Unsupervised Musicality Prediction of Pitch Sequences

Nikolas McNeal[1,4], Aniekan Umoren[2,4], Jennifer Huang[3,4], Tai Sing Lee[3,4]

[1]The Ohio State University, [2]Massachusetts Institute of Technology, [3]Carnegie Mellon University, [4]Center for Neural Basis of Cognition

CNBC CENTER FOR THE NEURAL BASIS OF COGNITION

Carnegie Mellon University

Motivation

Can a neural network perceive music like humans do?

Randall and Greenberg's 2016 study[1] demonstrated that "musicality" is a continuum, where sounds are not simply classified as musical or non-musical, but more or less musical than each other. Participants listened to 50 different pitch sequences and were asked to rate the musicality of each sequence on a scale.

Results showed a significant distinction 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.

In addition, Randall and Greenberg showed that smaller range, smaller mean-interval size, and smaller standard deviation of the mean were key features of audio sequences perceived as more musical. Contour, diatonic entropy, motive, and tonality did not significantly correlate with the participants' rankings.

We hypothesize that if a predictive neural network is trained on audio sequences, it will demonstrate musicality perception similar to humans.

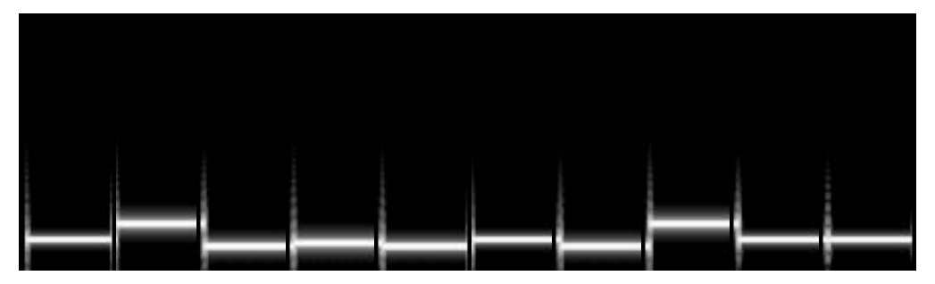
We test this by using a prediction network *PredNet*, originally designed for video prediction.

Methodology

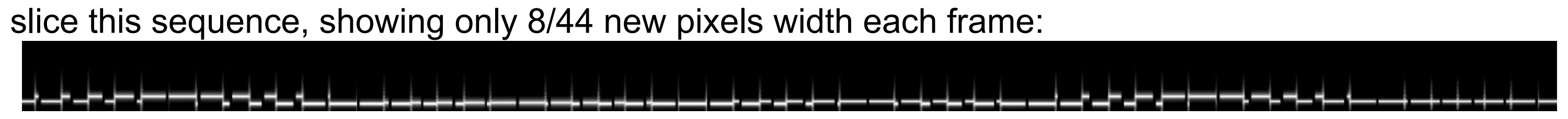
With PredNet[2], a predictive model is created in higher cortical areas, communicated through top-down connections to lower areas, and compared against the actual observations, before propagating the error forward to update the prediction.

For each frame *t*, PredNet generates a predicted frame *t+1*, compares to the real frame *t*, and continues.

