Auditory Grouping: Using Human Data to Produce a Grouping Algorithm

A Major Qualifying Project
Submitted to the Faculty of
WORCESTER POLYTECHNIC INSTITUTE
In partial fulfillment of the requirements for the
Degree of Bachelor of Science

Advisor: Professor James K. Doyle

Student Investigator: Zachary Wagner

4/28/2022

This report represents work of WPI undergraduate students submitted to the faculty as evidence of a degree requirement. WPI routinely publishes these reports on its web site without editorial or peer review. For more information about the projects program at WPI, see http://www.wpi.edu/Academics/Projects.

TABLE OF CONTENTS

ABSTRACT Page 3

ACKNOWLEDGEMENTS Page 4

LIST OF FIGURES Page 5

INTRODUCTION Page 6

MOTIVATION Page 7

BACKGROUND Page 7

NOVELTY Page 7

PROJECT DESIGN Page 8

EXPERIMENTAL DESIGN FOR DATA COLLECTION Page 8

Subjects *Page 8*

Design and Materials *Page 8*

COMPUTATIONAL MODEL DESIGN Page 9

Preprocessing *Page 9*

Targeting *Page 10*

Convolutional Neural Networks *Page 10*

Short-Time Fourier Transformations *Page 10*

Recurrent Neural Networks *Page 10*

Combined CNN’s and RNN’s *Page 10*

Cross Validation *Page 11*

COMPUTATIONAL METHODS Page 11

Spectogram Transformation *Page 11*

Supervised Pretraining *Page 11*

Google c/oud/TPU's *Page 11*

Fuzziness *Page 12*

Left/Right Aligning *Page 12*

ANALYSIS AND DISCUSSION Page 13

Demographic and Sorting Questions *Page 13*

Auditory Groupings *Page 17*

Average Groupings Per Task *Page 18*

Average Time Per Task *Page 20*

Correlating Demographic Data with Grouping Data *Page 21*

CONCLUSION Page 24

Suggestions for Future Research *Page 25*

REFERENCES Page 26

APPENDIX A : Informed Consent Page 27

APPENDIX B : Survey Flow and Questions Page 29

Abstract

Computerized encoding of audio information is highly complicated and not entirely understood. We utilized a hybridization of a machine learning algorithm and human auditory grouping and segmentation to further advance machine-based audio perception and grouping models. Human subjects listened to audio clips of musical selections and performed auditory grouping and segmentation of the clips. Data collected from the subjects' grouping/segmentation were utilized by our machine learning algorithm to enhance the algorithm's ability to emulate human auditory grouping and segmentation. A survey was also administered to collect information on demographics and musical experience. Overall, it was difficult to establish a direct correlation between the demographic data and the human-performed auditory grouping of the audio clips, with one exception concerning the number of groupings placed by subjects and the number of musical genres they enjoy. Suggestions for future research are discussed.

Acknowledgments

This paper represents the psychological science component of a larger project with both a psychology and computer science orientation. Credit for the short descriptions of computer science concepts included in this report go to my colleagues Cheng-Hsuan (Sean) Jan, William McDonald, and Yang (Robin) Chen. The reference from their report containing the computer science component of the project is cited throughout where it is relevant to the content discussed in my own design components and analyses. The full citation for their paper can also be found in the reference section of this report

LIST OF FIGURES

FIGURE 1 is a bar plot of the GENDER OF SURVEY PARTICIPANTS

FIGURE 2 is a bar plot of the AGE DISTRIBUTION OF SURVEY PARTICIPANTS

FIGURE 3 is a bar plot of the RACE/ETHNICITY OF SURVEY PARTICIPANTS

FIGURE 4 is a bar plot of the PREFERRED MUSICAL GENRES OF SURVEY PARTICIPANTS

FIGURE 5 is a bar plot of the MUSCIAL EXPERIENCE OF 36 SURVEY PARTICIPANTS

FIGURE 6 is a histogram of the LOGARITHMIC AVERAGE GROUPINGS/TASK

FIGURE 7 is a box and whisker plot of the LOGARITHMIC AVERAGE GROUPINGS/TASK

FIGURE 8 is a histogram of the LOGARITHMIC AVERAGE TIME/TASK

FIGURE 9 is a box and whisker plot of the LOGARITHMIC AVERAGE TIME/TASK

FIGURE 10 is a scatter plot the AVERAGE GROUPINGS/TASK AND NUMBER OF PREFERRED MUSICAL GENRES (with LINE OF CORRELATION)

FIGURE 11 is a scatter plot of the AVERAGE TIME/TASK AND NUMBER OF PREFERRED MUSICAL GENRES (with LINE OF CORRELATION)

Introduction

Human beings are able to categorize information in a number of different ways. Such human categorization of information also extends to the manner in humans interpret and perceive audio information. More specifically human beings are able to perceive audio information such as, for example, music, and intrinsically categorize and encode the audio information. One example of such human interpretation of music is referred to as auditory grouping.

Music can be described as a waveform having repetitive and similar sequences. The ability of humans to recognize these repetitive and similar sequences is crucial to the process of auditory grouping. Furthermore, such grouping of audio information enables humans to recognize and temporally partition a single piece audio information into a smaller number of musical segments. This process is referred to as segmentation.

Described differently, music segmentation can also be referred to as a process of finding the temporal boundaries or meaningful sections within a piece of music (Velik, 2008). Such temporal boundaries and/or meaningful sections include, but are not limited to, a chorus, a verse, a repeating melody or phrase, and the like. Segmentation of music has value in fields such as music recognition, speech recognition, music synthesis, music information retrieval, neural analysis and many others. Additionally, music segmentation can be used to assist with the development of theoretical systems for analyzing music. Furthermore, musical segmentation can be used as a tool in subjectively defining the manner in which human perception and intuition is utilized to interpret and perceive audio information. Musical segmentation is also well suited to use in algorithms for recognizing a particular individual's musical preferences, and subsequently providing suggested musical content to that particular individual.

