

New York University

Automatic Piano Music Transcription

GitHub: <https://github.com/shriyakalakata/automatic-piano-music-transcription>

Argy Sakti, Sean Wiryadi, Shriya Kalakata

May 07, 2024

Executive Summary

Problem Statement: Automatic music transcription, which involves converting audio signals into musical notation, is a challenging task due to the complexities involved in representing and analyzing audio data.

Goal: develop a machine learning model capable of predicting the activations of MIDI notes from an audio file containing a musical piece.

Technical Challenges

- Representing audio data in a suitable format
- Capturing long-range dependencies and musical intricacies
- Large number of classes (88 MIDI notes)
- Imbalanced data

Solution Approach

- Deep neural network (DNN) with dropout and early stopping
- Long Short-Term Memory (LSTM) networks
- Experimented with transfer learning

Value/Benefit

- Facilitate computational musicology and music information retrieval
- Enable applications like score generation, music education, and audio-to-score conversion

Related Works

Music Transcription Using Deep Learning

Luoqi Li
EE, Stanford University
lluoqi@stanford.edu

Isabella Ni
SCPD CS, Stanford University
nizhongyi@stanford.edu

Liang Yang
GS, Stanford University
lyang6@stanford.edu

Abstract—Music transcription, as an essential component in music signal processing, contributes to wide applications in musicology, accelerates the development of commercial music industry, facilitates the music education as well as benefits extensive music lovers. However, the work relies on a lot of manual work due to heavy requirements on knowledge and experience. This project mainly examines two deep learning methods, DNN and LSTM, to automatize music transcription. We transform the audio files into spectrograms using constant Q transform and extract features from the spectrograms. Deep learning methods have the advantage of learning complex features in music transcription. The promising results verify that deep learning methods are capable of learning specific musical properties, including notes and rhythms.

Keywords—automatic music transcription; deep learning; deep neural network (DNN); long short-term memory networks (LSTM)

and LSTM (Long Short-Time Memory), were built for realizing audio-to-score conversion. Acoustic signal recording polyphonic piano music in the format of .wav file is inputted into our algorithm, and the output is MIDI (Musical Instrument Digital Interface) file that can be easily converted into music score.

II. RELATED WORKS

AMT has been attempted since the 1970s and polyphonic music transcription dates to the 1990s [4]. Unsupervised learning approaches assume that prior knowledge is not required for music transcription. Non-negative matrix factorization (NMF), which is in some sense equivalent to Independent Components Analysis (ICA), was performed on the input magnitude spectrum to obtain the frequency spectra for each pitch and their corresponding activities in time domain [5] [6]. By learning enough examples, a basis set including spectra for all the pitches could be created. However, the learned dictionary matrix may not match perfectly with music notes, which causes interpretation problem at the output

10th International Society for Music Information Retrieval Conference (ISMIR 2009)

EVALUATION OF MULTIPLE-F0 ESTIMATION AND TRACKING SYSTEMS

Mert Bay Andreas F. Ehmann J. Stephen Downie
International Music Information Retrieval Systems Evaluation Laboratory
University of Illinois at Urbana-Champaign
{mertbay, aehmann, jdownie}@illinois.edu

ABSTRACT

Multi-pitch estimation of sources in music is an ongoing research area that has a wealth of applications in music information retrieval systems. This paper presents the systematic evaluations of over a dozen competing methods and algorithms for extracting the fundamental frequencies of pitched sound sources in polyphonic music. The evaluations were carried out as part of the Music Information Retrieval Evaluation eXchange (MIREX) over the course of two years, from 2007 to 2008. The generation of the dataset and its corresponding ground-truth, the methods by which systems can be evaluated, and the evaluation results of the different systems are presented and discussed.