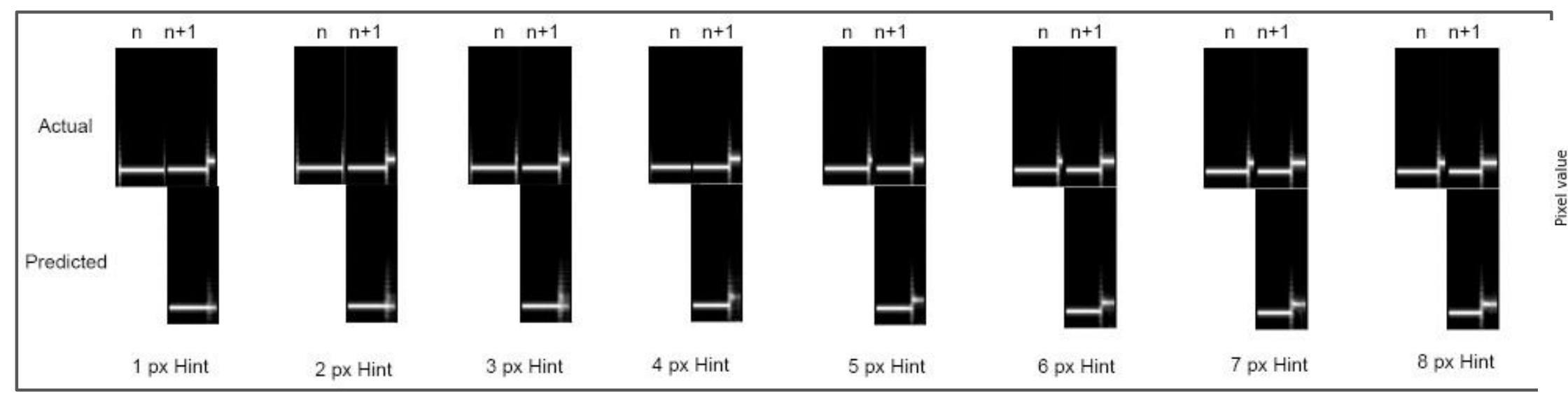
We generate mel spectrograms of each audio sequence. Training is performed on 8,052 3-second clips with harmonies. Testing is performed on ten musical single-note sequences and ten non-musical single-note sequences. We utilize a "sliding window" with overlap to create a series of frames. Shown below is one example of an audio sequence, where every vertical spike represents a note change.

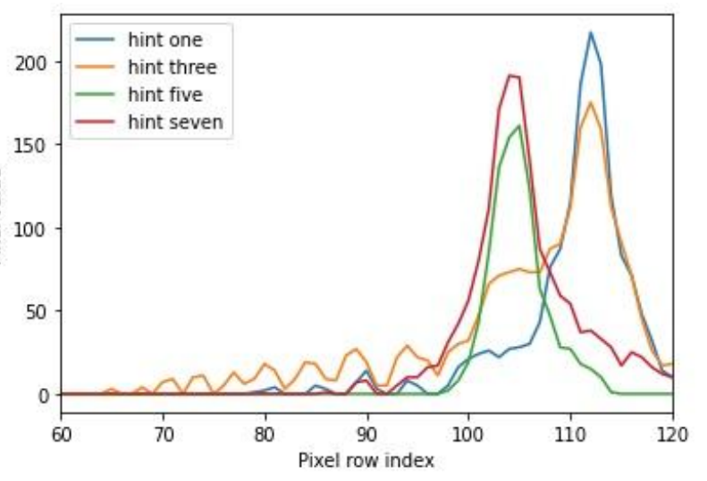


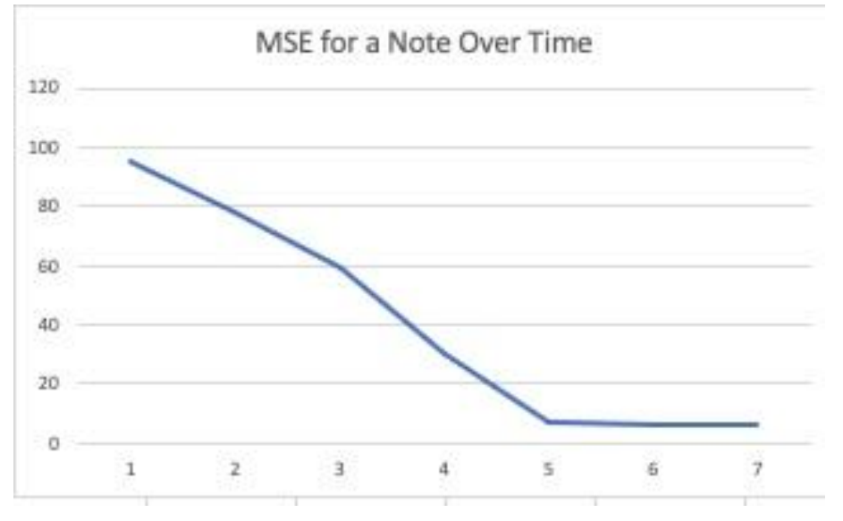
Since prediction networks require some overlap in order to make an accurate prediction, we evenly slice this sequence, showing only 8/44 new pixels width each frame:



However, because this sequence is sliced at even intervals, one particular note change may provide more or less pixels of "hint" than the others. The bottom left images depict this. The top row depicts ground-truth frames with increasing "hint," whereas the bottom row depicts the prediction.