The field of machine perception seeks to develop techniques for machines to perceive and interpret stimuli in ways similar to humans. In the field of machine perception, auditory perception refers to the set of problems related to how humans perceive audio, like music. A key problem in auditory perception is auditory grouping which is the way human perception breaks a continuous stream of perceived audio into chunks, or groups. In the context of machine perception, the auditory grouping problem requires taking in an arbitrary piece of audio and predicting the groups, represented by a start time and a duration, present in that audio. Additionally, grouping grammars can specify categories of auditory groups, and models over those grammars can predict the classifications of identified groups (Jackendoff and Lehrdal, 1981).

Our objective is to use the results of the auditory groupings performed by human subjects, and the voluntarily provided demographic data, to generate input data for our machine learning algorithm. Specifically, we intend to use our data collected from the subjects' grouping/segmentation of the audio clips with our machine learning algorithm to enhance the algorithm's ability to emulate human auditory grouping and segmentation. Data on demographics and musical experience will also be used to try to understand differences in grouping/segmentation behavior.

Motivation

The potential applications of an accurate auditory grouping model are countless. A model which can break a piece of music into perceived chunks could be used by musicians to better compose music by supporting algorithmic tools which operate on perceived phrases of notes as a unit. Other applications include big-data analysis of large auditory repositories, where individual review of pieces is out of scope. A music streaming service, for example, might prefer to shard their audio-streams along group-boundaries so temporal anomalies due to low-bandwidth stream latency occur during natural grouping breaks. Another application could be to support a new class of algorithm which operates on auditory data with understanding of human perception of that data; a compression algorithm could preserve information about groups as individual units, and deprioritize the noise between groups.

Background

Sophisticated computational models have been developed to replicate the ability of humans to recognize segments of music and perform auditory grouping and segmentation of music (Schlüter and Grill, 2015). Despite the ease with which humans are able to determine or recognize segments of music and ultimately divide or segment music, computational models for musical cognition are often comprised of complex algorithms which are extremely computationally intensive. Many existing computer-based approaches to musical cognition require complicated transformations of music prior to computer analysis. In one example, for a given musical selection, each note of the musical selection is first transformed into its respective spectrum. Each transformed note is then able to be algorithmically processed (Jackendoff and Lehrdal, 1981). Other computational models and/or approaches detect repeated patterns of features within a musical selection using clustering or novelty detection algorithms. By merging the ability of humans to recognize segments of music and perform auditory grouping and segmentation of music with a computer-based computational approach to music analysis, a hybridized and comprehensive model for auditory grouping can be realized.

Novelty

There is existing research on the auditory grouping problem in machine perception (Cambouropoulos, 2006). Existing techniques primarily focus on trying to construct a model of human perception of auditory grouping, either prescribing to Gestalt theory or intentionally in refutation of it, and then constructing a machine to interpret supplied waveforms towards the specification of that model (Szabó et al., 2016). Most of these models are unsupervised models, which is to say they do not test themselves against human-labelled data (Zhuang et al., 2020). It is our belief that for a task so fundamentally oriented towards subjective human perception, training (and then validating) a model on human generated data is critical. This validation against real perceptual actors has been beneficial in verifying other auditory perceptive phenomena, like the Iambic-Trochaic law (Spierings et al., 2017). Thus, we offer our specific contribution to the auditory grouping problem: applying supervised, contemporary, deep-learning convolutional models to human produced auditory grouping data.

Project Design

There are two components of the design of our project, the experimental design for data collection, and the exploration space for the machine learning modeling work. We address each in turn.

Experimental Design for Data Collection

Subjects

We recruited human subjects from two separate pools. The first pool from which we recruited survey participants is the Worcester Polytechnic Institute (WPI) Research Participation Pool comprised of WPI undergraduate students in psychology courses fulfilling a research participation requirement. The second pool from which we recruited survey participants is Amazon's Mechanical Turk (Mturk). Survey participants from the WPI Research Participation pool received course credit, while the survey participants from the MTurk pool received $3.50 for their participation.

A total of 375 participants completed our survey study from both Amazon's Mechanical Turk (Mturk) and the WPI SONA participant pool, with only 49 responses coming from SONA. From the 375 received survey results, 121 survey results were excluded from further analysis. Exclusion of the 121 survey results was based on factors indicating that the survey was not properly completed. The primary exclusion factor was a time of completion for the survey which was less than the time required to listen to the provided audio samples in the survey. The other exclusion criteria used in the data analyses were: too short of a timestamp indicating an impossibly short time spent on the survey, those who placed less than 2 groupings on a given audio clip, and those who placed more than 20 groupings per clip. After excluding the 121 survey results, we obtained survey results from 254 total participants. Of the 254 survey participants, 87 of the survey participants voluntarily provided demographic information. This was in part to design issues with our survey, as in the process of completion, subjects had to navigate between different browser tabs. Because the grouping task existed separately from Qualtrics, a link must be followed that then provided an identifier key to type into the Qualtrics program back on the original tab. This left a great deal of room for error, resulting in a low success count relative to total subjects. All participants provided informed consent prior to beginning this study. The informed consent statement is included in Appendix A: Informed Consent.

Design and Materials

For our study, we are aiming to collect survey results from participants performing auditory grouping tasks. We intend to use the results of the human-performed auditory grouping as input data for our machine learning algorithm. Our intent is to improve the capability and accuracy of the machine learning algorithm to recognize the same grouping cues as would be yielded from human-performed auditory grouping. Our materials will include 100 audio clips each of which is 10 seconds in length. The 100 audio clips were selected from music spanning a diverse selection of music types. The independent variable here is the auditory clip stimulus given to a participant, as well as any specific meta-data aspects of the sound bite. Our dependent variable is the way in which the clip has been grouped, first without any consideration for whether the groupings are “correct” or not.