1. INTRODUCTION

A key aspect of many music information retrieval (MIR) systems is the ability to extract useful information from

the MIREX [3] organized a multi-F0 evaluation task. This task can be considered as an evolution and superset of the previous MIREX audio melody extraction tasks. For more information on audio melody extraction, we refer the reader to [13]. The MIREX multiple-F0 task consists of two sub-tasks built around the two pitch representations mentioned earlier. The first subtask is called *Multiple-F0 Estimation* (MFE). In MFE, systems are required to return a list of active pitches at fixed time steps (analysis frames) of a polyphonic recording. The second subtask is called *Note Tracking* (NT). In the NT subtask, systems are required to return the note F0, onsets and offsets of note events in the polyphonic mixture, similar to a piano-roll representation. The MIREX multiple-F0 task attracted many researchers from around the world. In the 2007 MFE subtask, there were a total of 16 algorithms from 12 labs. For the NT subtask, there were 11 algorithms from 7 labs. In 2008, there were a total of 15 algorithms from 10 labs for MFE

Automatic Music Transcription: An Overview

Emmanouil Benetos *Member, IEEE*, Simon Dixon, Zhiyao Duan *Member, IEEE*, and Sebastian Ewert *Member, IEEE*

I. INTRODUCTION

The capability of transcribing music audio into music notation is a fascinating example of human intelligence. It involves perception (analyzing complex auditory scenes), cognition (recognizing musical objects), knowledge representation (forming musical structures) and inference (testing alternative hypotheses). *Automatic Music Transcription (AMT)*, i.e., the design of computational algorithms to convert acoustic music signals into some form of music notation, is a challenging task in signal processing and artificial intelligence. It comprises several subtasks, including (multi-)pitch estimation, onset and offset detection, instrument recognition, beat and rhythm tracking, interpretation of expressive timing and dynamics, and score typesetting. Given the number of subtasks it comprises and its wide application range, it is considered a fundamental problem in the fields of music signal processing and music information retrieval (MIR) [1], [2]. Due to the very nature of music signals, which often contain several sound sources (e.g., musical instruments, voice) that produce one or more concurrent sound events (e.g., notes, percussive sounds) that are meant to be highly correlated over both time and frequency, AMT is still considered a challenging and open problem in the literature, particularly for music containing multiple simultaneous notes¹ and multiple instruments [2].

The typical data representations used in an AMT system are illustrated in Fig. 1. Usually an AMT system takes an audio waveform as input (Fig. 1a), computes a time-frequency representation (Fig. 1b), and outputs a representation of pitches

IV-F, as well as methods for transcribing specific sources within a polyphonic mixture such as melody and bass line.

A. Applications & Impact

A successful AMT system would enable a broad range of interactions between people and music, including music education (e.g., through systems for automatic instrument tutoring), music creation (e.g., dictating improvised musical ideas and automatic music accompaniment), music production (e.g., music content visualization and intelligent content-based editing), music search (e.g., indexing and recommendation of music by melody, bass, rhythm or chord progression), and musicology (e.g., analyzing jazz improvisations and other non-notated music). As such, AMT is an enabling technology with clear potential for both economic and societal impact.

AMT is closely related to other music signal processing tasks [3] such as audio source separation, which also involves estimation and inference of source signals from mixture observations. It is also useful for many high-level tasks in MIR [4] such as structural segmentation, cover-song detection and assessment of music similarity, since these tasks are much easier to address once the musical notes are known. Thus, AMT provides the main link between the fields of music signal processing and symbolic music processing (i.e., processing of music notation and music language modeling). The integration of the two aforementioned fields through AMT will be discussed in Section IV.

Given the potential impact of AMT, the problem has also

MT3: MULTI-TASK MULTITRACK MUSIC TRANSCRIPTION

Josh Gardner,[†] Ian Simon, Ethan Manilow,[‡] Curtis Hawthorne, Jesse Engel
Google Research, Brain Team

ABSTRACT

Automatic Music Transcription (AMT), inferring musical notes from raw audio, is a challenging task at the core of music understanding. Unlike Automatic Speech Recognition (ASR), which typically focuses on the words of a single speaker, AMT often requires transcribing multiple instruments simultaneously, all while preserving fine-scale pitch and timing information. Further, many AMT datasets are “low-resource”, as even expert musicians find music transcription difficult and time-consuming. Thus, prior work has focused on task-specific architectures, tailored to the individual instruments of each task. In this work, motivated by the promising results of sequence-to-sequence transfer learning for low-resource Natural Language Processing (NLP), we demonstrate that a general-purpose Transformer model can perform multi-task AMT, jointly transcribing arbitrary combinations of musical instruments across several transcription datasets. We show this unified training framework achieves high-quality transcription results across a range of datasets, dramatically improving performance for low-resource instruments (such as guitar), while preserving strong performance for abundant instruments (such as piano). Finally, by expanding the scope of AMT, we expose the need for more consistent evaluation metrics and better dataset alignment, and provide a strong baseline for this new direction of multi-task AMT.¹

What have others done?