Left: A plot visualizing the MSE decrease as "hint" increases. This graph considers the average of all notes. In our results, we account for this variability of pixel "hint" by averaging the results of eight pixel shifts in our experiments.
Top right: An example of the distributions for a +8 pixel note change, for different hint amounts. The tail of lower hints is longer than higher hints (i.e., there is more uncertainty).

Prediction MSE Behavior

PredNet demonstrates higher prediction errors for non-musical sequences than musical sequences.

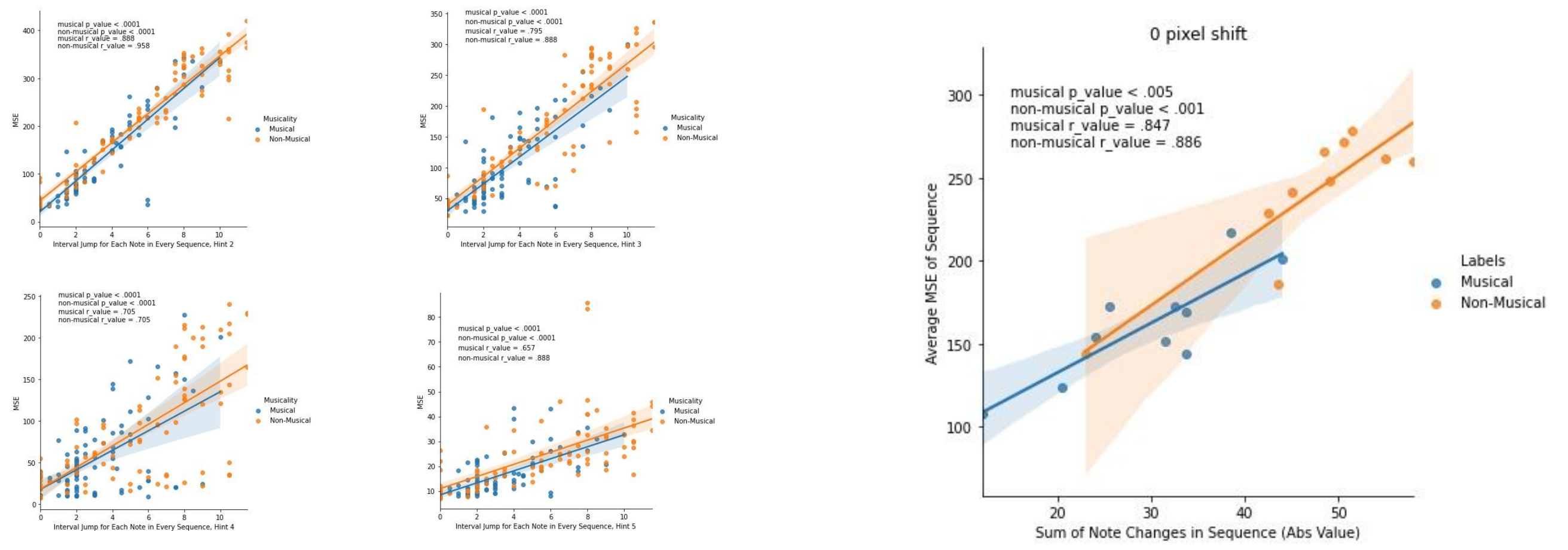


Figure 1.

a) **Left:** Example plots of the MSE vs interval "jump" size for every note in every sequence, for hints 2, 3, 4, and 5. MSE is higher for non-musical sequences than musical sequences, but this is in part due to the greater note jumps in the non-musical clips.
b) **Right:** A plot comparing the total interval size of every musical sequence to the total interval size of every non-musical sequence. There is a positive correlation between total interval size and MSE. This plot depicts zero pixel shift, but this does not necessarily mean every note has the same pixel hint, so plots of similar pixel shifts are very similar.

Figure 2 shows that the MSE for musical pieces decrease over time, even though the note step size tends to increase. This suggests that PredNet does remember the context of multiple notes when making prediction.