Additionally, an objective this study is a hybridization of data collection, to serve a primary and secondary goal. The primary goal, consisting of the bulk of the survey, will seek to gain consistency, repetition, and quantity of data from each individual participant to be fed to a machine learning algorithm. The secondary goal, consisting of a preliminary section of the survey focusing on demographic and other sorting questions, will seek to provide a basis by which traditional data analysis can be performed for different groups with the goal of using the survey data to gain insights into how/why people make the groupings the way they do.

The bulk of the survey is designed to collect masses of data for the algorithm. Survey participants are provided with 20 unique, 10 second audio clips. The 20 unique, 10 second audio clips are repeated 5 times such that the survey participants will listen to a total of 100, 10 second audio clips. The survey participants are then asked to click a device (e.g. computer key or mouse button) when they think they recognize a specific grouping. This will be visualized by a slider on a flat audio form that does not reflect the clip the participant is hearing. These survey pages were set to automatically move on to the next audio clip once the 10 seconds is through. This resulted in a total survey time of approximately 16.6 minutes, not including time spent on the optional demographics questions page.

The secondary goal of the survey was achieved by adding a section of demographic questions to the end of the survey. Survey participants were provided with the option to provide or not provide answers to the demographic questions. The demographic questions included inquiries regarding gender, age, music experience, music genre preferences, and other such relevant factors for later sorting of data. These questions serve as a basis for performing data analysis on the grouping data. In this way, the survey accomplished its primary goal of obtaining a large quantity data for training the algorithm, while also yielding a great deal of raw information to be later grouped and analyzed in accordance to the answers on the demographic section. The demographic questions are provided in Appendix B: Survey Flow and Questions.

Computational Model Design

It should be noted that a separate comprehensive report on the computational methodologies described herein has been prepared and is available (Jan and Chen, 2022).

Preprocessing

In order to achieve the most learnable surface from our data collection, techniques to stretch the data to a larger scope can be employed. These techniques are comparable to the techniques employed in image recognition like rotating or scaling images during classification. For our purpose, prior art indicates that pitch shifting and high/low-pass filters are relevant.
Notably, actual human perception may not be invariant under these modulations, but that is irrelevant if it offers a beneficial stepping stone for the learning algorithm; we'd still be testing on ground-truth data.

Targeting

Additional decisions can be made in regard to the output of the model. A new technique in onset detection which will be worthwhile for us to explore is to, instead of generating specific starting and ending times for groupings, the model would generate expected time-until-group and time-since-group for each audio frame, and use these to map a probabilistic model of groups.

Convolutional Neural Networks

Convolutional neural networks are neural networks that are mostly used to analyze digital images. In order to parse audio using convolutional neural networks, it is important to transform the data into something the neural network can understand. A few common ways to transform the data that we will try is Mel Spectrogram (MFCCs) Spectral Bandwidth Spectral Centroid Chromogram.

Short-time Fourier transforms

Transforming the audio through these means will allow the NN to parse the audio as an image. One strategy we will employ is using each one of these transforms in its own neural network. We would then compare the accuracy between models to determine which one is best suited for us. Another alternative we can use is making use of channels. CNNs have support for multiple color channels. We could stitch each one of these transforms together so the CNN can learn about all of the transforms with one model.

One instantiation of the CNN we can consider is a modification of the YOLO (You only look once) model for image boundary detection. A modification called the YOHO (You only hear once) is used for audio segmentation and is worth considering for the project.

Recurrent Neural Networks

Recurrent Neural networks are neural networks that connect some of the output back into the input of the neural network. This allows for the network to take into greater consideration of previous information and time. Similar to a CNN, RNN's need to have the data transformed into something the Neural Network can understand. We will employ the same transformations as the CNNs to transform the data.

A class of RNN's called LSTM (long short term memory) has been used successfully in auditory machine learning. A kind of LSTM called BLSTM (bidirectional long short term memory) has been used for audio segmentation which we will test out for our model.

Combination CNN's and RNN's

In most combination designs, the data was first passed through a CNN, then through an RNN before passing through a dense layer. In more recent combination designs, a recurrent convolutional layer has been implemented, making use of both CNN's and RNN's features in one layer.

Cross-validation.

While cross-validation is not a machine learning model itself, our small dataset encourages us to use cross-validation in our models. Cross-validation is a technique where you separate the data into N-equal parts and then train a model n times with a portion of the model comprised of the testing data and training data. This allows us to use the entire dataset as our training data and reduce bias in our model.

We will explore model designs employing a spread and combination of convolutional and recurrent models targeting a probabilistic model of grouping bounds. To accommodate our traditionally quite small dataset, we will explore data-preprocessing techniques to add samples and cross-validation metrics to more efficiently allocate test-data.

Computational Methods

Spectrogram transformation

MP3 files encode audio in samples, where each sample represents an instance in time. The sample rate of a clip loaded in librosa, the python library we used to parse data, is 22050. The clips we used to analyze groupings are exactly ten seconds long, meaning the data has a total of 220500 samples. Going through the samples using a dense neural network, a recurrent layer, or even a 1d-convolution would take a long amount of inference time. Transforming the data into a spectrogram, an image representation of the clip, would allow us to decrease inference time and also allow us to use more powerful layers like a 2d-convolution.

The spectrograms we used were a melspectrogram, a chromogram, a mel-frequency cepstrum (MFCC), and tempogram. All of the spectrograms were created with 128 bins, and transformed from power to decibels. Creating each of the spectrograms resulted in a 431 by 128 image. The images were then layered on top of each other to create a 431 x 128 x 4 array.

Supervised Pretraining

Due to the small size and limited scope of our dataset, we opted to try supervised pretraining on a larger dataset. The dataset we chose is a small section of the free music archive (FMA). The section we chose for the FMA contains 8000 music clips from numerous genres. The music clips are approximately 30 seconds long and each audio clip was split into 3 clips of 10 seconds each.

A model was trained to classify these music pieces into its various genres. After training was completed, the fully connected layers were removed from the model, and a new set of fully connected layers was appended on to resume training on the auditory grouping problem.