Representing audio data with numerical representation

- **Non-negative matrix factorization (NMF)**, decomposing into basis components representing frequency spectra for each pitch and their activations in time
- **Constant Q-Transform (CQT)**, to represent audio signal by decomposing the signal into frequency components over time, with a time domain that is segmented through a frame-based approach

Model Architectures

- Recurrent Neural Networks
- Deep Neural Networks
- Long Short-Term Memory

Model Evaluation

- Frame-based Evaluation
- Note-based Evaluation

Limitations

- With NMF, the learned dictionary matrix may not match perfectly with music notes, which causes interpretation problems at the output
- Scarce data on annotations of ground-truth music transcriptions
- Overlapping sound events incites challenges with harmonics overlap in frequency

Method/Approach

Data Preprocessing & Normalization

- Represented audio using **Constant-Q Transform (CQT)**
- **Normalized CQT** vectors within the limits of the training set
- **Divided the audio into frames** of 32 ms (frame-based approach)
- **Aligned CQT** vectors with MIDI annotations using **one-hot encoding** representing note activations

Regularization/Optimization Techniques

- **Dropout** - reduce overfitting
- **Early stopping**
- **Cyclical learning rates** (Max LR 0.1, Min LR 0.0001)
- **Mini Batch Size** (Tradeoff between Accuracy and Efficiency)

Model Architectures

- **Baseline Logistic Regression**
- **Deep Neural Network (DNN)**
- **Long Short-Term Memory (LSTM)**
- Used **binary cross-entropy loss** for multi-label classification
- Employed **Adam optimizer** for training

Implementation/Experimentation

Model Architectures

- Deep Neural Network (DNN)
 - Explored 1, 2, 3, 4 hidden layers with ReLU activations
- Long Short-Term Memory (LSTM)
 - Attempted transfer learning by initializing weights from pre-trained DNNs

Hyperparameter Tuning

- Grid search for optimal hyperparameters
- Tuned dropout rates: 0.05, 0.15, 0.25
- Tuned minibatch sizes: 250, 500, 1000, 1500
- Best configuration: dropout = 0.05, minibatch = 1500
- (for LSTM): window size

Results – Baseline Model

Baseline Logistic Regression

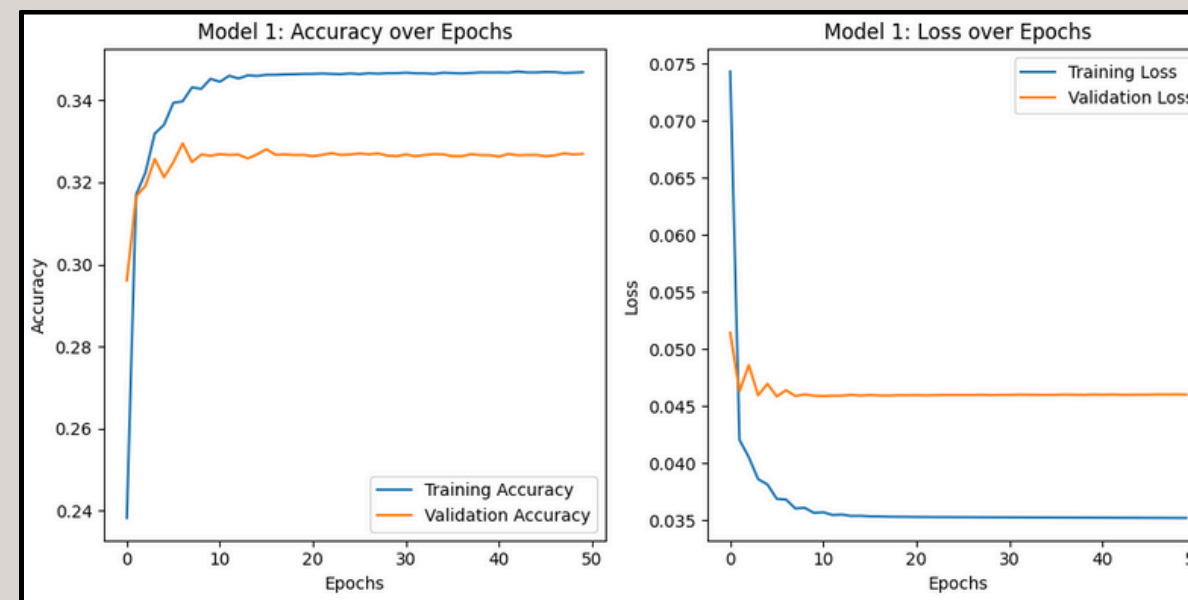
- Implemented a baseline model to serve as a comparison to our neural networks model
- **Accuracy = 10.75%**
- Extremely high recall of **85.17%**, at the cost of a low **precision of 10.96%**, indicates that the model overpredicts on note activations