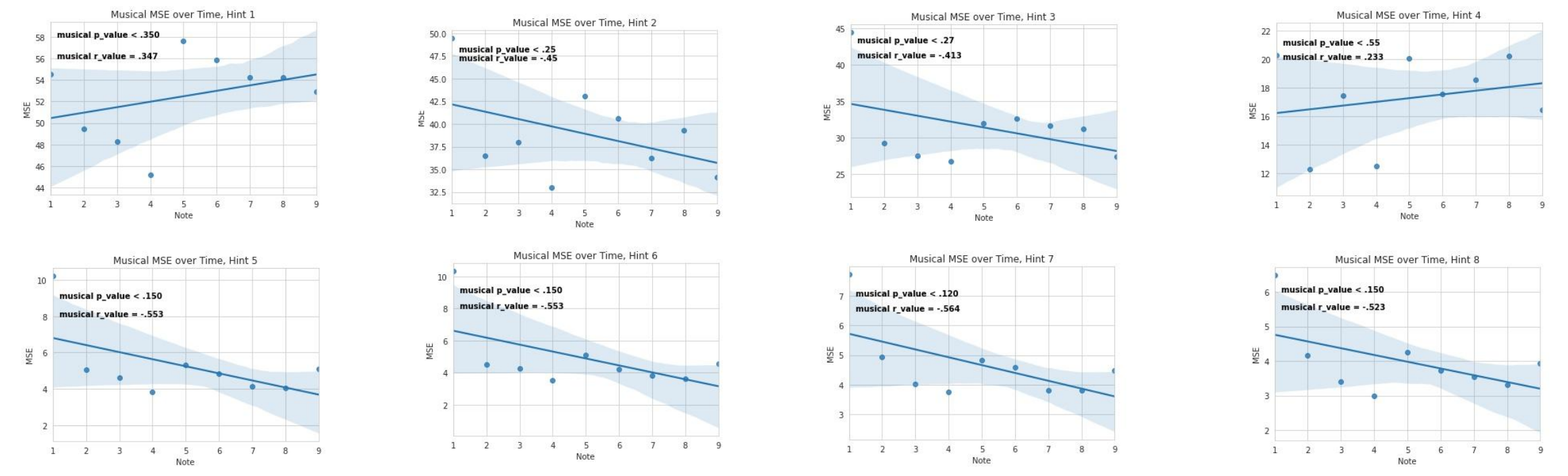


***Figure 2.* MSE for Musical Sequences.**
This figure of eight plots shows the MSE (y-axis) of each note number (x-axis) for the musical sequences. Each plot has a different amount of "hint," with the amount of hint increasing from one to eight, from top-left to bottom-right. Each note was divided by its average note jump.

