

AUDIO-TO-SCORE TRANSCRIPTION USING A TRANSFORMER DECODER, PRE-TRAINED FEATURES AND DATA AUGMENTATION

Eoin Cummins^{1,2}

Zhuoru Mo^{2,4}

Zhongyi Huang^{2,3}

Yaolong Ju²

¹ University College Dublin, Dublin, Ireland

² Great Bay University, Dongguan, Guangdong, China

³ Guangxi Normal University, Guilin, China

⁴ Shenzhen University, Shenzhen, China

Correspondence: juyaolong@gbu.edu.cn

ABSTRACT

This extended abstract presents our submission to the MIREX 2026 Audio-to-Score (A2S) Transcription task, specifically the blind A2S track. Our model transcribes polyphonic audio into human-readable musical scores using an autoregressive Transformer decoder conditioned on pre-trained MuQ representations. MuQ, a self-supervised music representation model, remains frozen during training. Data augmentation, specifically pitch shifting and time stretching, expands the paired audio–score examples in the Quartets training set to improve generalisation. We use beam search during inference.

1. INTRODUCTION

Audio-to-Score (A2S) transcription aims to convert musical recordings into human-readable scores, supporting music learning, performance, and analysis while reducing the effort required for manual transcription. It can be considered a subfield of Automatic Music Transcription (AMT), but targets complete musical notation rather than the MIDI or piano-roll representations commonly produced by AMT systems [1, 2].

Existing A2S approaches use either modular pipelines, which convert intermediate MIDI or piano-roll representations into scores, or end-to-end models that predict scores directly from audio [3]. Modular approaches can accumulate errors across processing stages, while training end-to-end models is constrained by the scarcity of paired audio–score datasets [4, 5]. Pre-trained music models offer learned representations relevant to transcription tasks such as pitch recognition and beat detection. Data augmentation provides a complementary way to increase training-data diversity and improve generalisation.

Our MIREX 2026 blind A2S submission combines a frozen, pre-trained MuQ music encoder [6] with an au-

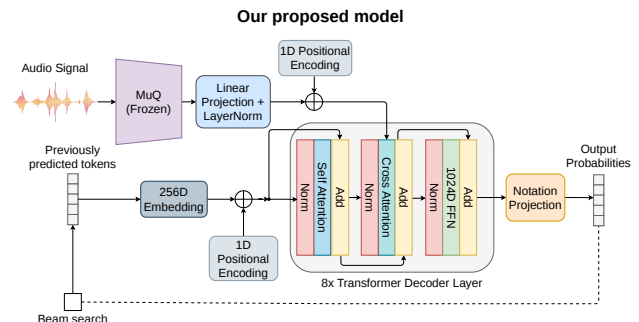


Figure 1. Architecture of our proposed audio-to-score transcription model. A frozen MuQ encoder extracts representations from the input audio, which are projected using a linear layer followed by LayerNorm to produce 256-dimensional representations and combined with positional encodings. Previously generated tokens are embedded and processed by an eight-layer autoregressive Transformer decoder, whose cross-attention conditions token generation on the encoded audio. The notation projection produces next-token probabilities. We use beam search to generate the final score during inference.

to-regressive Transformer decoder to generate scores in `**kern` format [7]. The decoder follows the architecture in [3], except with increased feedforward capacity and a Pre-Norm configuration. Data augmentation through pitch shifting and time stretching expands the paired audio–score examples in the Quartets training set to improve robustness and generalisation. Our setup follows [7] exactly except we use beam search instead of greedy decoding at inference time.

2. METHOD

2.1 Model

Our model depicted in Figure 1 combines a pre-trained audio encoder with an autoregressive Transformer decoder to generate `**kern` scores. We adopt MuQ [6] as our feature extractor, a self-supervised model pre-trained for music understanding. Input audio is resampled to 24,000



Hz, and MuQ is kept frozen. We use MuQ’s final-layer hidden-state sequence. We then use a linear projection followed by LayerNorm to reduce the MuQ embedding size from its original 1024 to 256. Finally, we add one-dimensional sinusoidal positional encodings and pass the resulting sequence to the Transformer decoder as cross-attention memory.

