchange*, 2018.

Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2023.

Filip Korzeniowski and Gerhard Widmer. End-to-end musical key estimation using a convolutional neural network. In *2017 25th European Signal Processing Conference (EUSIPCO)*, pages 966–970. IEEE, 2017.

Kyogu Lee and Malcolm Slaney. A unified system for chord transcription and key extraction using hidden markov models. In *ISMIR*, pages 245–250, 2007.

Liwei Lin, Qiuqiang Kong, Junyan Jiang, and Gus Xia. A unified model for zero-shot music source separation, transcription and synthesis. *arXiv preprint arXiv:2108.03456*, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Wei-Tsung Lu, Ju-Chiang Wang, and Yun-Ning Hung. Multitrack music transcription with a time-frequency perceiver. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Ethan Manilow, Prem Seetharaman, and Bryan Pardo. Simultaneous separation and transcription of mixtures with multiple polyphonic and percussive instruments. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 771–775. IEEE, 2020.

Ugo Marchand, Quentin Fresnel, and Geoffroy Peeters. Gtzan-rhythm: Extending the gtzan test-set with beat, downbeat and swing annotations. 2015.

Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Christopher Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler. Omras2 metadata project 2009. In *Proc. of 10th International Conference on Music Information Retrieval*, page 1. Citeseer, 2009a.

Matthias Mauch, Katy C Noland, and Simon Dixon. Using musical structure to enhance automatic chord transcription. In *ISMIR*, pages 231–236, 2009b.

Niko Moritz, Gordon Wichern, Takaaki Hori, and Jonathan Le Roux. All-in-one transformer: Unifying speech recognition, audio tagging, and event detection. In *INTERSPEECH*, pages 3112–3116, 2020.

Oriol Nieto, Matthew McCallum, Matthew EP Davies, Andrew Robertson, Adam M Stark, and Eran Egozy. The harmonix set: Beats, downbeats, and functional segment annotations of western popular music. In *ISMIR*, pages 565–572, 2019.

Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim, and Jonghun Park. A bi-directional transformer for musical chord recognition. *arXiv preprint arXiv:1907.02698*, 2019.

Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, volume 10, page 2014, 2014.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Hongje Seong, Junhyuk Hyun, and Euntai Kim. Video multitask transformer network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, pages 555–560. Miami, FL, 2011.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jun-You Wang and Jyh-Shing Roger Jang. On the preparation and validation of a large-scale dataset of singing transcription. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 276–280. IEEE, 2021.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. *Under review*, 2023.

Yiming Wu, Eita Nakamura, and Kazuyoshi Yoshii. A variational autoencoder for joint chord and key estimation from audio chromagrams. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 500–506. IEEE, 2020.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*, 2020.

Jingwei Zhao, Gus Xia, and Ye Wang. Beat transformer: Demixed beat and downbeat tracking with dilated self-attention. *arXiv preprint arXiv:2209.07140*, 2022.