INDIAN INSTITUTE OF TECHNOLOGY, BOMBAY

Tabla transcription using DNN

by

Pranav Sankhe

Supervisor: Prof. Preeti Rao

January 2019

0.1 Abstract

The tabla is a percussion instrument of the Indian subcontinent, consisting of a pair of drums, used in traditional, classical, popular and folk music. The audio of tabla is composed of many modalities. We wish to develop an end to end transcription system for tabla audio and to reconstruct the audio from the obtained transcription. We also intend to study the presence of (acoustic-prosodic) correspondences between the recitation and playing of tabla compositions like intensity and F0 variations, at the bol/stroke level (sort of like word-level in speech) and identify those that are well-correlated. In particular for music education, it would be useful to provide a rating and to identify the nature of the error in the learning process.

Recently data driven signal processing approaches have been gaining momentum to solve problems like the one elaborated above. The use of machine learning to solve complex non linear problems have been dominating the speech and music research which is evident in the recently published papers. In particular, we have explored the use of NMF (Non-Negative Matrix Factorization) to solve the problem of drum transcription for drum-only-recordings. While this approach works significantly well, having seen the developments in DNN models for Piano transcription, we are motivated to employ a similar architecture for tabla transcription.

There are few challenges in implementing a neural network architecture for tabla transcription which we will elaborate over in later sections. We intend to find a way around these challenges and come up with a suitable and robust transcription and evaluation system.

Contents

Chapter 1

Challenges

Here we enumerate over the challenges we face in developing a dual objective transcription system for percussion instruments.

- Interference of multiple instruments
- Playing Techniques
- Recording Conditions and Post Production
- Insufficient real world datasets

1.1 Interference of multiple instruments

The superposition of various instruments makes the recognition of a specific instrument difficult due to the overlaps in both spectral and temporal domain.

Percussive Instruments:

A basic drum kit includes drums of different sizes and well- distinguishable timbral characteristics. In a more advanced setup for studio recordings, similar drums with subtle variations in timbre often appear, resulting in sounds that are harder to differentiate. In the case of tabla, we have two indvidual tablas which the player uses to compose music. Seperating and identifying the audio from these two tablas is a research problem in its own right. This problem of interference is more severe when these sounds occur simultaneously.

Melodic Instruments:

The wide range of sounds produced from a drum kit or a tabla can potentially coincide

with sound components of many melodic instruments. Pre-processing steps for suppression of melodic instruments have been proposed but haven't been able to achieve substantial improvement.

1.2 Playing Techniques

Playing techniques is an important aspect of expressive musical performance. For drum instruments, these techniques include basic rudiments as well as timbral variations.

Approaches to model playing techniques:

- A study on the automatic identification of timbral variations of the snare drum sounds induced by different excitations has been done where a classification task is formulated to differentiate sounds from different striking locations (center, halfway, edge, etc.) with different excitations (strike, rim shot, and brush)
- The discrepancy between more expressive gestures on a larger dataset with combinations of different drums, stick heights, stroke intensities, strike positions, and articulations has been explored
- Different playing techniques for cymbal sounds have been investigated too. The Differentiation is based on either the position where the cymbal is struck (bell, body, edge), how a hi-hat is played (closed, open, chick), or other special effects such as choking a cymbal with the playing hand.

All of these studies showed promising results in classifying the isolated sounds, however, when the classifier is applied to the real-world recordings, the performance dropped drastically

1.3 Recording Conditions and Post Production

In practice, it is likely that we have to deal with convolutive, time-variant, and non-linear mixtures instead of linear superpositions of single drum sounds. The acoustic conditions of the recording room and the microphone setup lead to reverberation effects that might be substantial The recording engineer will likely apply equalization and filtering to the microphone signal Mostly, the resulting signal alterations can be modeled as convolution with one or more impulse responses. Non-linear effects such as dynamic compression and distortion might be applied to the drum recordings.

Consequences:

- Any methods involving machine-learning might deteriorate if the training data does not match the target data.
- Any methods involving decomposition based on a linear mixture model might be affected when the observed drum mixtures violate these basic assumptions.