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
Log Regression	<u>10.75</u>	10.96	85.17	19.41

$$Accuracy (Modified) = \frac{TP}{(TP + FP + FN)}$$

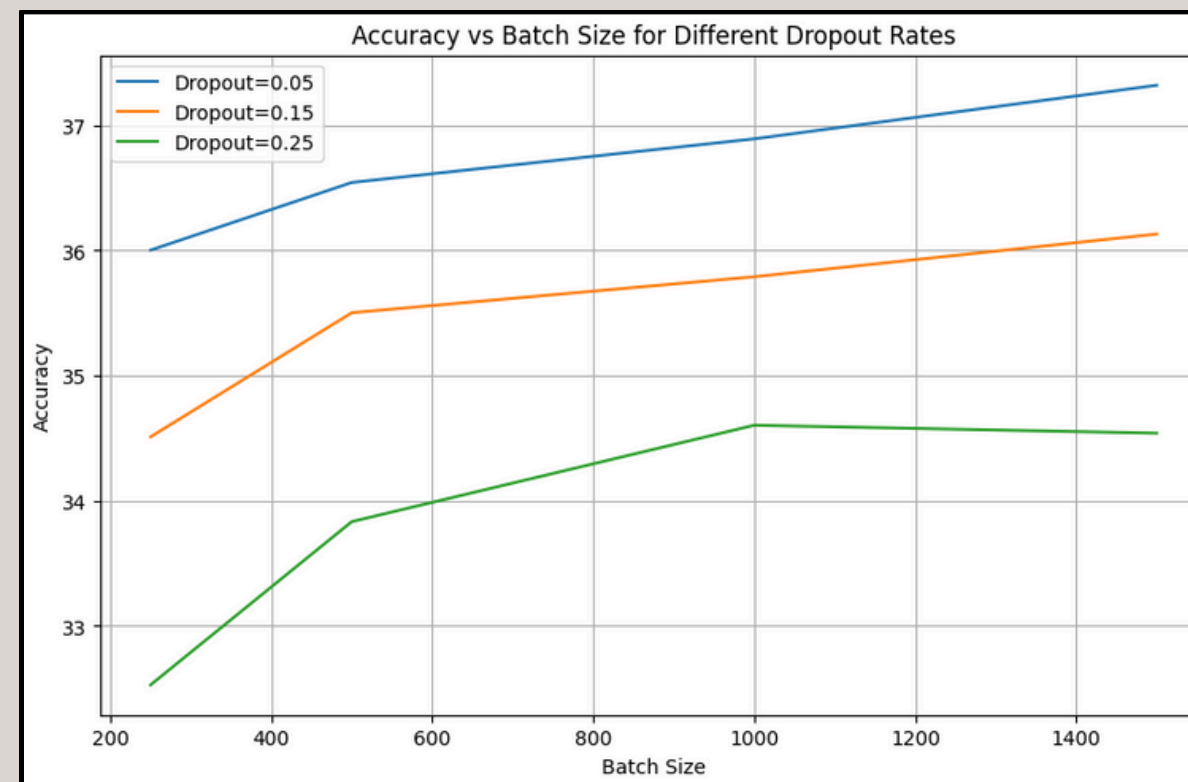
Results – DNN Models

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
DNN 1 Layer Base	<u>37.01</u>	68.64	44.54	54.02
DNN 2 Layer Base	35.67	65.13	44.10	52.59
DNN 3 Layer Base	34.53	64.65	42.56	51.33
