CARNEGIE MELLON UNIVERSITY

SENIOR UNDERGRADUATE THESIS

Unsupervised Musicality Prediction of Pitch Sequences

Author:
Jennifer HUANG

Advisor:
Dr. Tai Sing LEE

May 14, 2021

Abstract

In Randall and Greenberg's 2016 study, it was found that perception of the musicality of pitch sequences is inter-subjectively stable. We investigate the factors leading to this agreement by analyzing each sequence's predictability. PredNet (Lotter) is a deep convolutional recurrent neural network for unsupervised video prediction, which we adapt for audio prediction. The network is trained on a dataset consisting of single instrument melody audio file mel spectrograms and tested on the mel spectrograms of the audio files used in Randall and Greenberg's study. Varying degrees of contextual and top-down information are implemented in the network to determine the best predictors of music, allowing for insight into how musicality is linked to predictability.

Chapter 1

Introduction

1.1 Introduction to the Problem

Over the past few decades, deep neural networks have rapidly evolved to become effective models of the human brain on various tasks, from image classification to speech-to-text translation. While their performance on these tasks can meet or even exceed human performance, there are many capabilities of the human brain that even the most advanced deep learning techniques have yet to fully master. For instance, perception of beauty, the ability to classify something as more beautiful than another, is a quality which humans generally possess. However, little is known about whether deep neural networks are capable of perceiving beauty in the same way as humans.

In particular, we focus on the perception of beauty in music, which is sometimes known as musicality. The more musical a piece of music is, the more beautiful it is said to be. Many factors of a piece of music may influence a human's perception of its musicality, such as tempo, dynamics, and melody. Concentrating on melody, we analyze the predictability of the melody of a piece of music to determine if predictability impacts musicality perception.

In order to accomplish this, we make use of deep neural networks for video prediction, adapting them to be used for music audio prediction. In doing so, we gain a measure of music predictability from the network. Relating this measure of predictability to musicality, we can effectively determine if the predictability derived from neural networks is sufficient for judging musicality in a manner similar to that of humans.

1.2 Motivations and Background

1.2.1 Beauty and Musicality

Randall and Greenberg's research on musicality perception found that musicality is a continuum, where a series of sounds is not simply classified as musical or non-musical, but rather as more musical or less musical than another series of sounds. In their study, participants listened to 50 different pitch sequences and were asked to rate the musicality of each sequence on a scale from 1 to 5. Results demonstrated that there was clear, significant separability between the most musical sequences and the least musical sequences among all participants, indicating that musicality is a quality which is inter-subjectively stable while being a variable trait across different pitch sequences. Analyzing the pitch sequences considered most musical and least musical from a music theory standpoint yielded that smaller range, smaller mean-interval size, and smaller standard deviation of the mean were key features which led to a sequence being perceived as more musical. We will utilize the terms "least musical" and "non-musical" interchangeably throughout this paper.

In 2009, Schmidhuber proposed that subjective beauty of data is mainly based on its "interestingness", a measure of how easy it is to learn the data. Non-random, non-regular data which is able to be compressed in a way which makes it regular is classified as "beautiful". Thus, the beauty perception of data is guided by its predictability. Data which is too predictable is not non-regular, and thus can appear too boring or monotonous. On the other hand, data which is too unpredictable can appear too random and unstable, and thus unable to be compressed.

1.2.2 PredNet

PredNet is a predictive deep convolutional recurrent neural network used for video prediction (Lotter). Designed using the concept of predictive coding from neuroscience, it aims to continually generate and update a model of the input data in order to predict what will happen next by comparing predictions with the actual input. It does so in an unsupervised manner, with no labels being passed in with the data during training. Given a sequence of video frames, PredNet predicts the next frame. At each layer in the network, representation neurons output a layer-specific prediction and compare it against the target to produce an error, which is propagated vertically to the next layer.

We hypothesize that a predictive neural network which is trained to listen to music should have similar musicality perception of pitch sequences as the participants from Randall and Greenberg's study. This hypothesis is based off of Schmidhuber's theory of beauty, which implies that the musicality of a piece of music can be determined by its predictability.

In order to test our hypothesis, we utilize PredNet's ability to generate predictions of data to gain a measure of predictability of the 10 most musical and 10 least musical pitch sequences from Randall and Greenberg's study. Using this, we will be able to tell if predictability is a significant predictor of musicality.

1.2.3 Deep Learning for Audio

Much research has been done on how deep neural networks can be applied to audio tasks such as automatic speech recognition, speaker identification, and music genre classification. Work done on unsupervised audio classification using convolutional deep belief networks converts the audio data into a spectrogram, which the network then extracts features from in order to perform classification (Lee). In the same way, patches extracted by spectrograms produced by music, with various amounts of overlaps, are used in order to perform music classification using convolutional neural networks (Costa). In both of these research works, we see that spectrograms are commonly used when working with audio in deep learning, because converting audio to spectrograms allows the network to visualize the data and easily extract features from it.

Chapter 2

Methodology and Results

2.1 Methodology

2.1.1 Training, Validation, and Testing Data

The training and validation dataset used is the Medley-solos-DB dataset, which consists of WAV files of single instrument melodies, each being 2972 milliseconds long (Lostanlen). Each file consists of audio purely from one of the following instruments: clarinet, distorted electric guitar, female singer, flute, piano, tenor saxophone, trumpet, and violin.

This dataset was chosen because it contains isolated melodies from single instruments, much like the pitch sequences presented in the Randall and Greenberg musicality study. Furthermore, we choose to use audio from commonly heard musical instruments, rather than computer-generated audio similar to that of the testing data, in order to best model the musical experience of the participants in Randall and Greenberg's study. In particular, participants will generally have heard music from instruments like that of the training data, and are "tested" during the study on computer-generated music.

For testing, the 10 most musical and 10 least musical pitch sequence audio files from the Randall and Greenberg study are used. This allows the differences in prediction error by PredNet between the two opposite sides of the spectrum of musicality to be easily compared.

2.1.2 Data Preprocessing

The audio files must be converted into a format appropriate for PredNet to use. As a video prediction network, PredNet expects a sequence of video frames in order to predict the next video frame. Thus, we must "visualize" the audio files and transform each into a sequence of video frames.

In order to do this, we convert each audio file into a mel spectrogram. The mel spectrogram is a spectral-temporal representation of sound, with frequency in the mel scale as the y axis and time as the x axis. The mel scale, which is a frequency scale in which humans perceive pitches to be equal distance from each other, is used in order to more closely simulate human perception of the audio. Furthermore, the amount of loudness for a given frequency at a given time is represented by each mel spectrogram pixel's brightness, on a relative scale from -80 to 0 dB where 0 dB corresponds to the maximum loudness. As such, all values in the mel spectrogram are normalized, changing the range of values from [-80,0] to [0,255] in order to be properly input into PredNet.

Taking the mel spectrogram for each audio clip, we generate a "video" from it by using a "sliding window" technique. We slice the spectrogram at regular intervals along the x axis, shifting to the right by a fixed amount for each slice. The size of each slice, as well as the amount of overlap between consecutive slices, is adjusted during the experiments conducted. PredNet is then given these mel spectrogram frame "videos" as input.

2.2 Experiments

2.2.1 Experiment 1: Description

In the first experiment, each spectrogram is sliced into non-overlapping frames, with each frame capturing 0.25 seconds of the audio. Every five consecutive frames of the spectrogram are passed into a four-layer PredNet, and the prediction error is computed by calculating the mean squared error between the final fifth prediction for the frame sequence and the actual final fifth frame of the frame sequence. This PredNet model does not incorporate loss from its upper layers; instead, it completely relies on the loss from its bottommost layer. This version of the network is known as PredNet L_0 (Lotter).

After training on the Medley-solos-DB dataset for 50 epochs, the final fifth frame predictions for each five frame sequence in each spectrogram are concatenated together horizontally to create a "reconstruction" of the original spectrogram (Figure 2.1). Since our first fifth frame prediction happens at the 1 second mark of the audio, only part of the original spectrogram is reconstructed. As such, the non-reconstructed frames of the original spectrogram are not comparable to the predicted frames and are thus cropped out of the images in Figure 2.1.

2.2.2 Experiment 1: Results

From Figure 2.1, we can see that the predicted frame reconstruction of the spectrogram closely resembles that of the original spectrogram, except that all frames are shifted to the right by 1 frame. This leads to the conclusion that this PredNet model has simply learned to output exactly what it saw in the previous frame as its prediction for the next frame.

The average prediction error between the 10 most musical and 10 least musical sequences over the predicted frames is shown on the left side of Figure 2.2. A paired-t test indicates that the difference between the most musical and least musical prediction error distributions is statistically significant, with the most musical prediction error being smaller than the least musical prediction error.