A possible strategy to counter these challenges might be data augmentation. In our case the amount of training data could be greatly enhanced by applying diverse combinations of audio processing algorithms including reverberation, distortion and dynamics processing.

1.4 Insufficient real world datasets

Size: The most common issue of all the existing drum transcription datasets is the insufficient amount of data. Since these datasets are created under very different conditions, they can- not be easily integrated into one large entity. Given that tabla is an Indian classical instrument, the dataset of tabla recordings is limited in size.

Complexity: The existing datasets have the tendency of over-simplifying the ADT problem. Only the drum sequences with basic patterns are presented in the dataset.

Diversity: Most of these datasets do not cover a wide range of music genre and playing style. The limitation in terms of diversity can hinder the system's capability of analyzing a wider range of music pieces. Particularly, the lack of any singing voice in the corpora ENST-Drums and IDMT-SMT-Drums indicates their insufficiency. Tests on tracks containing singing voice revealed that this poses a big problem, especially for RNN-based ADT methods.

Homogeneity: Since each dataset is most likely to be generated under fixed conditions the audio files within the same dataset tend to have high similarities.

Chapter 2

Data Augmentation

Data Augmentation in the aspect of images have been considerably researched on, as againt for music. There are few papers which we will be following and will also actively ponder on other possible augmentations.

2.1 Paper 1

A set of augmentations can be implemented by considering the constrain preserving the labels. In line with recent research in speech recognition, Schluter Et al., observed that pitch shifting happens to be the most helpful augmentation method. Combined with time stretching and random frequency filtering, they achieved a reduction in classification error between 10 and 30, reaching the state of the art on two public datasets. Here's the link to their paper in PDF format: Click Here.

2.1.1 Data-independent Methods

A way to increase a model's robustness is to corrupt training examples with random noise. We consider dropout setting inputs to zero with a given probability and additive Gaussian noise with a given standard deviation. This is fully independent of the kind of data we have, and we apply it directly to the mel spectrograms fed into the network.

2.1.2 Audio-specific Methods

Pitch shifting and time stretching the audio data by moderate amounts does not change the label for a lot of MIR tasks. We implemented this by scaling linear-frequency spectrogram excerpts vertically (for pitch shifting) or horizontally (for time stretching).

A much simpler idea focuses on invariance to loudness: We scale linear spectrograms by a random factor in a given decibel range, or, equivalently, add a random offset to log-magnitude mel spectrograms.

As a fourth method, we apply random frequency filters to the linear spectrogram. Specifically, we create a filter re- sponse as a Gaussian function.

2.2 Paper 2

McFee Et al. have implemeted a software framework in `Python` for data augmentation for music which remains to be explored in depth but we believe it will provide valuable tools to carry out augmentation efficiently. Here's their paper regarding the same in PDF format: Click Here

Chapter 3

Datasets

Tabla Solo Dataset:
The Tabla Solo Dataset is a transcribed collection of Tabla solo audio recordings spanning compositions from six different Gharanas of Tabla, played by Pt. Arvind Mulgaonkar. The dataset consists of audio and time aligned bol transcriptions.

Mridangam Stroke Dataset
The Mridangam Stroke dataset is a collection of 7162 audio examples of individual strokes of the Mridangam in various tonics. The dataset comprises of 10 different strokes played on Mridangams with 6 different tonic values.

Mridangam Tani-avarthanam dataset: In Carnatic music, Tani-avarthanam is the solo performance by the percussion ensemble following the main piece of the concert. [No time labels]

Datasets remain to be explored extensively.

Chapter 4

Architecture

Here's a brief overview of the ideas of interest derieved from the Magenta's Transcription system for piano recordings. Acute details remian to be added and we plan to do that anon.

Model

In all the previous attempts at ADT, all the frames are considered independent (except RNN) and are given equal importance which is not the case. The frame which contains the onset is more important than the other frames and this inspired to add an onset detector to the system

Data Representation

The input data is the mel spectrogram of the raw audio signal. This raw audio signal consists of all the drums (as well as other instruments) playing together. I wish to explore the possibility of employing BSS methods on audio signal to identify the individual drum sources. After separating the individual drum sources, we can construct a n-channel input for our model. The motivation to separate sources is the huge decrease in accuracy upon introduction of other instruments.