

MERT-BASED AUDIO-TO-SCORE TRANSCRIPTION WITH A TRANSFORMER DECODER

Antonio Hidalgo-Centeno Carlos Penarrubia Jose J. Valero-Mas
Pattern Recognition and Artificial Intelligence Group, University of Alicante, Alicante, Spain
a.hidalgo@ua.es, carlos.penarrubia@ua.es, jjvalero@dlsi.ua.es

ABSTRACT

Audio-to-Score (A2S) transcription converts polyphonic music recordings directly into complete symbolic scores, including pitches, durations, voices, and score-level notation. We submit an end-to-end A2S system built from a pre-trained music representation model and a Transformer decoder. The acoustic encoder is MERT-v1-330M, a self-supervised model pre-trained on large-scale music data, which is fully fine-tuned on the Quartets training set. Its representations are mapped to `**kern` tokens by an eight-layer Transformer decoder trained from scratch, enabling autoregressive generation of complete multi-spine scores. The system is trained exclusively on the 24,351 excerpts of the Quartets training split, with no additional training data beyond the external pre-training of MERT. On the Quartets test set, the model achieves a Symbol Error Rate of 2.7% and an MV2H score of 95.8%, showing that large-scale music pre-trained representations transfer effectively to A2S transcription. For the MIREX evaluation, the system accepts standard audio files, performs the required audio conversion automatically, and produces one `**kern` score per input recording. The implementation is packaged for reproducible, plug-and-play execution.

1. INTRODUCTION

Audio-to-Score (A2S) transcription converts a recording directly into a score-level representation, estimating not only pitches and onsets but also key and time signatures, clefs, note values, and voices [1]. End-to-end encoder-decoder models avoid the error propagation of multi-stage pipelines [2,3], but they need recordings paired with annotated scores, which are scarce [4,5].

Our MIREX 2026 A2S submission mitigates this scarcity by replacing the convolutional encoder of a state-of-the-art A2S Transformer [1] with the pre-trained music encoder MERT [6], adapted end-to-end on the Quartets dataset. It is the best configuration of a broader study of six pre-trained audio encoders under three adaptation strate-

gies (full fine-tuning, LoRA, and frozen encoder), which will be presented at the main ISMIR 2026 conference.

2. SYSTEM DESCRIPTION

The system is a transfer-learning pipeline composed of a pre-trained encoder E_θ , a layer-aggregation module $a(\cdot)$, and a Transformer decoder D_ϕ that maps an audio recording x to a token sequence $\hat{y}$:

$$\hat{y} = D_\phi(a(E_\theta(x))). \quad (1)$$

During adaptation, θ is initialized from the pre-trained weights and optimized jointly with the randomly initialized ϕ .

Encoder. We use `m-a-p/MERT-v1-330M` [6] (about 315 M parameters), a Transformer encoder with a 1D convolutional front end on raw waveforms. It was pre-trained by its authors on about 160,000 hours of music with a self-supervised masked-prediction objective combining an acoustic teacher based on EnCodec [7] and a musical teacher based on the Constant-Q Transform. Input audio is downmixed to mono and resampled to 24 kHz. Following [8], the hidden states of all encoder layers are combined with a learned weighted sum $a(\cdot)$ and linearly projected to the decoder dimension. *Adaptation strategy:* all encoder parameters are fine-tuned together with $a(\cdot)$ and the decoder (full fine-tuning; neither frozen nor LoRA).

Decoder and output. The decoder is an eight-layer Transformer [9] with four attention heads per layer and cross-attention over the encoder output, trained from scratch following [1]. Scores are encoded in `**kern`, after removing non-musical metadata and graphical directives, which yields a vocabulary of 3,718 tokens. Two special tokens serialize the multi-spine structure: `<COC>` (change of voice) and `<COR>` (progression to the next step). The generated tokens are detokenized into a multi-spine `**kern` file, including its header.

Training. Only the training split of Quartets [1] is used: it contains string-quartet fragments by Haydn, Mozart, and Beethoven with audio synthesized with FluidSynth. The piece-level split yields 24,351 training, 6,089 validation, and 7,611 test excerpts. We use Adam (learning rate 10^{-4} , up to 500 epochs), 10% dropout, and teacher forcing with 20% noisy tokens. Early stopping (patience of 25 epochs) is applied to the validation Symbol Error Rate (SER), and the checkpoint with the lowest validation SER is kept.



System	SER ↓	MV2H ↑	MUSTER ↓
CNN encoder (scratch)	9.6	85.2	43.6
MERT, frozen	14.1	83.3	46.5
MERT, LoRA	5.8	90.3	43.3
MERT, full fine-tuning	2.7	95.8	42.3

Table 1. Quartets test results (%). The submitted system is in bold.

Inference. The entry point `transcribe.py` reads a folder of audio files (`.flac`, `.wav`, `.mp3`, `.ogg`) and writes one `<audio_name>.krn` file per input. The input is automatically adapted to match Quartets (16-bit FLAC, 22,050 Hz, about 30 s), i.e. resampled and converted if needed. Decoding is greedy with the seed fixed to 42, so execution is deterministic and requires no manual intervention.

