Automatic Music Transcription Using Artificial Intelligence

Advisors
Yung-hsiang Lu & Yeon Ji Yun

Ojas Chaturvedi, Kayshav Bhardwaj, Shreeya Sarurkar, Emily Li, Junyong Lee, Arnav Kalekar

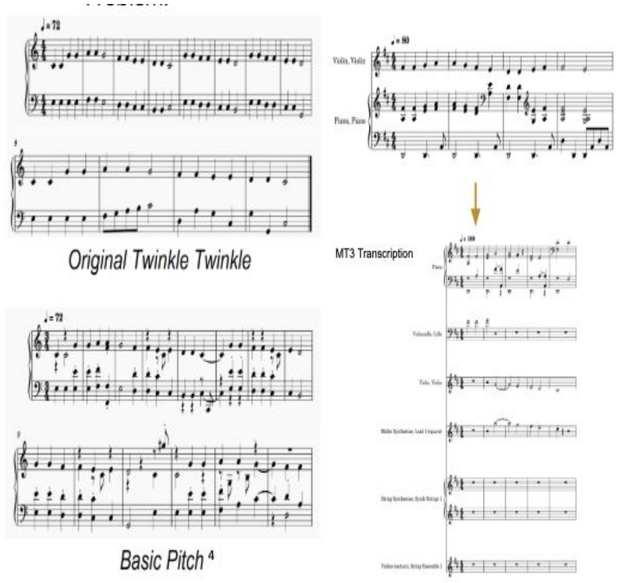
Abstract

Automatic Music Transcription (AMT) converts audio into symbolic formats like sheet music or MIDI. Traditionally time-intensive for human transcribers, AMT can accelerate music creation, education, and research by saving time and enhancing accessibility. Beyond music, its applications extend to assistive hearing technologies, such as addressing the "cocktail party problem" by isolating individual sound sources from noise. Techniques that distinguish instruments by acoustic traits could similarly separate voices, improving hearing aids and speech recognition in complex auditory settings.

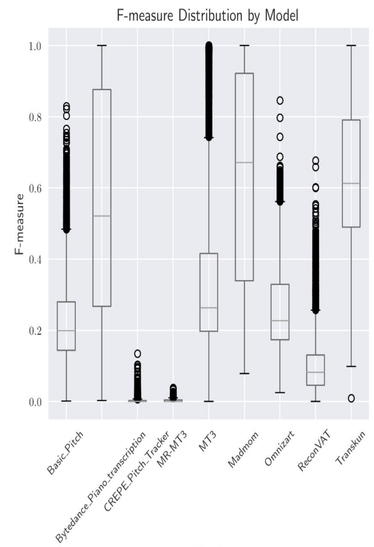
To explore AMT's current capabilities, we conducted a literature review of existing software, identifying Google's MT3 model as among the most promising. However, many tools struggle with polyphonic or multi-instrumental music, hallucinate nonexistent instruments, or fail to isolate melodies. In response, we organized an AMT competition in April 2025, where participants submitted programs that converted up to 100 classical recordings (with at most three instruments) into MIDI files. Submissions, hosted on GitHub, were evaluated on instrument accuracy, pitch, and onset/offset precision.

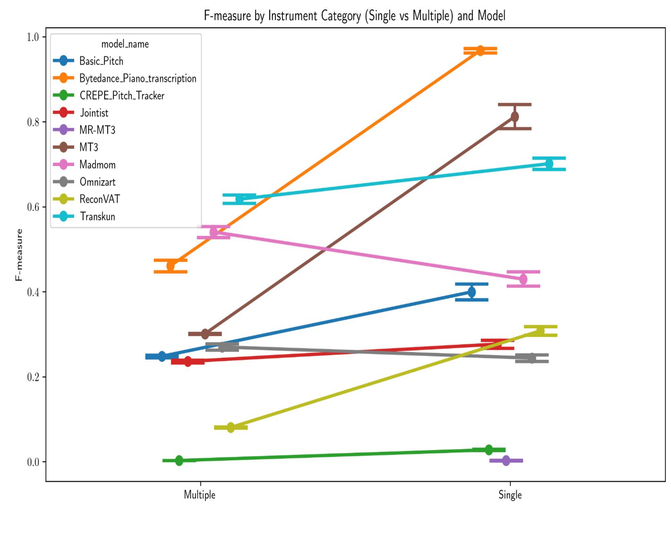
Each model excelled in specific areas but lacked generality. Some differentiated instruments well yet misidentified pitches, while others captured rhythm accurately but confused instrument types. Future work aims to combine these strengths into a more robust model. One approach under exploration is a model-switcher, which dynamically selects transcription models based on the music's characteristics. Ultimately, incorporating features like musical complexity could further improve accuracy and adaptability. In the future, the progress made in music transcription will not only impact the music industry, but improve overall quality of audio and speech transcription.

