(1) A clear **problem definition:** what are you doing and why, what are the expected inputs and the outputs

(2) A high-level **overview of the model:** you don't need to present specifics of the architecture for example, but do define components you plan to use and how those will connect with one another

(3) A **description of the dataset** you will use for the task: is it labeled? how big? show sample of data)

(4) A description of how will you **evaluate the performance:** what is the error metric, why is it sensible)

(5) Breakdown of the **timeline** and who on the team would do what

-

(6) What **literature** will you read or have read: just a list, you don't need to go through it in any detail

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9222310
https://arxiv.org/pdf/2010.01815.pdf

PROPOSAL STARTS ON THE NEXT PAGE

CPSC 532S Project Proposal

The Problem

Automatic Music Transcription (AMT) is a task involving converting audio data into music notation [2]. This is a difficult problem due to uncertainty caused by complicated multi instrumental music and noisy lower fidelity recordings. There are also many qualities in music that we could measure such as rhythm (which is the pulse of the music), pitch (which is the frequency of the note), dynamics (which is the volume and the change in volume) and timbre (the perceived sound quality of a music note). These qualities can be hard to measure, especially since human musicians are often imprecise. For example, flutes often sound sharp (which is slightly higher than their normal pitch) or flat (which is slightly lower than their normal pitch) based on the changing temperature of the room. Rhythm is also often more felt than precisely notated and timing can be inconsistent. Existing AMT models tend to focus on single instruments [2] [4] [5] while most music in general uses multiple instruments. We seek to develop neural network models to transcribe polyphonic (multiple notes played simultaneously) and multi-instrumental music into sheet music using a neural network architecture using attention on a large dataset. Our model will take in raw waveform audio data, convert it into a representation that can be parsed by our model such as spectrogram, and then output MIDI data (digital representations of note occurrences) which can then be used to generate sheet music. We seek to develop a reliable and accurate music transcriber that could be used by music teachers and music students for music education purposes.

Datasets

There are two labelled datasets we plan to train and test on, each with slightly different goals. The first is the MAESTRO dataset [7], which is composed of 200 hours of piano performances, and has both the audio waveforms and finely aligned MIDI labels. MIDI contains the necessary information to recreate each musical note played by the pianist either digitally or in human-readable form. The information includes onset and offset data (when the key was pressed and when the note duration), pitch/frequency information (which note was actually pressed), as well as velocity (the volume of the note over time). The MAESTRO dataset is very commonly used in piano transcription works [1] [4] [5].

Canonical Composer	Canonical Title	Split	Duration (sec)
Alban Berg	Sonata Op. 1	train	759.51847125
Alexander Scriabin	24 Preludes Op. 11, No. 13-24	train	872.640588096
Alexander Scriabin	5 Preludes, Op.15	validation	400.557825938

Table 1. Sample metadata from the MAESTRO dataset [7]

Since we also want to evaluate our model on multi-instrumental music, we will also be using the MusicNet dataset [6], which is composed of 330 classical music recordings, totalling 34 hours. There are both solo and ensemble recordings of twelve instruments (piano, violin, viola, clarinet, etc.). For every note, this dataset also stores the onset and offset data and pitch/frequency, as well a label for which instrument plays the note. Velocity measurements are unfortunately not included.

Composer	Composition	Movement	Ensemble	Source	Duration (sec)
Mozart	String Quartet No 14 in G major	1. Allegro vivace assai	String Quartet	European Archive	356
Mozart	Clarinet Quintet in A major	4. Allegretto con variazioni	Clarinet Quintet	William McColl	591
Bach	WTK I, No. 19: Prelude and Fugue in A major	1. Prelude	Solo Piano	Kimiko Ishizaka	91

Table 2. Sample metadata from the MusicNet dataset [6]

Model Overview

We plan to use elements from the most state of the art piano transcription model [8] and combine some elements from recent developments in multi-instrument transcription and instance segmentation. The model from [8] currently takes in a mel-spectrogram translated from raw audio and has four acoustic model blocks to calculate four aspects (note onset, note offset, frame classification, and velocity) needed to properly generate MIDI data. Each block is built using convolution and biGRU layers. The outputs are then combined and fed into additional biGRU layers to generate the final note onset and frame classification predictions.