Google c/oud/TPU's

The datasets for supervised pretraining and auditory groupings were too large to fit in ram. This meant that data had to be loaded through chunks during training. Normally this would be simple to do on CPU or GPU training, but that requires data to be stored on google cloud if you want to train on TPUs. Because training on TPU's is significantly faster than training on GPUs and CPUs, it was necessary to get training done by a reasonable time.

To overcome these issues, we first encoded the spectrograms as an rgba image, then we stored the images in TFRecord files. TFRecord files allow for multiple instances of training datum to be store in a single file, allowing the TPU to get multiple instances of training datum with a google cloud call.

Fuzziness

Instead of encoding each group as a single 1 in an array, we could surround the groups with l's creating three contiguous l's representing a group. This helps make the outputs less sparse and the machine will output nicer probability distributions.

Left/Right aligning

Because our testing data was all 10 seconds long, we created our model to accept 10 second audio clips. All of our training data however were 8 seconds long. To overcome this, we first aligned the audio clips with the start of the 10 seconds and padded the end with silence. We then aligned the audio clip with the end of the 10 seconds and padded the beginning with silence. Left/Right aligning gives us the benefit of doubling our training data and also allows us to train without unfairly biasing one end of the audio clips over the other.

Analysis and Discussion of Human Subjects Data

*Machine learning analyses included in colleagues' report (Jan and Chen, 2022).

Demographic and Sorting Questions

Using the survey format conducted through SONA Systems and Amazon's Mechanical Turk (Mturk), with a financial incentive, a total of 375 participants completed our survey study. Survey results were generated from January 21, 2022 through March 3, 2022.

Further, the same 87 survey participants also indicated which musical genres are common to their favorite playlists. The total survey time was approximately 16.6 minutes, not including any time spent responding to the voluntary responses to the demographic and other sorting questions.

All of the demographic and other sorting questions are included in Appendix B: Survey Flow and Questions.

The gender breakdown of the 87 survey participants was 33.3% female,64.2% male and 2.5% non-binary.

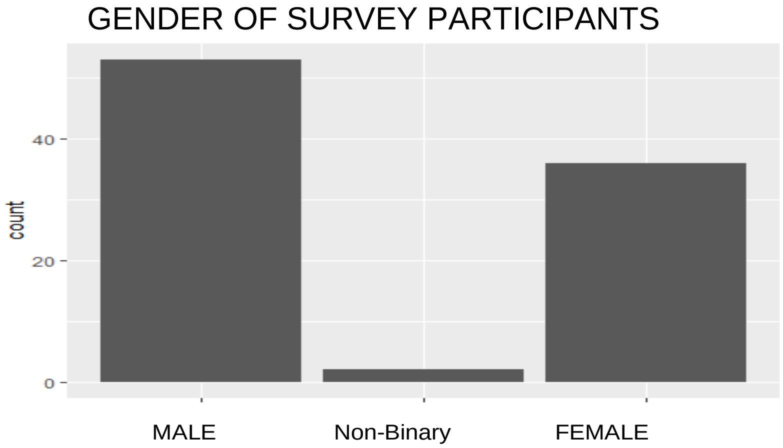
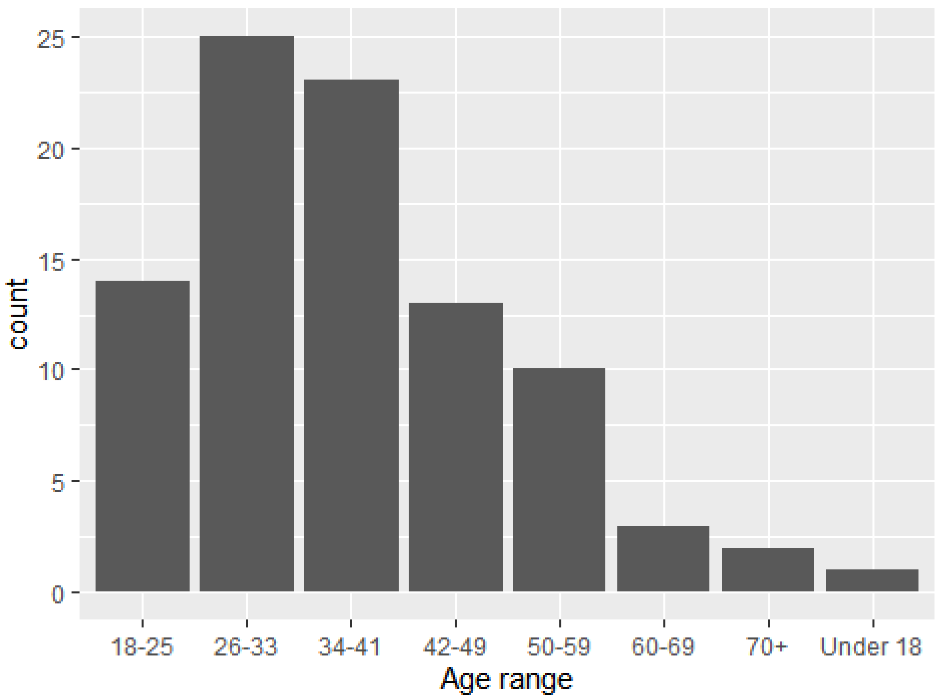


FIGURE 1

The observed breakdown of gender for the 87 participants is nearly identical to the reported gender distribution of the WPI student body, from which many of the survey participants were selected. According to the website datausa.io, the gender breakdown of the enrolled WPI student body is reported as: 63.5% Male and 36.5% Female.

