**CS:4980, Deep Learning**

**NIMA HAMIDI**

*Project Proposal*

**AUTOMATIC MUSIC STRUCTURE ANALYSIS USING DEEP NEURAL NETWORK**

**BUILDING A MODEL THAT TAKES A RECORDED MUSICAL EXAMPLE AND EXTRACTS THE FORMAL STRUCTURE OF THE PIECE, UTILIZING NEURAL NETWORK.**

Over last couple of years I have been working on an automatic tool to analyze music in a temporal form. In order to understand the concept, I will describe the traditional approaches of music analysis.

**Music Analysis:**

There is a long tradition of analyzing music to understand the segmentation, considering the perception of time. Very simple form of A -B -A is a piece of music that is consist of three main statements or musical segments. The first statement “A” is followed by a contrasting idea “B”. The “A” section comes back at the end to conclude the musical expression. This way of analysis can go further and deconstruct a complex piece of music into several segments. The temporal relationship between segments define unique auditory experience.

A

A

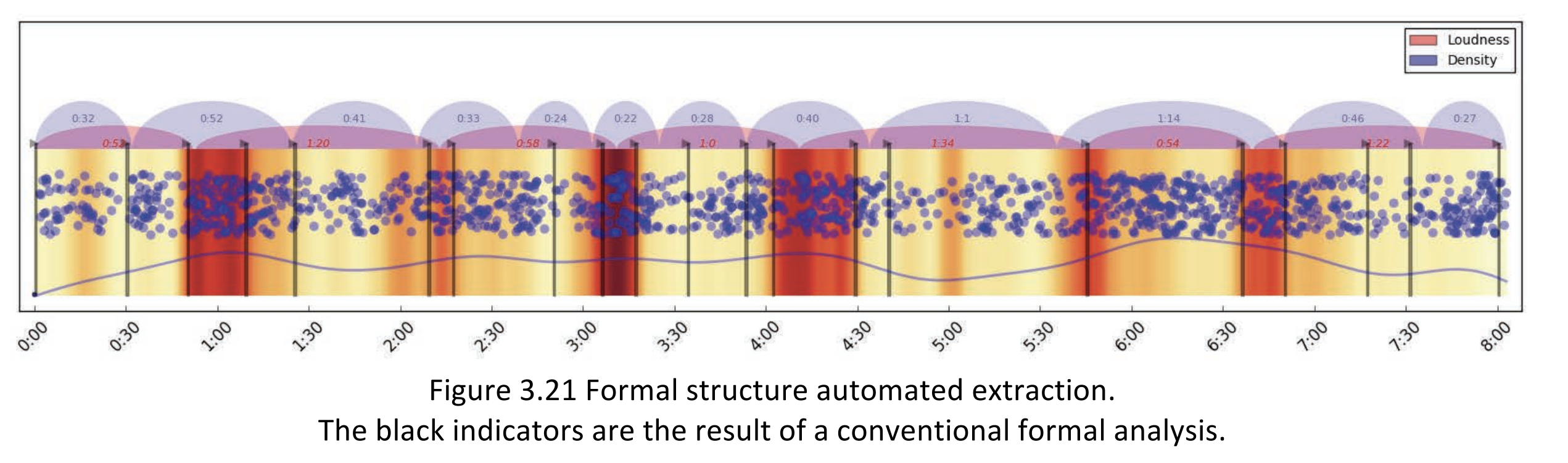
B

*A piece of music in the form of A-B-A*

Traditionally there are a large number of high level feature in order to come up with such a segmentation. Harmony, phrasing, voicing, instrumentation, dynamic, pitch structure, and rhythm are some of this musical feature. Extracting some of these features are very challenging and eventually not producing the best result.

**Visualizing temporality in Music:**

I propose that three mid-level spectral feature of loudness, rhythmic density (onset detection) and roughness (Timbre) are critical to shape temporal experience. In my recent research I utilized two of these features to propose a visualization that convey a temporal understanding of the intended piece. I compare the result of this automatic tool to a traditional analysis, done using music score and human aural skills (Black vertical lines).

****

Two features of loudness (heat map) and density (blue scatter plot) are mid-level feature. Using some simple algorithms and statistical smoothing function I extracted red and blue segmentation that represent two different features individually. As you can see in Figure 3.21[[1]](#footnote-1) The black vertical lines, resulted from a traditional manual analysis, are coinciding with the proposed segmentation by the algorithm most of the time. However, to convey a holistic temporal structure, it seems relevant to propose the segmentation, considering all the intended features combined.

**Project Proposal**

I propose to build a model that takes a recorded musical example and extracts the formal structure of the piece utilizing Neural Network. The process is as the following.

1. Extract three mid-level feature from audio file, utilizing signal processing approaches.
2. Generate two-dimensional matrix that carries three set of features. This can be treated as a three channel image data which contains loudness, density and roughness instead of RGB channels.
3. Design and tune a neural network such as *Autoencoder* that is proper for unsupervised learning tasks.
4. Extract the temporal segmentation using clustering techniques.

**Unsupervised Learning - Autoencoder:**

An autoencoder is an artificial neural network used for unsupervised. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for the purpose of dimensionality reduction. The autoencoder idea is motivated by the concept of good representation, in other words it states that not all the representations or features of the input are suitable to perform the tasks such as classification. [[2]](#footnote-2)

**Software - Details**

The first portion of the code will extract three features of loudness, density and roughness in a form of three channel, two-dimensional array. For this portion I will use signal processing libraries such as *Librosa, Essentia, and numpy.* The second portion is a autoencoder neural network that reconstructs the input array each iteration. The cost function is the difference between the original and the reconstructed piece. The network learns by reducing the cost function to minimum. After several epochs, some of the hidden layer might include proper data in order to extract the temporal structure. For this portion I will used *Tensorflow, Keras,* or the combination of these two.

**Papers Studied:**

Deep Learning for Sequential Pattern Recognition, Pooyan Safari

Musical Structure Segmentation with Convolutional Neural Networks, Tim O’Brien

Composing Music with Recurrent Neural Networks, Daniel Johnson

**Future Work:**

A successful result from this project will provide the proper knowledge and tool to design a more complex system, capable of reconstructing musical pieces. At this point, composing music utilizing neural networks, such as convolutional neural network and RNN, mimicking a specific style or composer is feasible. Since these models are looking at small musical elements such as rhythmic pattern, harmonic progression and melodic figures, the final result is missing a holistic expression. I proposed this issue can be addressed by using my proposed approach. Instead of looking at small musical feature, considering a temporal understanding of a large number of musical examples, one might propose an optimal structure that fulfills holistic expression. In other word, this will provide a tool that takes a large number of musical examples, and learn the temporal pattern carried over all the examples. The machine can compose a new piece that reflects the temporal feature, learned from a collection of musical example. Indeed, considering style or a specific composer might be relevant.

1. Visualizing Temporality – Nima Hamidi – PhD Dissertation in Music Composition [↑](#footnote-ref-1)
2. Deep Learning for Sequential Pattern Recognition, Pooyan Safari [↑](#footnote-ref-2)