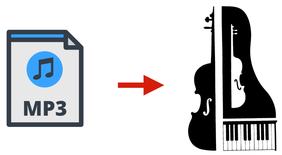
Human Inspired Music Compression through Note Transcription

Zachary Hoffman, Shabhum Chandak, Tsachy Weissman
Department of Electrical Engineering Stanford University

Motivation

Big Picture:

- Music files as compared to other forms of media can take up significant storage
- Especially when stored in lossless formats such as .WAV files
- Whether it is:
  - a music repository,
  - video game soundtrack,
  - personal music library,
  - or music sharing,
- There is a need for music that comes in smaller form factors that maintain the same level of quality



Background information:

- Music creators and children playing on iPads have already been using a widely accepted file format that work to solve this problem
- MIDI (Musical Instrument Digital Interface) is a simple file format used on applications such as Apple's Garage band
- MIDI represents audio in its composite notes and instruments
- It is easily edited and comes in a small form factor
- Because a MIDI reader serves as a decoder, audio quality can be as high or low as the recipient wants given the same MIDI file
- So how do we take a continuous digital audio signal and convert it to its composite parts?

Strategy for Music Compression

What is "Human Satisfaction?"

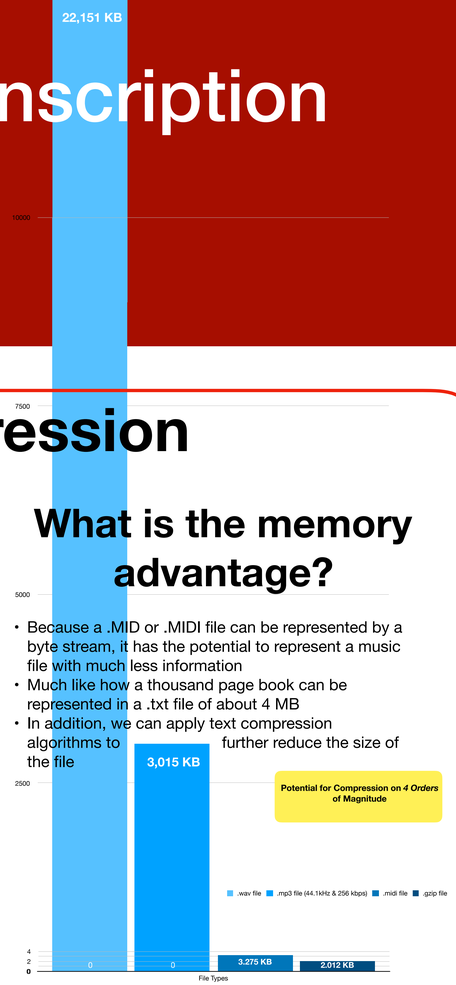
- Because music functions as a media to be consumed by people, the most important metric to evaluate the efficacy of this codec is with human discretion
- Here "Human Satisfaction" refers to:
  - a combination of the perceived quality of the audio file in comparison to the original .wav (uncompressed) file
  - in addition to the accuracy of notes and style of play for the recreated file
- Because this is a highly subjective metric, success for this codec is not universal or singularly represented by one figure

What is a MIDI file?

- A MIDI file is a byte stream that contains the information for a score of music.
- This includes:
  - instrument number(s),
  - pitch,
  - velocity,
  - and duration of each note
- It is a widely used file format for music creation and transcription playback.

What is the memory advantage?