AGE DISTRIBUTION OF SURVEY PARTICIPANTS Figure 2



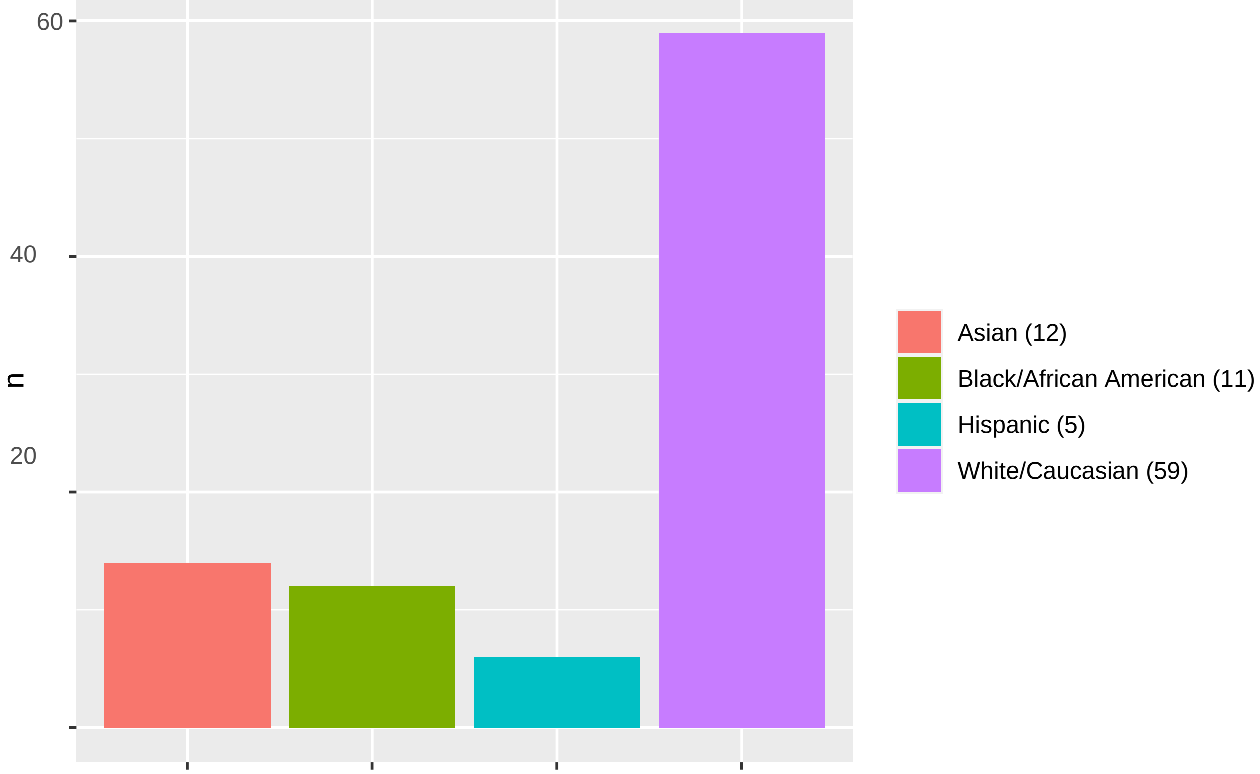
The age demographics for the WPI student body, according to the website datausa.io, is given as 91% of the WPI student body are under the age of 30. Clearly, such a concentration of people under the age of 30 makes it challenging to obtain a diversely aged sample of survey participants from the WPI student body pool.

Additionally, the age demographics for the pool of MTurk respondents is given as 37% of people on MTurk are in their 30’s, another 17% are in their 40’s, and roughly 11% are in their 50’s (Moss, 2020). Despite *some* variability, and certainly more age variability than the WPI student body pool, MTurk still skews significantly younger than the U.S. population as a whole. This skew makes it challenging to obtain a diversely aged sample of survey participants from the MTurk pool.

The breakdown of race/ethnicity for the 87 survey participants who responded to the voluntary demographic questions, is graphically illustrated in the following bar plot.

RACE/ETHNICITY OF SURVEY PARTICIPANTS

Figure 3



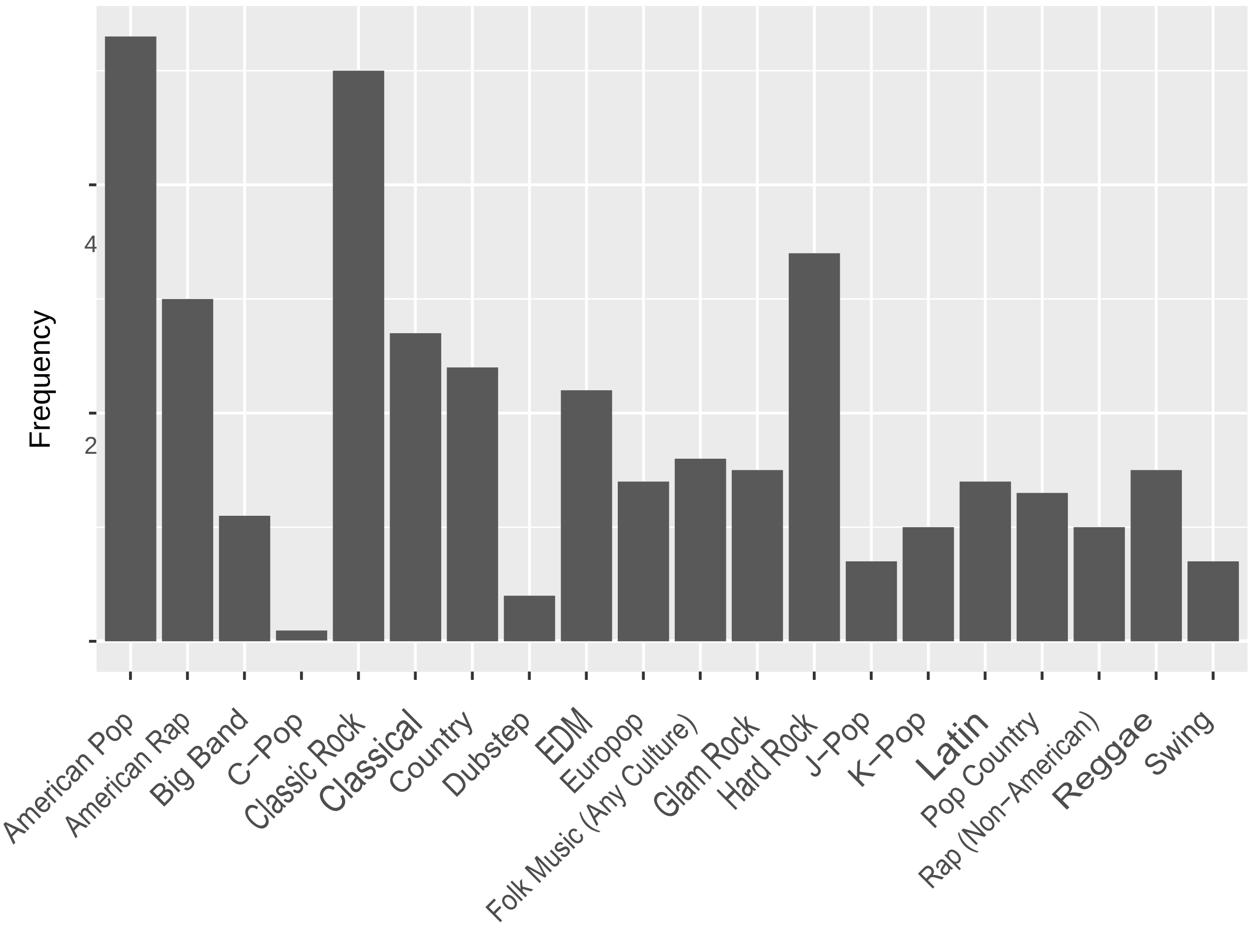
The observed breakdown of race/ethnicity as a percentage of the 87 survey participants is 72.8% White/Caucasian, 14.8% Asian, 13.5% Black/African American and 6.2% Hispanic.