DNN 4 Layer Base	30.89	64.37	37.26	47.20
DNN 1 Layer Dropout 0.05 Batch 1500	<u>37.32</u>	69.91	44.46	54.36
DNN 1 Layer Tuned	37.06	69.92	44.09	54.08
DNN 2 Layer Tuned	<u>37.59</u>	67.20	46.03	54.64
DNN 3 Layer Tuned	36.58	67.13	44.56	53.57
DNN 4 Layer Tuned	35.40	66.93	42.90	52.28

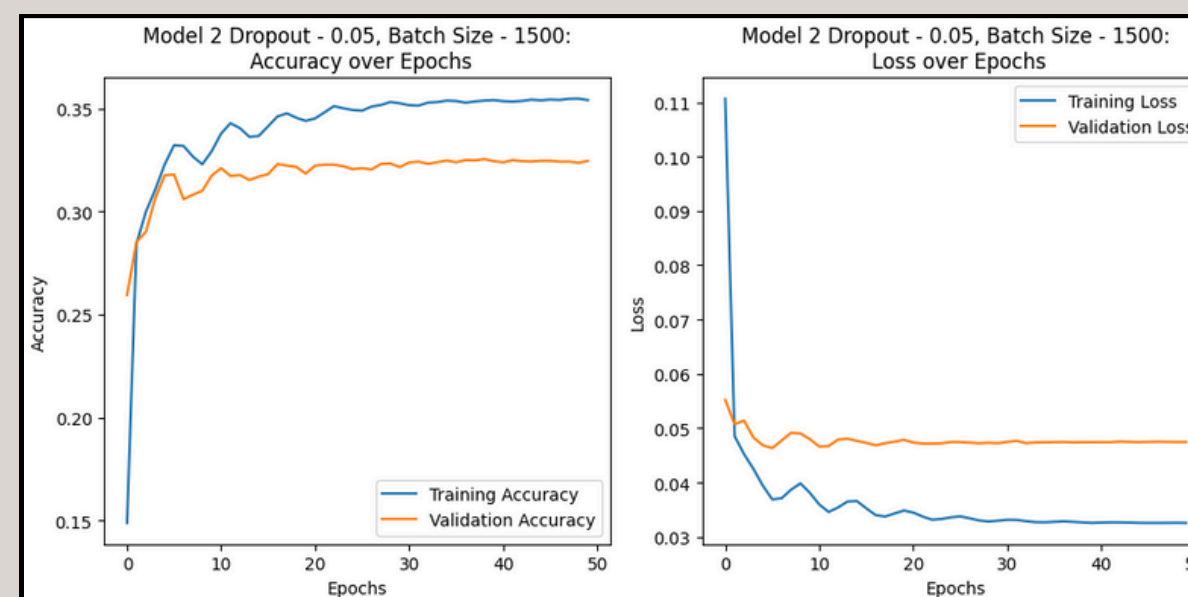


Performance of best
“baseline” DNN
Model (1 layer)

Accuracy = 37.01%



Finetuning Dropout
Rate and Batch Size
based on best
“baseline” DNN
Model



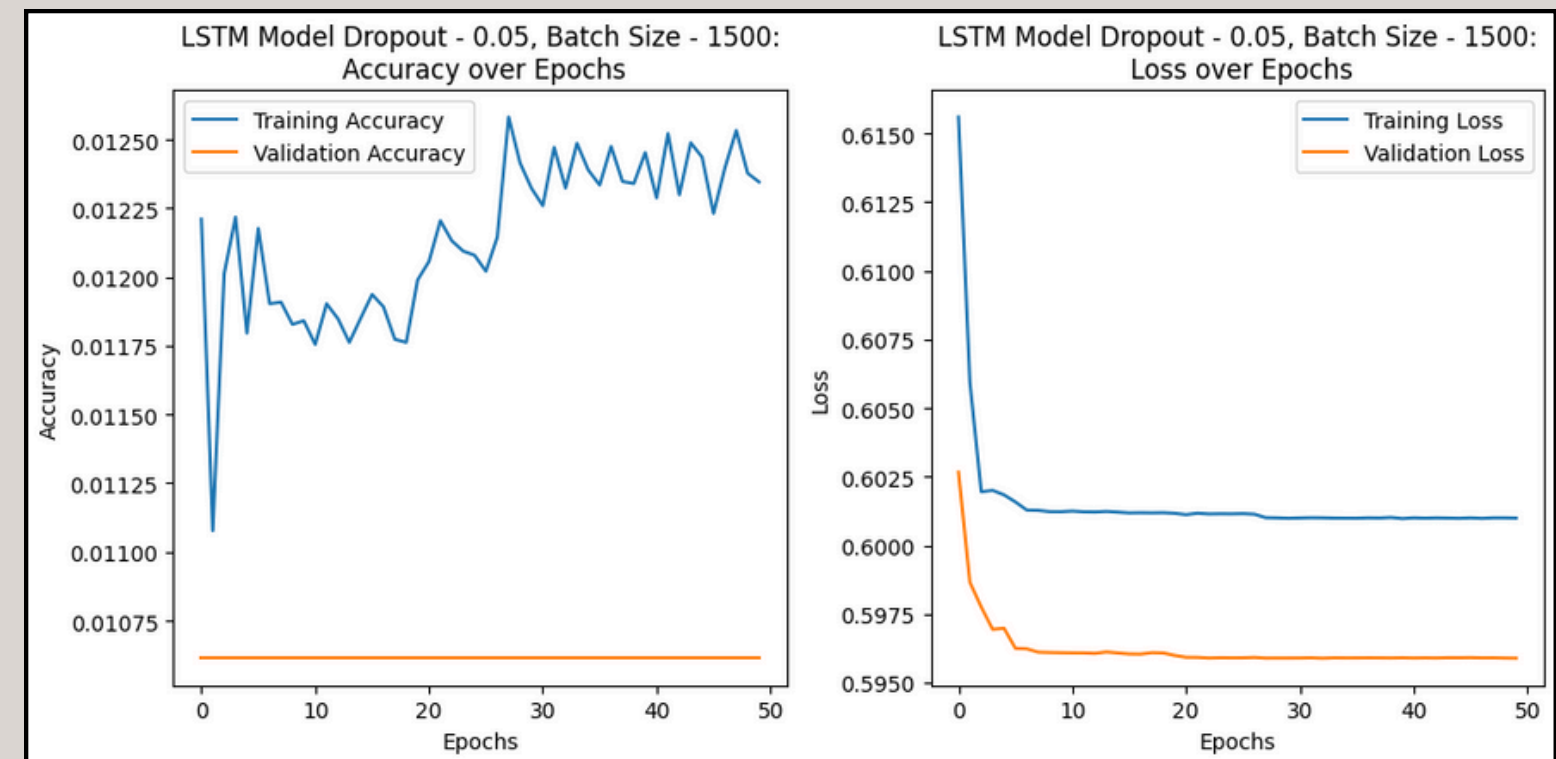
Performance of best
tuned model
(2 layers) with best
hyperparameter
configuration

Accuracy = 37.59%

Results – LSTM

- Poor Performance: 2.67%
- Weight Sharing/Transfer Learning (Weights from best performing DNN)
- Issues:
 - Data Reshaping
- Optimization Methods:
 - Learning Rate
 - Hidden Layers
 - Window Size
- Improvements:
 - Investigate reshaping
 - Bidirectional LSTM

Model Name	Accuracy (%)	Precision (%)	Recall (%)	F1-Score
LSTM	<u>2.67</u>	-	-	-



LSTM Model Performance

Conclusion

- Automatic piano transcription is challenging due to complexities in converting audio to notation
- Explored deep neural networks (DNNs) and LSTMs to capture patterns and long-range dependencies

Future Works:

- Attention-based models and transformers
- Investigate data augmentation and representation learning techniques
- Leverage multi-modal approaches combining audio with symbolic/visual data

Conclusion

- Limitations in achieving high accuracy
- Exploration of architectures and techniques provided valuable insights
- Need for innovative approaches to effectively capture audio information and translate to notation

New York University

Thank You

Argy Sakti, Sean Wiryadi, Shriya Kalakata