These results support the conclusion from Randall and Greenberg's study that mean-interval size is a predictor of musicality. Simply predicting that the next frame will be exactly the same as the frame that came before results in a statistically significant difference between prediction error for the most musical and least musical sequences. The prediction error between two spectrogram frames having bands around the same frequency area (small interval size) is lower than that of two spectrogram frames having bands around different frequency areas (large interval size). Thus, lower prediction error, which we find to be correlated with the most musical pitch sequences in this case, corresponds to smaller mean-interval size.

To investigate this hypothesis, we calculate the mean interval size of each of the 10 most musical and least musical pitch sequences, and choose three sequences from each category which have mean interval size between 4 and 5. The prediction error between the most musical and least musical sequences in this selective group is shown on the right side of Figure 2.2. A paired-t test indicates that the difference between these selected most musical and least musical distributions is not statistically significant, confirming our hypothesis that mean interval size is a key factor in these predictions and in the ability to distinguish between most musical and least musical sequences.

2.2.3 Experiment 2: Description

Based off of the results from Experiment 1, we investigate whether the presence of the first three frames in each five frame sequence impacts the predictions of PredNet, or if PredNet is simply using the fourth frame to predict the fifth. In order to test this, we pass every two consecutive frames of the spectrogram into PredNet L_0 for training, validation, and testing and compare this prediction error with the prediction error we get when passing in every fifth consecutive frames. This is done for 0.25 second non-overlapping frames, 0.35 second non-overlapping frames, and 0.5 second non-overlapping frames.

2.2.4 Experiment 2: Results

As shown in Figure 2.3, the prediction error of PredNet L_0 is not significantly impacted by the presence of only one previous frame versus four previous frames. This was found to be true for all frame widths experimented with, indicating that contextual information beyond the one frame directly before the predicted frame is not particularly useful for this model in generating predictions.

2.2.5 Experiment 3: Description

Next, we utilize the more comprehensive version of PredNet, PredNet L_{all}, which has a 0.1 weight on the loss of the upper layers (Lotter). Maintaining the same number of layers, four, as before, this PredNet L_{all} model's loss used during the training process is a weighted sum of the loss of the bottom layer times 0.7 and the loss of each of the upper layers times 0.1.

Furthermore, in this experiment we decide to input the entire sequence of 0.25 second frames for each audio file into both the PredNet L_0 and PredNet L_{all} models, creating one comprehensive spectrogram frame "video" for each audio file rather than several smaller "videos" for each audio file. This is done to more closely resemble the methodology used in the original PredNet paper (Lotter).

2.2.6 Experiment 3: Results

The comparison between the PredNet L_0 and PredNet L_{all} training loss graphs over 50 epochs can be seen in Figure 2.4. As shown, PredNet L_{all} is more effective in minimizing training loss and is more stable during training, indicating that top-down information in PredNet is useful in generating predictions for music.

2.2.7 Experiment 4: Description

In all previous experiments, the predicted frame corresponding to the final time step in each sequence was the only one compared against the actual frame in the calculation of prediction error; however, PredNet is also producing predictions for each of the intermediate time steps in the input sequence. In this experiment, we monitor the predictions for each time step that PredNet L_{all} produces, with the entire sequence of frames for each audio file as input (same as from Experiment 3).

We train, validate, and test PredNet L_{all} with a variety of input frame sizes and overlaps to determine the most effective frame representation for the model to perform spectrogram frame prediction. Specifically, we test the baseline of non-overlapping 0.25 second frames, 60% overlapping 0.25 second frames, and 80% overlapping 0.5 second frames.

2.2.8 Experiment 4: Results

The training and validation results from this experiment are shown in Figure 2.5. As can be seen, the model with 60% frame overlap and 0.25 second frame width does not converge as well as the other two models during training, which is reflected in the validation prediction error graph. This may be due to the fact that 0.25 second frames are relatively narrow, and the 60% overlap means that consecutive frames are generally very similar to one another. As such, PredNet does not have enough information to perform proper, meaningful predictions. On the other hand, the model with 80% frame overlap and 0.5 second frame width performs the best.

All of our validation prediction error results, as shown in the right column of Figure 2.5, indicate that, excluding the results from the first and second frames, predictions do not get significantly better over time as the context of previous spectrogram frames seen increases (from left to right on the x axis). The prediction error only decreases drastically from the first frame to the second frame, which is because the first frame does not have any contextual information and just outputs a blank prediction, yielding a high prediction error.