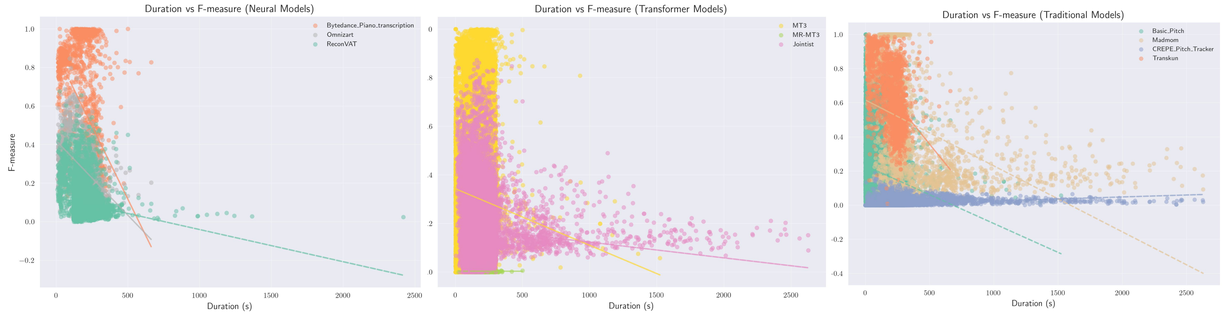
Existing Model Outputs



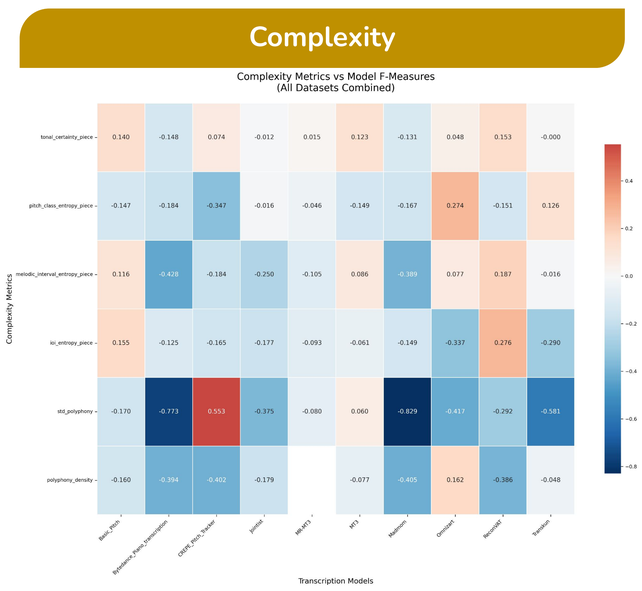
Model Metrics

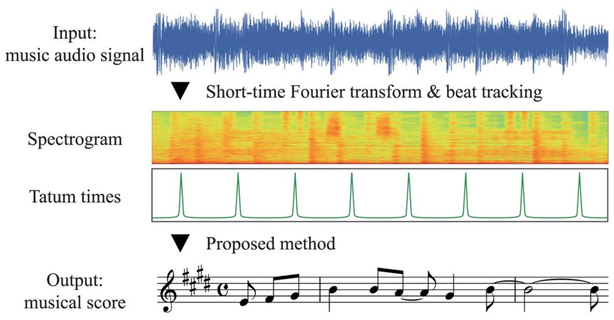


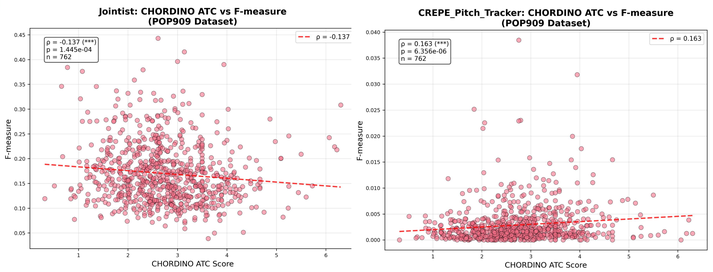




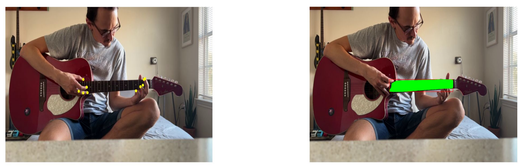
Complexity

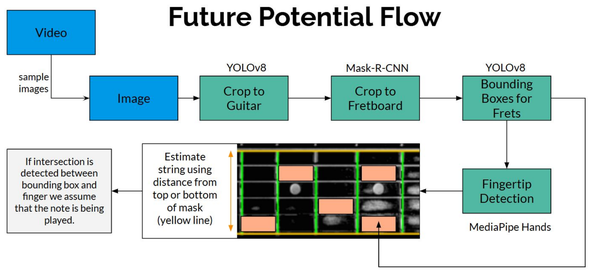






Computer Vision





Acknowledgements

1. Nishikimi, Ryo et al. "Audio-to-Score Singing Transcription Based on a CRNN-HSMM Hybrid Model." APSIPA Transactions on Signal and Information Processing 10 (2021): e7. Web.
2. Gardner, Josh, et al. "MT3: Multi-task multitrack music transcription." arXiv preprint arXiv:2111.03017 (2021).
3. Wu, Yu-Te, et al. "Omnizart: A general toolbox for automatic music transcription." arXiv preprint arXiv:2106.00497 (2021).
4. Bittner, Rachel M., et al. "A lightweight instrument-agnostic model for polyphonic note transcription and multipitch estimation." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

Thank you to the NSF for funding the AI in Music group and their projects, including the Automatic Music Transcription. Additionally, thank you to the advisors of this project, Professor Yung-hsiang Lu and Professor Yeon Ji Yun.