The observed breakdown of race/ethnicity observed for the 87 participants is approximately consistent with the race/ethnicity of the WPI student body, from which the survey participants were selected. According to the website datausa.io, the race/ethnicity of the enrolled WPI student body is reported as: 56.5% White, 7.28% Hispanic/Latino, 5.99% Asian, and 2.89% Black or African American.

Although there are slight differences between the observed breakdown of race/ethnicity as a percentage of the 87 survey participants and reported statistics on the race/ethnicity of the WPI student body, such differences can be attributed to the voluntary aspect of the demographic portion of the survey, compared with a census reporting of all enrolled students at WPI.

Additionally, the 87 survey participants responded to the question “What genres are common to your favorite playlists? (Select all that apply)” as shown below. The survey participants were allowed to select as many of the 20 provided musical genres as desired.

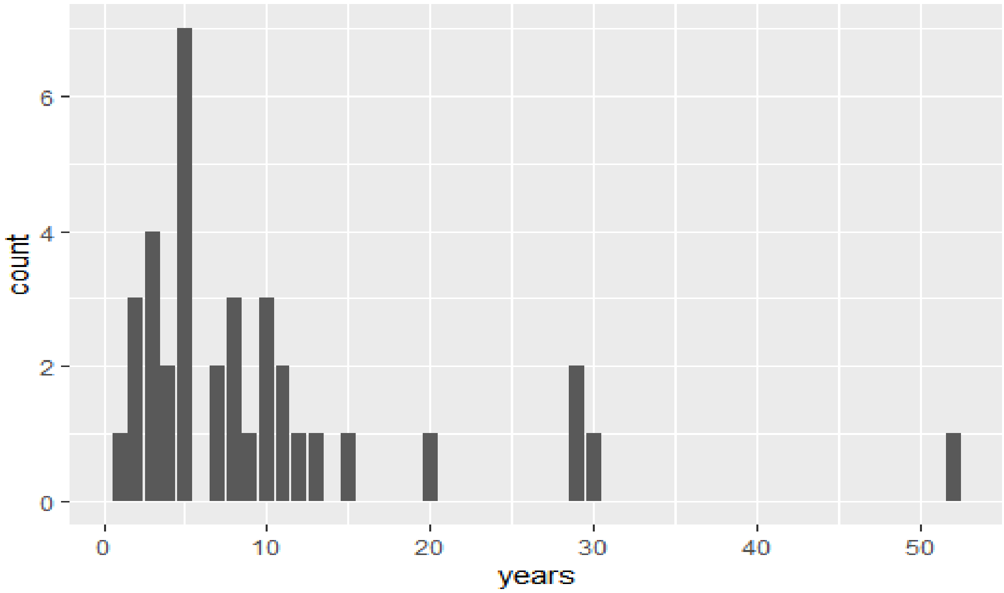
PREFERRED MUSICAL GENRES OF SURVEY PARTICIPANTS Figure 4



We observed that the three musical genres most frequently selected as a preferred musical genre were American Pop, Classic Rock, and Hard Rock.

The 87 survey participants also responded to the question, “If you play a musical instrument, how many years of experience do you have?” From the 87 survey participants providing demographic information, 36 of the survey participants indicated that they did play a musical instrument. The years of musical experience for the 36 survey participants is illustrated in the below bar plot.

MUSICAL EXPERIENCE OF 36 SURVEY PARTICIPANTS Figure 5



We observed that approximately 41.2% (36 of 87) of the survey participants providing demographic information had musical experience, and 28 of the 36 survey participants had more than one year of musical experience.