Furthermore, the average prediction error between the selected most musical and least musical sequences (with average interval size between 4 and 5, selected from Experiment 1) over the predicted frames is calculated (Figure 2.6). A paired-t test indicates that the difference between these most musical and least musical prediction error distributions is statistically significant. From this, we can conclude that the difference in most musical and least musical prediction error given top-down information and the context of the entire pitch sequence, with 80% frame overlap and 0.5 second frame width, is significant.

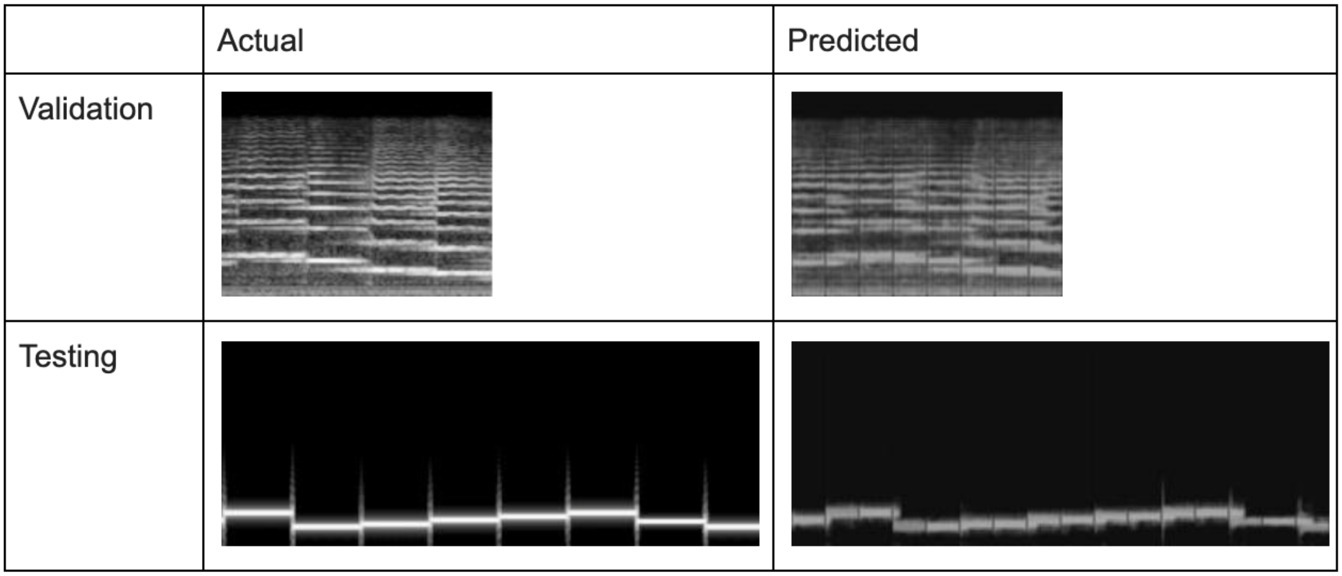

FIGURE 2.1: Experiment 1 - Concatenation of actual original frames (left column) compared with concatenation of predicted frames (right column), on one validation spectrogram and one testing spectrogram.