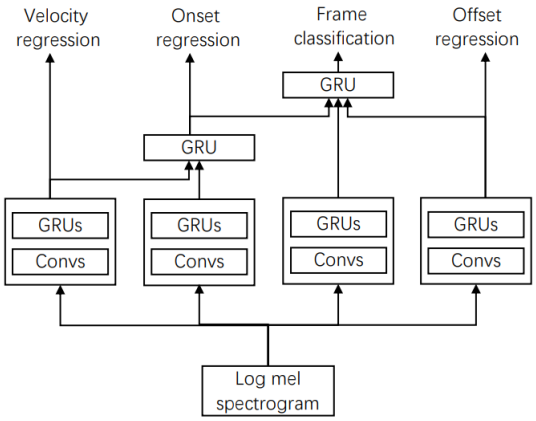

Figure 1. High-level diagram of piano transcription model [8]

We want to add improvements introduced by multi-instrument models to this architecture [1] [5]. [1] in particular addresses the additional problem of identifying the instrument for every note, which is beyond the scope of this project, but innovation from this architecture may still be useful in helping our model be more flexible to sounds from different instruments. We want to replace the biGRU layers with self-attention (or use a combination of biGRU and attention), as it is believed self-attention would be better able to capture longer-range dependencies in the frequency data [1], something which is especially important when you are working multiple instruments. If this is successful, we may also want to add additional convolutional encoding layers before the acoustic models, inspired by the multi-instrument models [1] [5], as it may help capture information about instrument-specific sounds that can be shared by all the acoustic models. Since the MusicNet dataset does not include velocity measurements [6], our model will not have the velocity acoustic model. Velocity measurements are ultimately not necessary for production of sheet music, but may be useful in predicting the onset of a note [6]. Time permitting, we may also want to explore conditioning the model further on predicted instrument identification data, generated by a pre-trained model [3].

While the original piano transcription model converts the raw audio waveform to mel-spectrograms as model input [8], multi-instrument models convert it to a similar data representation called Combined Frequency and Periodicity (CFP) [1] [5], which we will be using. Both mel-spectrograms and CFP have input data dimensions KxN, where N is time-step bins, and K is some form of frequency bins [1] [8].

Evaluation

We will evaluate the MIDI output with frame and note evaluation with onset, offset and velocity. We use several evaluation metrics: Precision, Recall, and F1. F1 is a better measure than just using accuracy since the dataset is highly imbalanced. For example, there is a lot more silence in the dataset than non zero pitch values [1]. Despite the fact that F1 is computed using Precision and Recall, it is still useful to output Precision and Recall. For example, high precision could be desired as it would eliminate unwanted noise while higher recall would be desired for

editing the transcription since deleting undesired notes might be easier than adding in undetected notes [1]. We will use these metrics to compare the state of the art piano transcription model [8], and multi-instrument model [1].

Timeline

Gantt Chart

Task	Task Name	Start Date	End Date	% Completed	Member/s	3/15/2021	3/19/2021	3/23/2021	3/27/2021	3/31/2021	4/4/2021	4/8/2021	4/12/2021	4/16/2021	4/20/2021		
1	**Literature Review of automatic music transcription**	3/15/2021	3/22/2021	30%	James/Jacques/Aritro												
2	**Identify what modifications can be done to existing work**	3/15/2021	3/22/2021	30%	James/Jacques/Aritro												
3	**Data processing**	3/22/2021	3/31/2021	0%	James												
4	**Model implementation (training + testing)**	4/1/2021	4/12/2021	0%	Jacques/Aritro												
5	**Report**	4/13/2021	4/20/2021	0%	James/Jacques/Aritro												

Literature

1) Wu, Y. T., Chen, B., & Su, L. (2020). Multi-Instrument Automatic Music Transcription With Self-Attention-Based Instance Segmentation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 2796-2809.

2) Benetos, E., Dixon, S., Duan, Z., & Ewert, S. (2018). Automatic music transcription: An overview. *IEEE Signal Processing Magazine*, *36*(1), 20-30.

3) Hung, Y. N., & Yang, Y. H. (2018). Frame-level instrument recognition by timbre and pitch. *arXiv preprint arXiv:1806.09587*.

4) Hawthorne, C., Elsen, E., Song, J., Roberts, A., Simon, I., Raffel, C., ... & Eck, D. (2017). Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*.

5) Wu, Y. T., Chen, B., & Su, L. (2019, May). Polyphonic music transcription with semantic segmentation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 166-170). IEEE.

6) Thickstun, J., Harchaoui, Z., & Kakade, S. (2016). Learning features of music from scratch. *arXiv preprint arXiv:1611.09827*.

7) Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C. Z. A., Dieleman, S., ... & Eck, D. (2018). Enabling factorized piano music modeling and generation with the MAESTRO dataset. *arXiv preprint arXiv:1810.12247*.

8) Kong, Q., Li, B., Song, X., Wan, Y., & Wang, Y. (2020). High-resolution piano transcription with pedals by regressing onsets and offsets times. *arXiv preprint arXiv:2010.01815*.

Office hours notes: