Proposal v3

Tyler Jones

May 2025

1 Abstract

This project explores Automatic Music Transcription (AMT), the task of converting audio recordings into symbolic representations such as sheet music or guitar tablature. While AMT has seen significant progress, especially with the rise of deep learning, accurately transcribing polyphonic music remains a persistent challenge [1, 2]. This work builds upon recent breakthroughs like the Onsets and Frames model [3] and Transformer-based architectures [7, 5], both of which significantly enhance transcription accuracy by modeling temporal and spectral patterns more effectively. The project aims to design a system that takes audio input, detects note onsets and frequencies, and outputs corresponding musical notes and MIDI files. Leveraging `librosa` and `CREPE` for signal analysis and `music21` for symbolic representation, the system will be evaluated against existing benchmarks in piano transcription. Future directions include incorporating self-supervised learning techniques [6] to reduce the need for annotated data and expand generalization across instruments and musical styles.

2 Introduction

Automatic Music Transcription (AMT) is the process of converting an audio recording into a symbolic musical representation, such as sheet music, a MIDI file [1], or tablature. It stands at the intersection of signal processing, music theory, and machine learning, and has broad applications in music education, digital archiving, interactive composition, and accessibility for the hearing impaired [2].

Despite decades of research, AMT remains a challenging task, particularly in polyphonic contexts where multiple notes occur simultaneously. Human listeners can identify melodies, harmonies, and rhythms with ease due to their knowledge of musical structure and timbre, but replicating this perceptual ability computationally requires robust models of pitch, timing, and instrument characteristics [4].

Recent advances in deep learning have significantly improved transcription accuracy. Models like Onsets and Frames [3] introduced architectures capable of jointly modeling the onset, offset, and pitch of notes using convolutional and recurrent layers. More recently, Transformer-based approaches have shown

promise by capturing long-range temporal dependencies and better modeling music as a sequence generation task [5, 7].

This project aims to design a streamlined AMT system focused on monophonic and simple polyphonic audio inputs. Using `librosa` and `CREPE` for audio analysis and `music21` for symbolic output, the system will extract note onsets and fundamental frequencies, map them to musical pitches, and export both sheet music and MIDI representations. In addition to leveraging supervised learning techniques, this project also explores the potential of self-supervised methods to reduce dependence on labeled datasets [6].

3 Related Work

3.1 Classical Approaches to Music Transcription

Early work in automatic music transcription relied heavily on signal processing techniques such as the short-time Fourier transform (STFT), constant-Q transform (CQT), and harmonic-percussive source separation (HPSS) to detect note onsets and estimate fundamental frequencies [4]. These classical methods typically assumed idealized conditions, such as monophonic input or isolated instrument recordings, which limited their applicability to real-world music.

Hand-crafted features like spectral flux and pitch salience were often used to inform note event detection, but they struggled in noisy environments or complex polyphonic textures [**?**]. While these systems laid important groundwork, they lacked the flexibility and adaptability of modern machine learning models.

3.2 Deep Learning in AMT

The introduction of deep learning marked a significant turning point in AMT research. Neural networks enabled models to learn hierarchical representations of audio signals, eliminating the need for hand-crafted features. Notably, the Onsets and Frames model by Hawthorne et al. [3] proposed a dual-objective architecture that separates onset detection from frame-wise pitch tracking, significantly improving transcription accuracy for piano recordings.

More recent developments have focused on integrating attention mechanisms and sequence modeling. Transformer-based models, such as those developed under Google's Magenta project, have shown strong performance on complex musical passages by modeling music as a structured sequence of tokens [5]. These models are able to attend to long-range dependencies and have demonstrated generalization across instruments and styles.

3.3 Self-Supervised and Unsupervised Learning

Labeling large datasets of transcribed music is both time-consuming and expensive. To address this, recent work has explored self-supervised learning paradigms that enable models to learn from unlabeled audio. Techniques such as BYOL (Bootstrap Your Own Latent) and contrastive predictive coding have

been adapted to the audio domain with promising results [6]. These methods provide a foundation for building more robust, general-purpose transcription systems that are less dependent on labeled training data.

3.4 MIDI and Symbolic Output

A significant challenge in AMT is converting detected pitch and timing information into meaningful symbolic representations. Libraries like `music21` have made it easier to programmatically generate and manipulate Western music notation and MIDI files [**?**]. However, aligning these symbolic outputs with expressive performance elements (e.g., tempo rubato, articulation) remains an active area of research.

3.5 Positioning This Project

This project integrates foundational signal processing techniques (via `librosa`) with modern transcription tools to produce clean symbolic outputs from simple monophonic and homophonic recordings. While not as complex as fully end-to-end neural transcription systems, this project aims to strike a balance between interpretability and automation, providing a baseline system that can be iteratively improved with more sophisticated learning techniques. The emphasis on exporting both MIDI and sheet music further positions the project as a practical, usable tool for music learners, educators, and hobbyist composers.