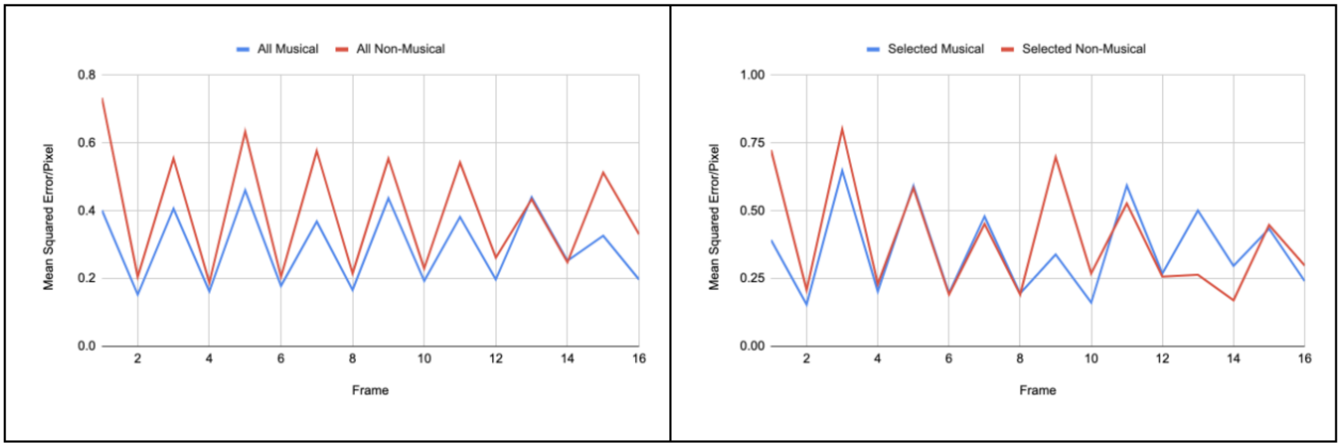

FIGURE 2.2: Experiment 1 -
Left: Average prediction error for the 10 most musical and 10 least musical sequences over consecutive predicted frames.
Right: Average prediction error for the selected most musical and selected least musical sequences (with average interval size between 4 and 5) over consecutive predicted frames.
Note: Each of the prediction error graphs in this figure and the following figures are created by plotting the average prediction error for Frames 1,2,3,... of the spectrogram for the most musical and least musical sets and then linearly interpolating between the data points.