The decoder follows the eight-layer architecture of [3], with two modifications: we increase the feedforward dimension from 256 to 1024 and change the normalisation order from Post-Norm to Pre-Norm to improve training stability [8]. We retain the word-level tokenisation used in the original architecture and generate score tokens autoregressively, conditioned on the encoded audio and preceding tokens.

2.2 Decoding

At inference time, we replace the greedy decoding used in [7] with beam search of width 4. The resulting `**kern` token sequence is detokenised and written to a file.

3. TRAINING

We train on the Quartets dataset [9], which pairs synthesised audio from string quartets by Haydn, Mozart, and Beethoven with `**kern` scores. We use the established split of 70% training, 15% validation, and 15% testing, following [3].

Data augmentation is applied offline to the training split through pitch shifting and time stretching. For each audio–score pair, we generate six augmented variants, corresponding to pitch shifts of $\{-3, -2, -1, 1, 2, 3\}$ semitones. Each variant independently receives a time-stretch factor sampled from $\{0.9, 0.95, 1.05, 1.1\}$. The target score is transposed accordingly, and its tempo metadata is updated to reflect the time stretching. Both audio transformations use Rubber Band.¹ The original examples are retained alongside their augmented variants, resulting in a sevenfold expansion of the training set.

We train with cross-entropy loss using the AdamW optimiser, a constant learning rate of (5×10^{-5}) , weight decay of (10^{-3}) , and a batch size of 8. Early stopping monitors the validation loss with a patience of five epochs and a minimum improvement of (10^{-3}) . We retain the checkpoint with the lowest validation loss.

4. RESULTS AND DISCUSSION

On the Quartets test split under our evaluation setup, our model achieves a symbol error rate (SER) of 4.23% using beam search with width four, compared with the previously reported 4.98% using greedy decoding [7]. Both decoding strategies use the same trained checkpoint. Beam search reduces SER by 0.75 percentage points, corresponding to a relative reduction of 15.1%.

As a supplementary decoding experiment, we also evaluate beam search using the separately trained SheetSage-

A2S model reported in [7] on the SheetSage-A2S test split. It achieves an SER of 20.92%, matching the previously reported greedy-decoding result at two decimal places. Beam search therefore improves the Quartets result, while performance on SheetSage-A2S remains essentially unchanged.

5. REFERENCES

- [1] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, “Automatic Music Transcription: An Overview,” *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.
- [2] F. Jamshidi, G. Pike, A. Das, and R. Chapman, “Machine Learning Techniques in Automatic Music Transcription: A Systematic Survey,” p. 9, 2024, a Preprint. [Online]. Available: <https://arxiv.org/abs/2406.15249>
- [3] M. Alfaro-Contreras, A. Ríos-Vila, J. J. Valero-Mas, and J. Calvo-Zaragoza, “A Transformer Approach for Polyphonic Audio-to-Score Transcription,” in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2024, pp. 706–710.
- [4] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, “Data representations for audio-to-score monophonic music transcription,” *Expert Systems with Applications*, vol. 162, p. 113769, 2020.
- [5] 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 *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2022, pp. 4603–4607.
- [6] H. Zhu, Y. Zhou, H. Chen, J. Yu, Z. Ma, R. Gu, Y. Luo, W. Tan, and X. Chen, “MuQ: Self-Supervised Music Representation Learning With Mel Residual Vector Quantization,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 3653–3664, 2025.
- [7] E. Cummins, Z. Huang, A. D’Hooge, Z. Mo, and Y. Ju, “Audio-to-score transcription using pre-trained features, data augmentation, and the new sheetsage-a2s dataset,” *arXiv preprint arXiv:2608.06165*, 2026.
- [8] R. Xiong, Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, Y. Lan, L. Wang, and T. Liu, “On layer normalization in the transformer architecture,” in *Proceedings of the 37th International Conference on Machine Learning*, 2020, pp. 10 524–10 533.
- [9] M. A. Román, A. Pertusa, and J. Calvo-Zaragoza, “A Holistic Approach to Polyphonic Music Transcription with Neural Networks,” in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2019, pp. 731–737.

¹ <https://breakfastquay.com/rubberband/>