4 Risks and Contingencies

While this project is designed to be modular and scalable, several specific risks could affect its successful completion.

4.1 Audio Onset Detection Inaccuracy

Accurate onset detection is foundational for identifying note events. If the `librosa` onset detection function fails to detect onsets in complex or softly articulated passages, the transcription output may be misaligned or incomplete.

Contingency: I will evaluate multiple onset detection methods (e.g., spectral flux, energy-based detection) and allow for adjustable onset sensitivity thresholds in the implementation. Annotated debugging visualizations will also be used to refine detection.

4.2 Ambiguity in Pitch Estimation

Estimating fundamental frequencies from polyphonic or noisy audio can result in octave errors or spurious harmonics, especially in non-piano instruments or heavily reverberant recordings.

Contingency: I will begin with monophonic and clean homophonic examples to validate baseline performance, and apply harmonic masking or spectral

subtraction techniques to reduce interference. An interactive manual correction interface could also be explored if time permits.

4.3 Alignment of Transcription to Notation or MIDI

Even with accurate pitch and onset detection, aligning note events into readable sheet music or MIDI output may result in quantization errors or poor rhythmic interpretation (e.g., swing vs. straight rhythms).

Contingency: I will implement adjustable quantization levels and make use of `music21`'s built-in tools for rhythm analysis and rewriting. I'll also test outputs on both standard and syncopated inputs to assess generalizability.

4.4 Computational Load or Latency

Real-time transcription is outside the project's scope, but even offline transcription could suffer from long processing times on large audio files or high-resolution spectrograms.

Contingency: I will precompute and cache intermediate representations (e.g., CQT, onset envelope) where possible, and benchmark performance on sample files early in the timeline to inform optimization.

4.5 Difficulty in Evaluating Transcription Quality

Ground truth labels for musical transcription are hard to obtain outside of curated datasets. Without a clear standard of accuracy, evaluation may be subjective.

Contingency: I will use small manually annotated audio excerpts and compare generated MIDI output against ground truth. Additionally, I'll collect qualitative feedback by having musicians review selected outputs for musicality and correctness.

5 Timeline

This project, in addition to the semester I have already spent developing its foundational components, will span the summer prior to the 2025–2026 academic year and continue through the following two semesters.

Over the summer, I will focus on evaluating the feasibility of training a machine learning model specifically for pitch detection in polyphonic music. This will involve initial experiments using existing datasets and pre-trained models, as well as testing training configurations and model architectures to determine whether a custom model would outperform existing solutions.

Depending on the outcome of this feasibility study, I will begin developing a dataset tailored to the needs of automatic music transcription. This dataset will be composed of audio-MIDI pairs created by generating sheet music in MuseScore, exporting the corresponding MIDI files, and recording complementary

guitar tracks in Reaper using direct input (DI). Recording via DI will allow me to manipulate the audio post-recording using amp simulators such as Neural DSP's Archetype Tim Henson or John Petrucci plugins. This setup will enable a wide range of pitch, distortion, and ambient configurations without requiring multiple takes, thereby streamlining the dataset creation process.

Once the dataset is complete and has been processed for optimal compatibility with AMT tasks, I will begin developing the core signal processing components of the system. These components will include onset detection and frequency estimation algorithms, which will be implemented and iteratively refined by adjusting parameters and evaluating their performance against ground truth data.

Subsequently, I will implement a system for assigning note onset and offset times as well as rhythmic values. This system must be both accurate and flexible, as it will form the basis for interpreting the audio data in musically meaningful terms. Once note values and rhythms are reliably extracted, I will use the `music21` library to generate MIDI files that represent the transcribed music.

With a functional transcription pipeline in place, I will proceed to the fine-tuning phase, during which I will perform parameter optimization and performance benchmarking. This phase will ensure the system is not only accurate but also efficient and scalable.

If time permits, I will focus on enhancing the usability and visualization of the system, potentially by creating a basic user interface or providing graphical outputs that help visualize the transcription process and results.

References

[1] Emmanouil Benetos, Simon Dixon, Zhiyao Duan, and Sebastian Ewert. Automatic music transcription - an overview. *Journal of Music Information Retrieval*, 2019.

[2] Chirag Bhatt and Manish Thakkar. Automatic music transcription: challenges and future directions. *Journal of Intelligent Information Systems*, 44(3):565–578, 2014.

[3] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1802.06182*, 2018.

[4] Anssi P. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 2004.

[5] Google Magenta. Transcription with transformers, 2021.

[6] Daisuke Niizumi, Daiki Takeuchi, and Yuki Mitsufuji. Byol for audio: Self-supervised learning for general-purpose audio representation. *arXiv preprint arXiv:2110.14449*, 2021.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in NeuAal Information Processing Systems (NeurIPS)*, 2017.