Auditory Groupings

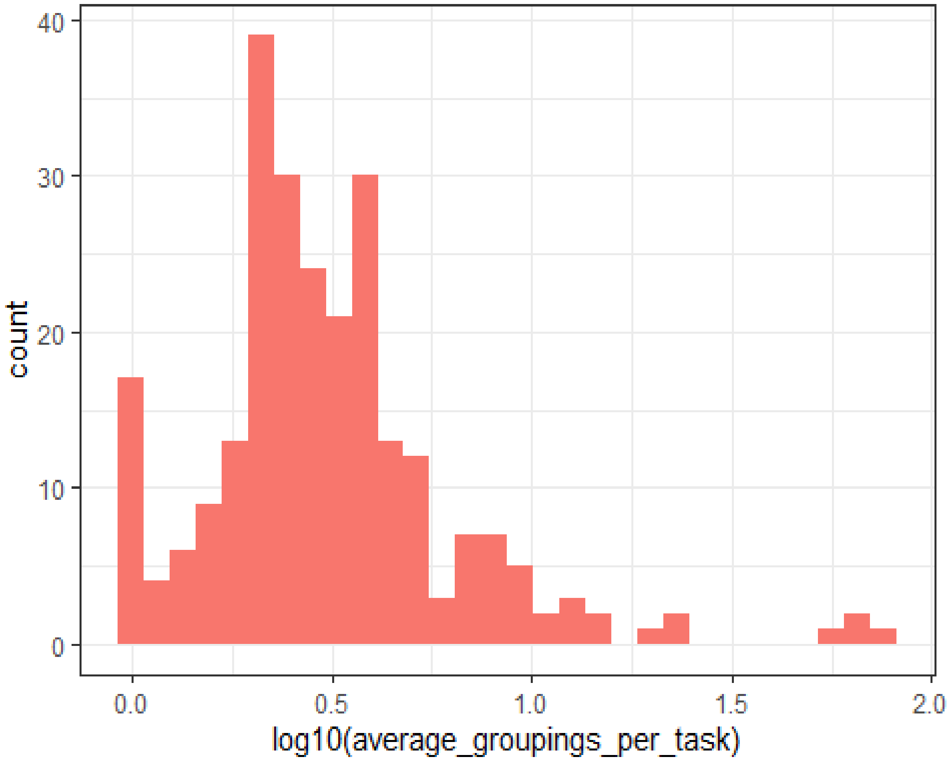
When performing the auditory grouping and segmentation, while listening to the 100, 10 second, audio clips, each of the 254 human subjects were provided with a graphical representation indicating the progression of each 10 second audio clip. When a survey participant perceived a grouping point within the audio clip, the survey participant would depress a computer keyboard key. Although straightforward, we observed that the process of recognizing a grouping point within the audio clip and immediately depressing the computer keyboard key was challenging. Several subjects also mentioned in an open-ended survey response that the task of indicating perceived auditory groupings was difficult to learn. 75 subjects gave responses to the question asking, “Please take a moment to describe in more detail how you decided to group each music clip. (percussion, melody, instrumentals, etc.).” Of these 75, only some were useable. Interestingly, there was very little repetition in how people chose to group the clips, with a wide variety of strategies stated. We also found that when we performed the auditory grouping task, prior to creating the survey, there was a fairly steep learning curve in completing the grouping process. 31 subjects reported that they found the grouping task to be difficult, which is just under half of the subjects that answered the open-ended questions. The main 2 points of feedback were that the rapid pace of the successive tasks made precise groupings difficult, and second that some of the clips were somewhat esoteric so placing confident grouping markers was hard to determine.

Average Groupings Per Task (Groupings per each 10 Second Audio Clip)

The 254 survey participants who listened to the 100, 10 second, audio clips and indicated when they recognized a specific grouping in each 10 second audio clip yielded the following results.

The entire data set corresponding to the 254 responses related to average groupings per task for the 100, 10 second, audio clips is included in the supplemental information filed along with the present report.

HISTOGRAM OF LOGARITHMIC AVERAGE GROUPINGS/TASK Figure 6



(DESCRIPTIVES BELOW ARE NOT LOGARITHMICALLY ADJUSTED)

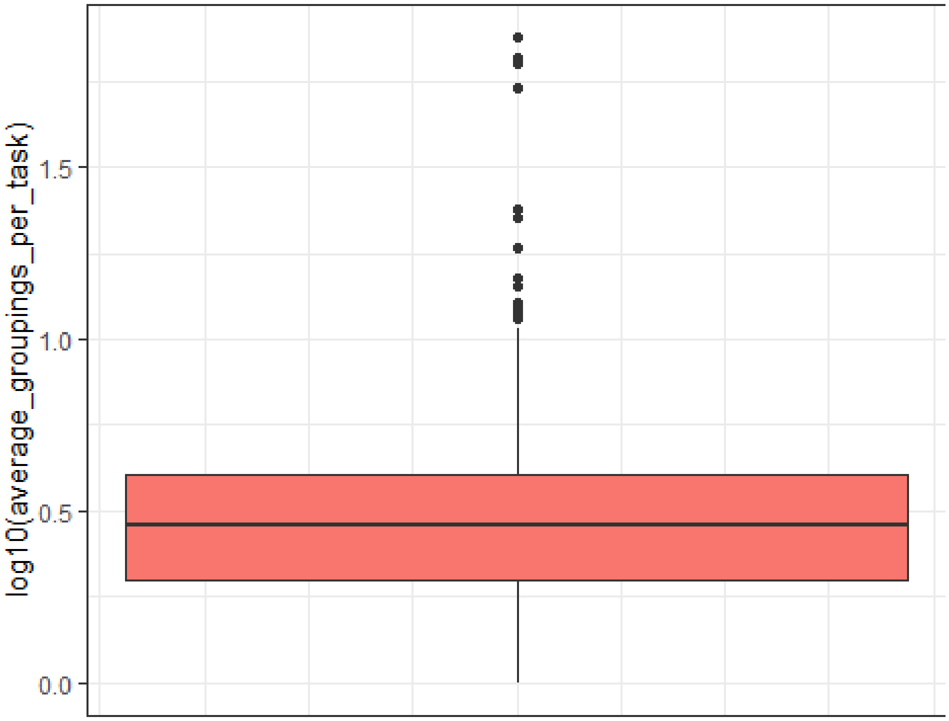
```
Summary(Average_Groupings_per_task)

Min.     1st Qu.    Median     Mean    3rd Qu.    Max.
1.000    2.000      2.868      4.641    4.000      75.684
```

Due to the variation in the data corresponding to the average number of groupings for each 10 second audio clip, referred to herein as a “task”, it was necessary to depict the average groupings per task in logarithmic form as shown in Figure 6, above.