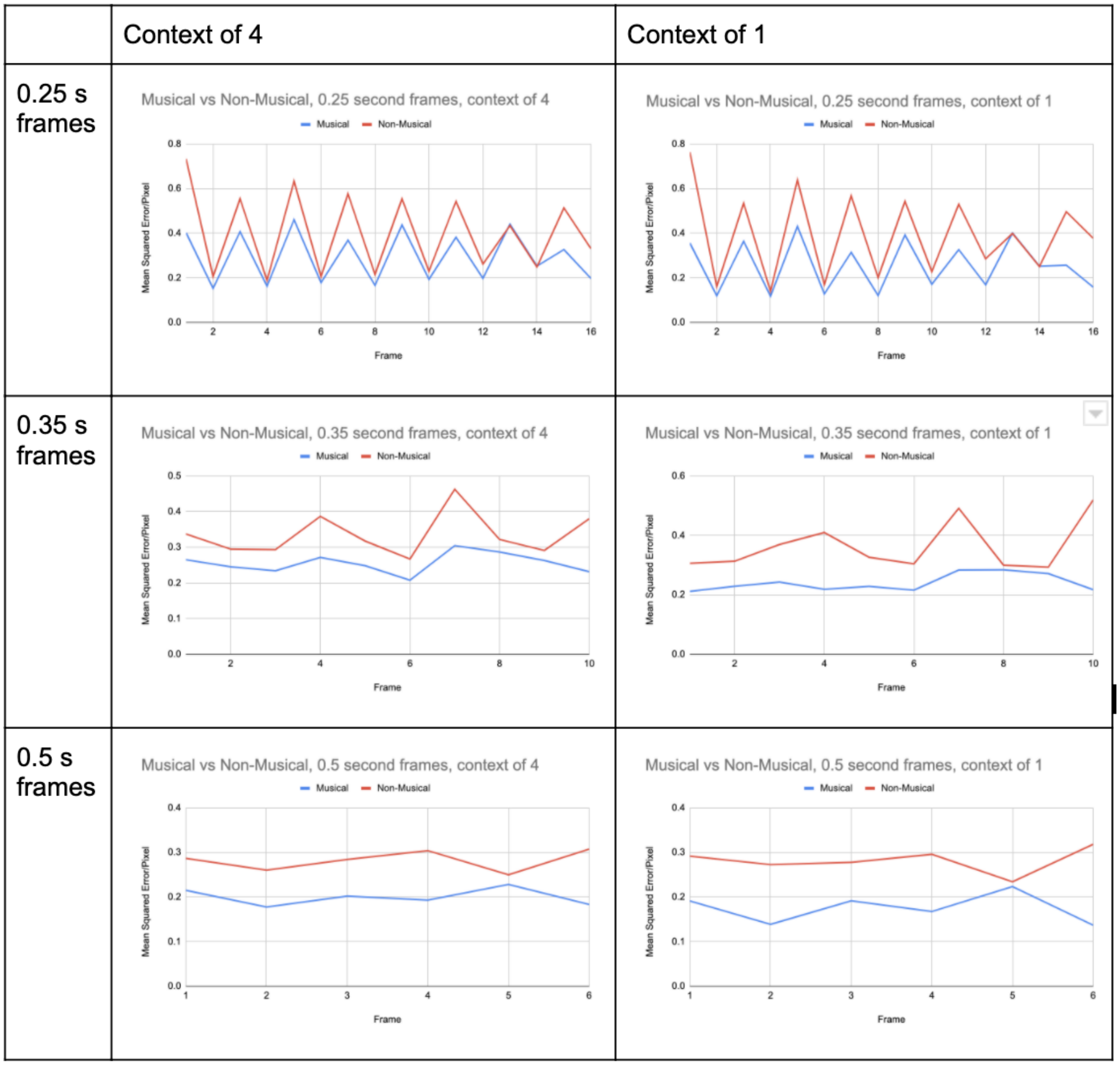

FIGURE 2.3: Experiment 2 - Average prediction error for the 10 most musical and 10 least musical sequences over consecutive predicted frames, comparing context of four previous frames (left column) with context of one previous frame (right column).

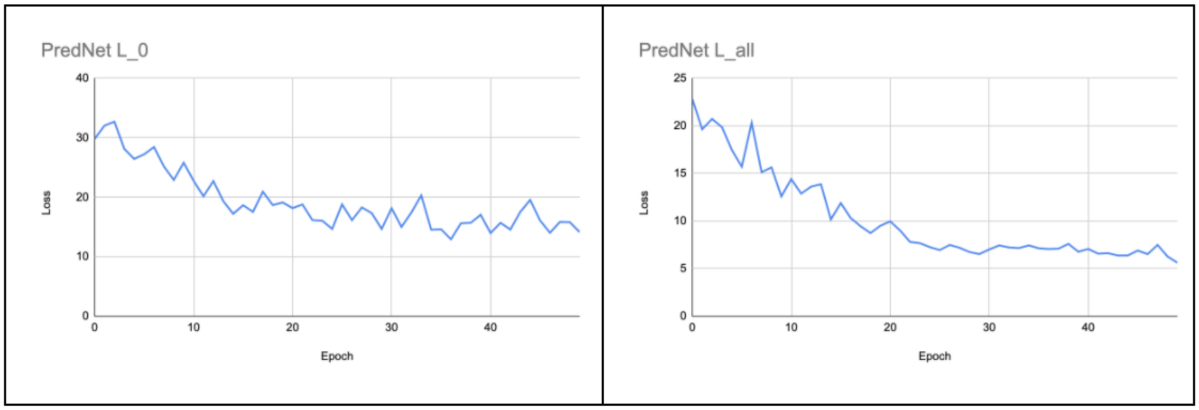

FIGURE 2.4: Experiment 3 - Training loss over epochs for PredNet L_0 (left) and L_{all} (right) on complete spectrogram frame sequences.