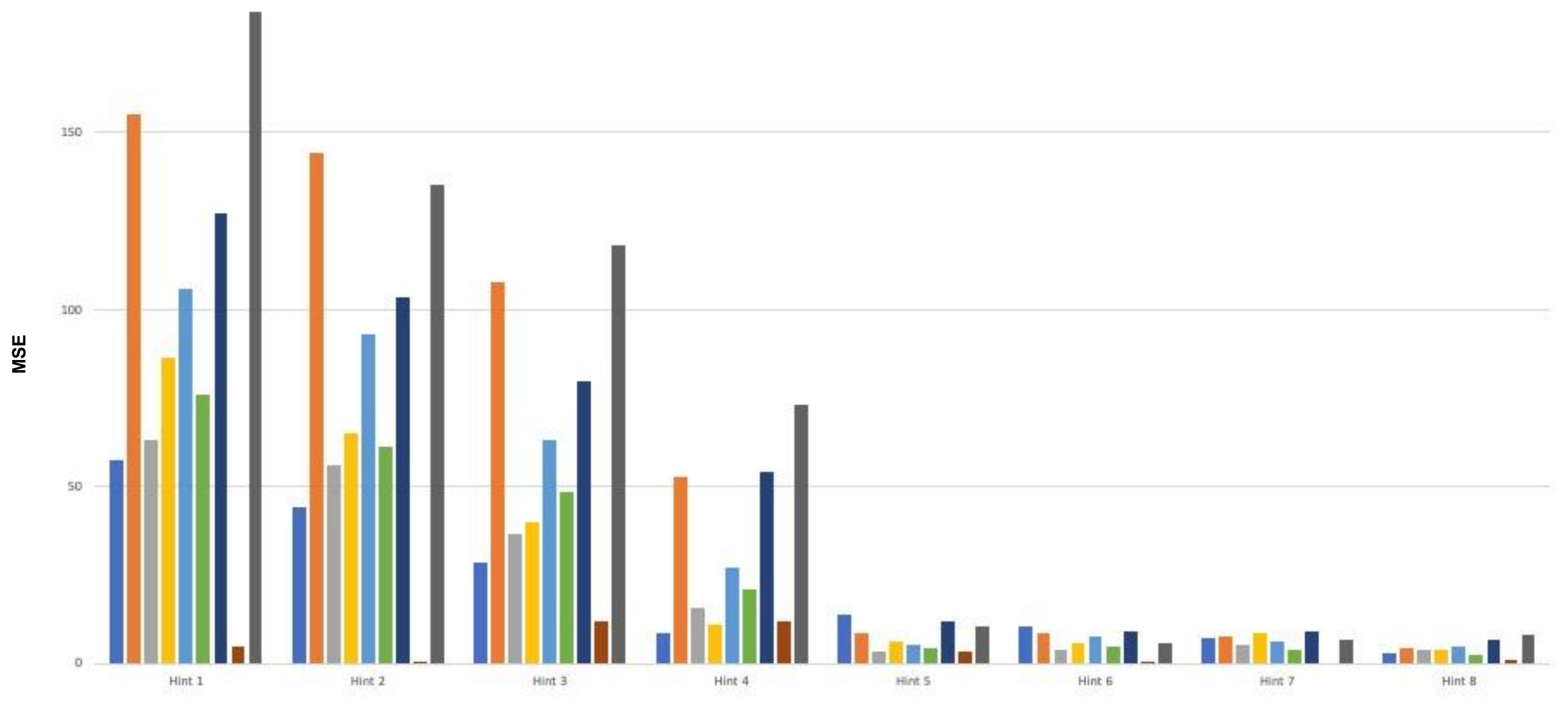


***Figure 3.* Difference between Sequences' Non-Musical MSE and Musical MSE**
Figure 3 shows the difference between non-musical and musical (non-musical - musical) prediction error for the nine notes, for different levels of hints. A smaller hint produces greater errors and greater difference. There is a general tendency for the difference to increase when more notes are listened for different levels of hints.

Note that these results are confounded by the fact the average step size at each note change over time is not uniform. The average jump between Note 2 and Note 3 is 6.4 for non-musical and 2.425 for musical, while the average jump between Note 9 and Note 10 is 6.1 for musical and 2.95 for non-musical. Nevertheless, the difference in MSE is still greater for note 9-10 than note 2-3, which again indicates a longer-term memory effect than just based on the note jump (Figure 1).

Note: We use *mean squared error* (MSE) to calculate our prediction errors. The MSE is calculated after our pixel values are multiplied by 255 (to convert to int type), so while our MSE values are scaled up, they are directly proportional.

Attributes MSE

MSE increases positively with "Mean," "Standard Deviation," and "Range."
In addition, MSE increases as ZCONT(contour, or similarity to the archetype) and ZPPM (compressibility, or "partial predictability of motive") increase. This is unexpected as some studies indicate higher contour/predictability is more musical. However, in the human subject data, non-musical tends to have higher contour and PPM for these random sequences, consistent with the PredNet results.