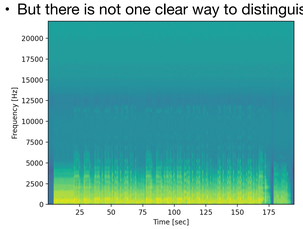
- Because a .MID or .MIDI file can be represented by a byte stream, it has the potential to represent a music file with much less information
- Much like how a thousand page book can be represented in a .txt file of about 4 MB
- In addition, we can apply text compression algorithms to further reduce the size of the file



Challenges

Training:

- When training, the first step to retrieve the notes would be a simple Fourier transform
- Below is a spectrogram that illustrates the Fourier transform on a piece of piano audio used in my training dataset
  - Where warmer colors represent higher amplitude
  - And the vertical axis are the frequencies
- However, instruments have undertones, often harmonic with the note played
- As a result, we see discrete chunks at multiple frequencies at any given time interval
- But there is not one clear way to distinguish



magenta

- Also as a continuous function there is no clear distinction between the onset of a note and its duration

A possible solution:

- A Convolutional Neural Network would be a natural solution
- Google's Magenta project based off of open source TensorFlow models includes a model that can account for the onsets and frames of a piano
- It does so by merging two parallel networks: one for recognizing the onset of a note and the other for the duration
- The next step is to train the model on different instruments and have it be able to differentiate different instruments

A Human-centric Problem:

- The next issue arises with our training methodology
- Magenta's "onsets and frames" model was made to transcribe piano music in a similar manner and purpose as human scribes
- However, our goal orbits around the goal of a satisfied listener
- Therefore, when training the model on new instruments there is no "mean-square-error-esque" metric for which to convolve the model.

An Imperfect Workaround:

- This led me to having to manually determine which model yielded the best codec to satisfy our human
- Here rises an issue with our "Human Satisfaction" metric
- More work is needed to determine more concrete metrics for which one can train a model so that it is practical to work on a range of music that do not exclusively include:
  - Multiple instruments
  - Synthetic instruments
  - Vocals

Results

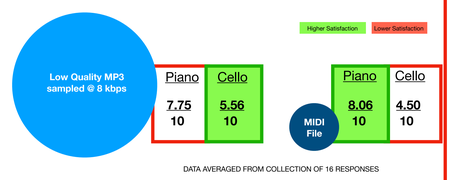
Methodology:

- After determining a good candidate for a cello model 30 second clips were collected: high quality mp3, low quality mp3, and MIDI to mp3 file
- Human respondents were asked to rank their "satisfaction" with the recreation of the song
  - comparing the original, high quality, mp3 to the the low quality mp3
  - And comparing the original mp3 to the transcribed MIDI file
- Satisfaction was ranked on a scale of 0 to 10 - where 0 is least satisfied and 10 is most satisfied

Results:

- The Low Quality MP3 represented the absolute smallest form of the file that is currently conventional
- Although it still is larger than the MIDI file by an order of 3 magnitudes
- Despite the size difference the perceived satisfaction of the audio files were quite similar between the two file formats
- Indeed the Magenta Piano model actually exceeded the diminished quality MP3 file
- This leads me to believe that at least for simple music, with few instruments and no vocals, a transcribed music codec is a substantial improvement on memory and potentially on the quality of a compressed (non .wav) file

Average Human Satisfaction Rating for Low Quality MP3 vs Transcribed MIDI



Needed Improvements and Next Steps

- The Human Satisfaction survey provided feedback that the cello model needs improvement, asking of us to look for a better model or different training parameters
- Although a MIDI file represents a substantially smaller file format, it is only good so far as a means for file transfer because other than Garageband and Audacity and other similar applications, most personal devices cannot handle the file format, requiring a multi-step decoder on both ends
- Once again, the next step is to work on a multi-instrument model

References

Cello Cliparts #49168 [Photograph]. (n.d.). http://clipart-library.com/clipart/27502.htm

Freepik. (n.d.). *Mp3 free icon* [Photograph]. https://www.flaticon.com/free-icon/mp3_180828

Having to wait a long time for a file to download on dialup [Photograph]. (n.d.). https://www.reddit.com/r/nostalgia/comments/9vqsyj/having_to_wait_a_long_time_for_a_file_to_download/

Make Music and Art Using Machine Learning. (n.d.). Magenta. https://magenta.tensorflow.org/

Stanford | ENGINEERING
Electrical Engineering