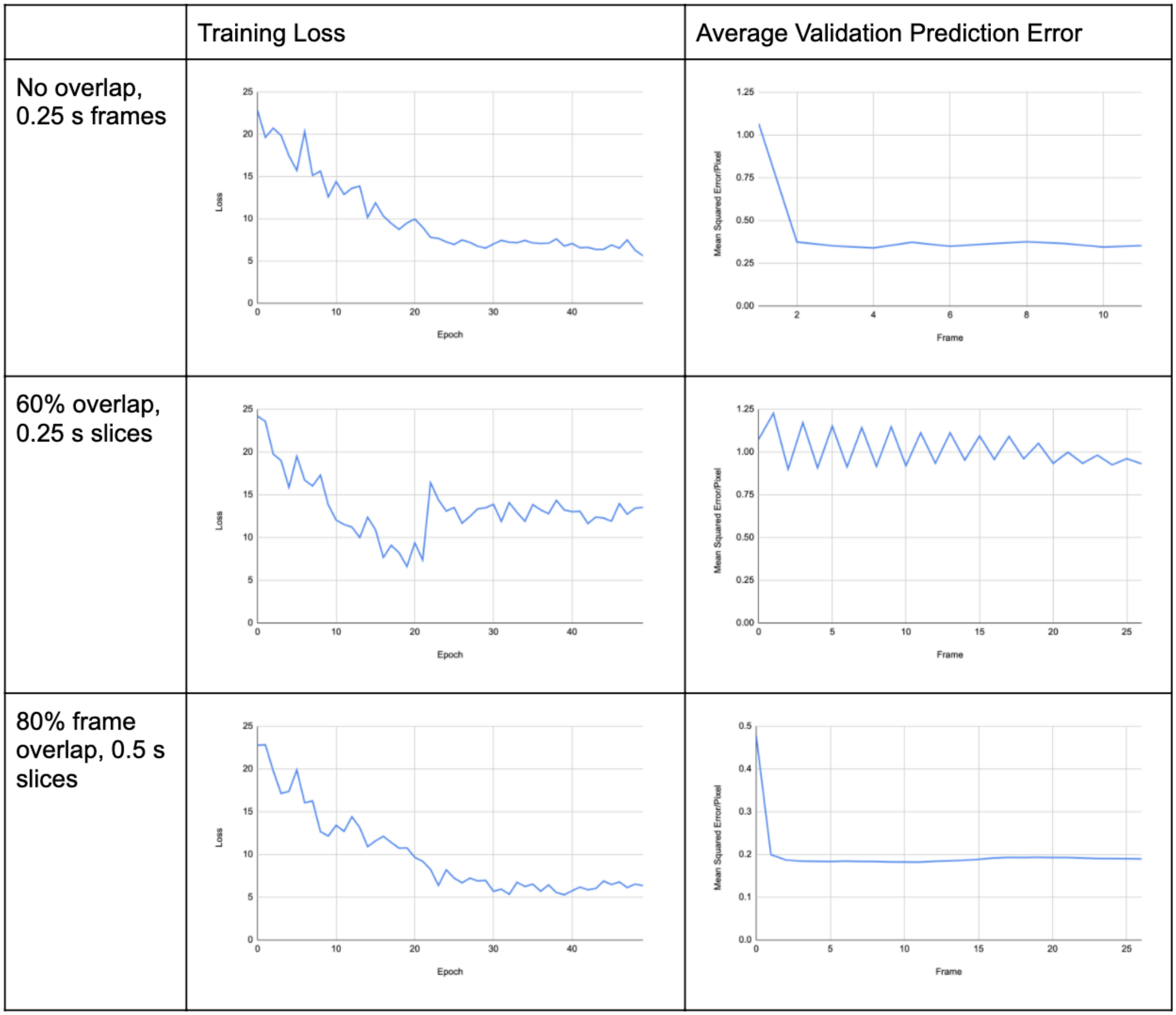

FIGURE 2.5: Experiment 4 - Training loss over epochs (left column) and average validation prediction error for each consecutive predicted frame (right column).

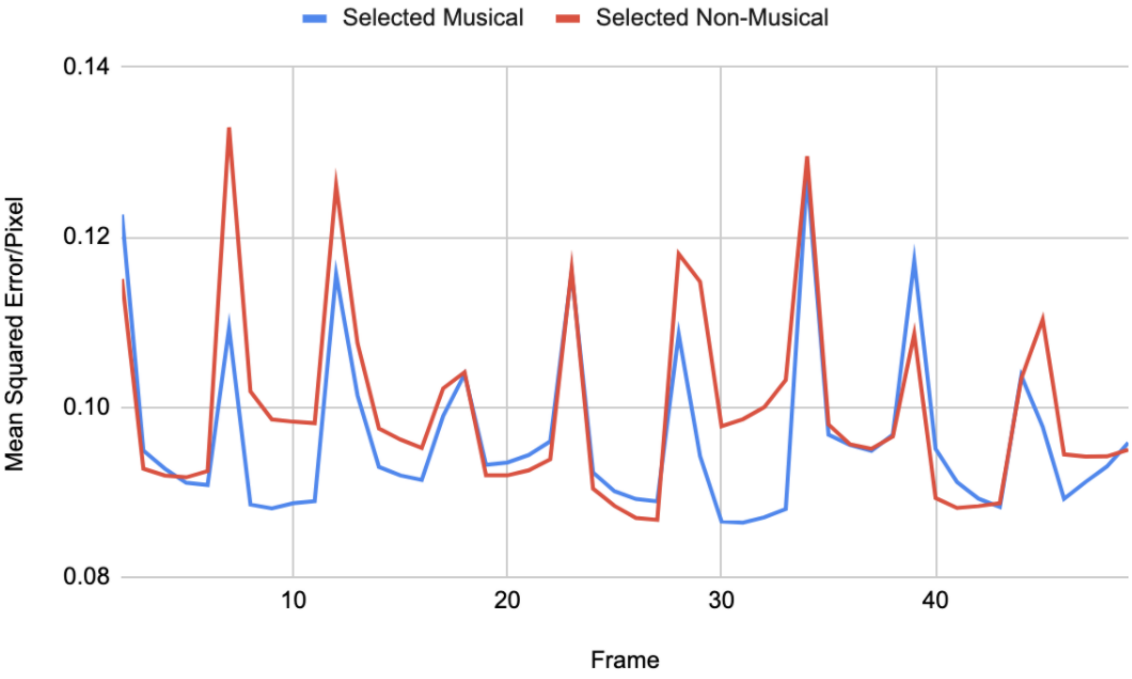

FIGURE 2.6: Experiment 4 - Average prediction error between the selected most musical and selected least musical sequences (with average interval size between 4 and 5) over consecutive predicted frames.