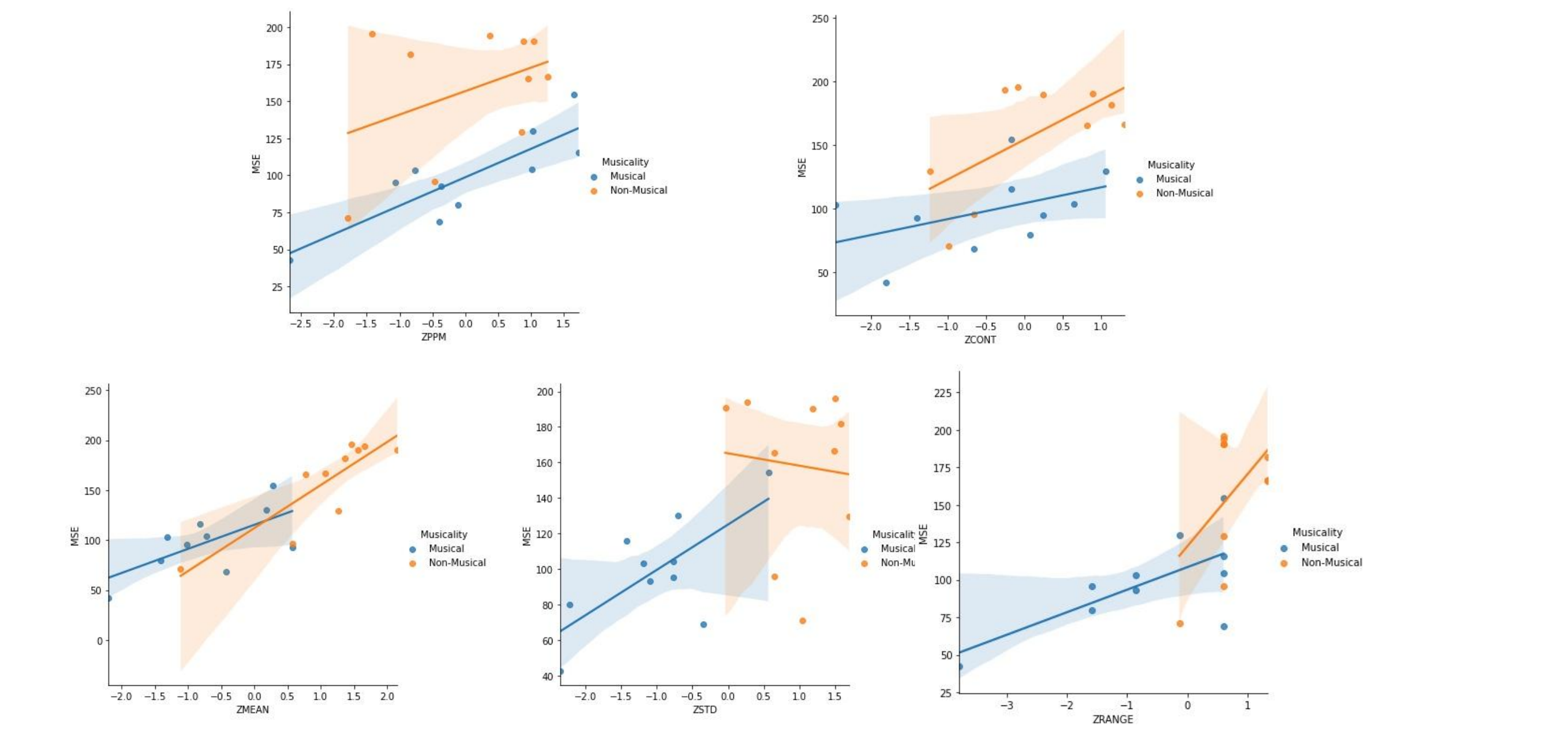


Figure 4.

a) **Top-left:** ZPPM vs MSE. *PPM* measures the "compressibility" of the audio, or an estimation of its predictability based on motive frequency
b) **Top-right**: ZCONT vs MSE. *Contour* measures overall shape (pitches, rhythms, tempi, timbre)
c) **Bottom-left:** ZMEAN vs MSE. *Mean* measures mean interval size (average note jump)
d) **Bottom-middle:** ZSTD vs MSE. *STD* measures the standard deviations of mean in a clip
e) **Bottom-right:** ZRANGE vs MSE. *Range* measures the range of intervals in a clip

Summary and Discussion

We tested a neural network's ability to predict audio and found that it performs better on musical sequences than non-musical sequences.

In addition, our results coincide with Randall and Greenberg's finding that only smaller range, smaller mean-interval size, and smaller standard deviation of the mean significantly correlate with musicality. We observe the MSE for musical pieces decrease over time, even when the note step size increases, suggesting that PredNet's stacked LSTMs effectively remember and consider context of previous notes. We also observe the difference between non-musical MSE and musical MSE increase as notes progress, though the variance in this result may be due to the non-uniform average step size at each note change.

This could, in part, relate to Jurgen Schmidhuber's theory of computational beauty[3], which states that there may be a link between predictable data and things that are "beautiful." Non-random, non-regular data which is able to be compressed in a way which makes it regular is classified as "beautiful," or in our case, "musical."

Future work could include testing on harmonies/chords/ensembles, rather than simply the single-note sequences.

References

[1] 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.

[2] 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.

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