3. DEVELOPMENT RESULTS

Table 1 reports results on the Quartets test partition. SER is computed on token sequences, MUSTER [10, 11] after conversion to MusicXML, and MV2H [12] after conversion to MIDI. These are development figures, *not* results on the MIREX evaluation set.

Full fine-tuning of MERT reduces SER from 9.6% to 2.7% (about 70% relative) and raises MV2H from 85.2% to 95.8% with respect to a convolutional encoder trained from scratch. A frozen encoder is worse than that baseline and LoRA lies in between, hence the choice of full fine-tuning. MUSTER improves less (43.6% to 42.3%); we hypothesize that this is related to the format conversion it requires, but we have not verified it. The gains hold for every composer: SER drops from 12.1% to 3.9% on Beethoven, from 8.5% to 2.1% on Haydn, and from 8.2% to 2.1% on Mozart. Beethoven remains the most demanding subset, likely due to a larger and sparser token distribution. Since Quartets audio is synthesized, performance on real recordings may differ.

4. RESOURCES AND AVAILABILITY

Table 2 declares the resources required by the task. Code and a Dockerfile (PyTorch 2.7.1, CUDA 12.8) are available at <https://github.com/ahidalgocenteno/a2s-pretrained/tree/mirex-2026-a2s-pretrained>. The checkpoint is hosted at <https://huggingface.co/ahidalgocenteno/a2s-pretrained> and downloaded automatically on first run; offline execution is supported by pre-populating the Hugging Face cache and setting `HF_HUB_OFFLINE=1`.

5. CONCLUSIONS

We presented an end-to-end A2S system that couples a fully fine-tuned MERT-v1-330M encoder with a Transformer decoder trained from scratch, using only the Quar-

Item	Value
Training data	Quartets train split: 24,351 excerpts
Additional data	None; MERT was pre-trained externally by its authors
Parameters	322 M
Training compute	7 d 17 h 1 min 57 s, NVIDIA RTX 5090 (32 GB)
Inference time	6 h 17 min 11 s on the evaluation set, RTX 5090
Peak GPU memory	22.65 GB (inference)
Hardware limits	Within limits (≤ 32 GB VRAM, ≤ 24 h)

Table 2. Resources declaration for the MIREX 2026 A2S task.

tets training split as task data. On Quartets it reaches 2.7% SER and 95.8% MV2H.

6. ACKNOWLEDGMENTS

This research was supported by the Spanish Ministry of Science and Innovation through the LEMUR research project (PID2023-148259NB-I00), funded by MCIU/AEI/10.13039/501100011033/FEDER, EU, and the European Social Fund Plus (FSE+). The second author is supported by the University of Alicante through the FPU Program (UAFPU22-19).

7. REFERENCES

- [1] M. Alfaro-Contreras, A. Ríos-Vila, J. J. Valero-Mas, and J. Calvo-Zaragoza, “A transformer approach for polyphonic audio-to-score transcription,” in *ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 706–710.
- [2] V. Arroyo, J. J. Valero-Mas, J. Calvo-Zaragoza, and A. Pertusa, “Neural audio-to-score music transcription for unconstrained polyphony using compact output representations,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4603–4607.
- [3] W. Zeng, X. He, and Y. Wang, “End-to-end real-world polyphonic piano audio-to-score transcription with hierarchical decoding,” in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 7788–7795, aI, Arts & Creativity.
- [4] A.-M. Christodoulou, O. Lartillot, and A. R. Jensenius, “Multimodal music datasets? challenges and future goals in music processing,” *International Journal of Multimedia Information Retrieval*, vol. 13, no. 3, p. 37, 2024.
- [5] A. Galan-Cuenca, J. J. Valero-Mas, J. C. Martínez-Sevilla, A. Hidalgo-Centeno, A. Pertusa, and J. Calvo-Zaragoza, “Muscat: A multimodal music collection for

automatic transcription of real recordings and image scores,” in *Proceedings of the 32nd ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2024, p. 583–591.

- [6] Y. Li, R. Yuan, G. Zhang, Y. MA, X. Chen, H. Yin, C. Xiao, C. Lin, A. Ragni, E. Benetos, N. Gyenge, R. 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 *International Conference on Representation Learning*, vol. 2024, 2024, pp. 12 181–12 204.
- [7] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, “High fidelity neural audio compression,” *Transactions on Machine Learning Research*, vol. 2023, 2023, publisher Copyright: © 2023, Transactions on Machine Learning Research. All rights reserved.
- [8] Z.-C. Chen, C.-L. Fu, C.-Y. Liu, S.-W. D. Li, and H.-y. Lee, “Exploring efficient-tuning methods in self-supervised speech models,” in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 1120–1127.
- [9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.
- [10] E. Nakamura, E. Benetos, K. Yoshii, and S. Dixon, “Towards complete polyphonic music transcription: Integrating multi-pitch detection and rhythm quantization,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 101–105.
- [11] Y. Hiramatsu, E. Nakamura, and K. Yoshii, “Joint estimation of note values and voices for audio-to-score piano transcription,” in *ISMIR*, 2021, pp. 278–284.
- [12] A. McLeod and M. Steedman, “Evaluating automatic polyphonic music transcription,” in *International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 42–49.