Chapter 3

Conclusions

3.1 Comparison with Prior Work

Compared with prior work related to deep learning and music audio, our work uses a similar data preprocessing technique by transforming the music audio files in a spectrogram. However, the motivation and task which we train the neural network to perform is quite unique from any other literature in the field.

Overall, the results of the experiments correlate with and confirm the conclusions made by Randall and Greenberg in their musicality work. In particular, the idea that mean interval size is one of the key predictors of musicality is reinforced by Experiment 1.

Based on our analysis of the predictability of each sequence using PredNet, the final results of Experiment 4 allow us to conclude that the predictability of a pitch sequence influences the perception of its musicality. Thus, we have found evidence that Schmidhuber's theory, which states that the perception of beauty in data can be linked to how predictable the data is, can be applied to music, specifically single note sequences. In particular, we conclude that a predictive neural network trained to listen to music can predict the most musical sequences better than the least musical sequences.

3.2 Future Work

We have learned that predictions of audio can vary greatly based on factors like frame size, amount of overlap, amount of context, and amount of top-down information. Further work could be done to determine the most optimal way to encode audio information in order to generate the most accurate predictions.

Furthermore, analysis of the predictability of note sequences was done using PredNet, a video prediction network. As such, there are key differences between the spectrogram frames which were input into PredNet during this study and the video frames which PredNet was tailored towards in Lotter's paper.

Namely, video frames naturally have an x and y axis which both capture an instantaneous moment in time, while spectrogram frames have an y axis capturing the frequencies of the sounds heard and an x axis capturing a range of time. Thus, other methods to visualize the audio files could be investigated in the future in order to best accommodate the usage of PredNet. On the other hand, other prediction networks besides PredNet could be used to generate predictions of the audio.

Additionally, an experiment which confirms and furthers our work could be constructed. Randomly generating note sequences and passing them into our trained PredNet model will give us a measure of predictability for each sequence. Human participants could then listen to these same randomly generated note sequences and judge their musicality, allowing us to determine if the humans' musicality judgements are correlated in any way with the relative predictability level output by PredNet.

Finally, we could experiment with how the style of the music input into PredNet during training impacts the accuracy of its predictions during testing. For example, we could determine if training PredNet with African music results in greater prediction accuracy when tested on more African music, compared with the prediction accuracy when tested on Chinese music. Extending these findings from PredNet to a theory of how the human brain perceives music, such an experiment can yield insights into how musicality perception is impacted by one's upbringing or culture.

Bibliography

Costa, Yandre M.G., Luiz S. Oliveira, and Carlos N. Silla Jr. (Mar. 2017). "An evaluation of Convolutional Neural Networks for music classification using spectrograms". In: URL: https://dl.acm.org/doi/10.1016/j.asoc.2016.12.024.

Lee, Honglak et al. (Dec. 2009). "Unsupervised feature learning for audio classification using convolutional deep belief networks". In: URL: https://dl.acm.org/doi/10.5555/2984093.2984217.

Lostanlen, Vincent et al. (2018). "Medley-solos-DB: a cross-collection dataset for musical instrument recognition". In: URL: http://doi.org/10.5281/zenodo.1344103.

Lotter, William, Gabriel Kreiman, and David Cox (Mar. 2017). "Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning". In: URL: https://arxiv.org/pdf/1605.08104.pdf.

Randall, Richard and Adam Greenberg (July 2016). "Principal Components Analysis of Musicality in Pitch Sequences". In: URL: https://www.researchgate.net/publication/344596330_Principal_Components_Analysis_of_Musicality_in_Pitch_Sequences.

Schmidhuber, Jürgen (Jan. 2007). "Simple Algorithmic Principles of Discovery, Subjective Beauty, Selective Attention, Curiosity Creativity". In: URL: https://arxiv.org/pdf/0709.0674.pdf.