Analysis of the data, including the results for the Mean value, Median value, First Quartile and Third Quartile values (values of which are not logarithmically adjusted) enables the team to recognize outlier data which may or may not be valid data for use by the machine learning algorithm.

To further identify potential outlier data, a Box-and-Whisker plot of the logarithmically-adjusted average groupings per task yields the following:

BOX-AND-WHISKER PLOT OF LOGARITHMIC AVERAGE GROUPINGS/TASK Figure 7

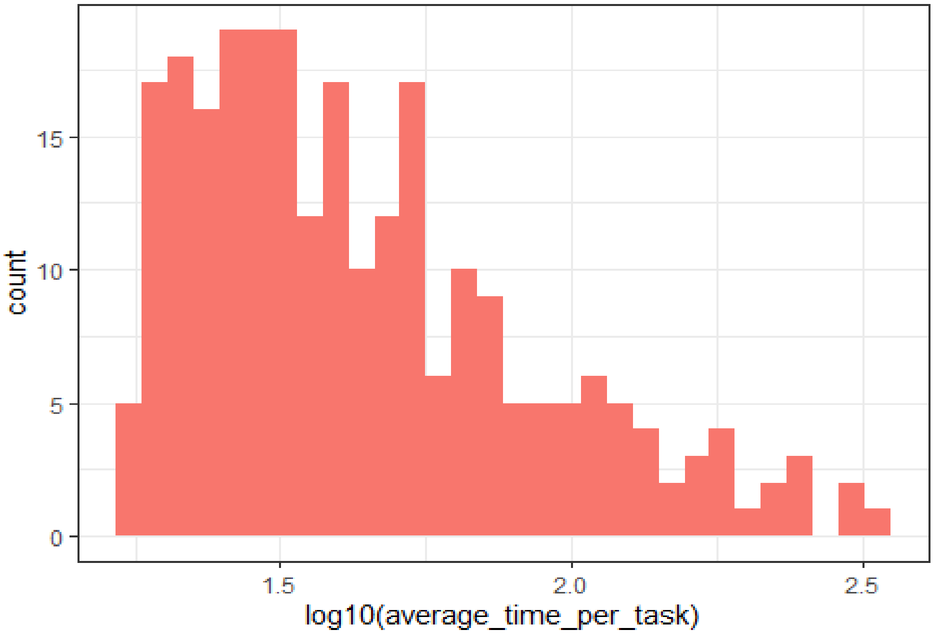


The logarithmically-adjusted Box-and-Whisker plot of Figure 7 clearly indicates the several outliers are present in the results from the 254 survey participants, and that such outlier data values may need to be considered for validity prior to inclusion of the outlier data values in the data set for the machine learning algorithm.

Average Time Per Task (Average Time between each Grouping in each 10 Second Audio Clip)

For the 254 survey participants who listened to the 100, 10 second, audio clips and indicated when they recognized a specific grouping in each 10 second audio clip, we collected data regarding the duration of time between each indication of a grouping. Although each audio clip is only 10 seconds in length, survey participants were allowed to pause or rewind an audio clip to facilitate their identification of a perceived auditory grouping and to ensure that their desired indication of such a grouping location was accurately recorded.

HISTOGRAM OF LOGARITHMIC AVERAGE TIME/TASK Figure 8



(DESCRIPTIVES BELOW ARE NOT LOGARITHMICALLY ADJUSTED)

```
Summary(Average_time_per_task)

Min.         1st Qu.      Median       Mean       3rd Qu.       Max.
17.05        25.42        38.06        57.19      64.08         328.60
```

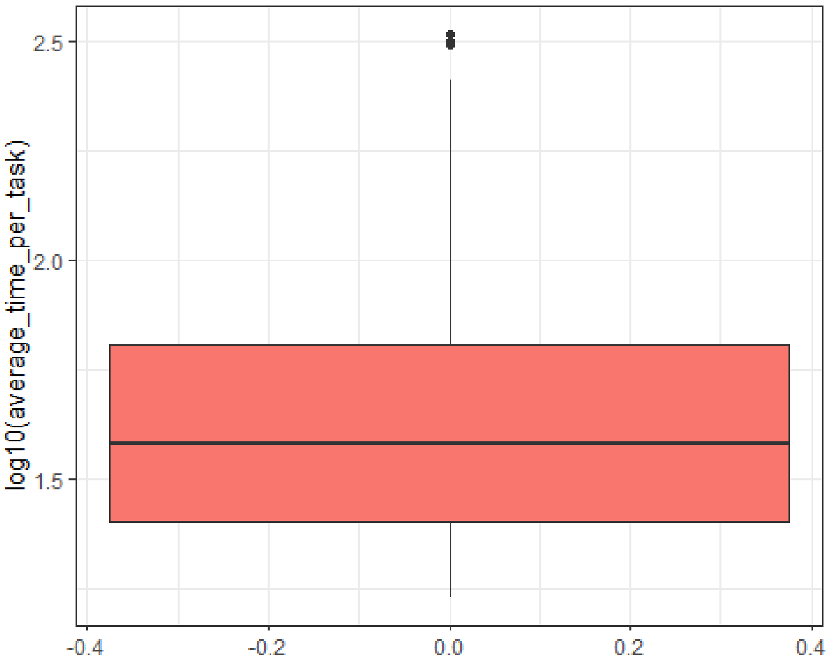
Analysis of the data, including the results for the Mean value, Median value, First Quartile and Third Quartile values (all of which are not logarithmically adjusted) enables the team to recognize outlier data which may or may not be valid data for inclusion in the data set to be used by the machine learning algorithm.

Because the survey participants were allowed to pause or rewind an audio clip, the average time per task is actually longer in duration than the 10 second length of the audio clip.

Due to the breadth and spread of the data corresponding to the average time per task for each 10 second audio clip, it was again necessary to depict the average time per task in logarithmic form as shown in Figure 8, above.

To further identify potential outlier data, a logarithmically-adjusted Box-and-Whisker plot of the average time per task (duration between indicated groupings) yields the following:

BOX-AND-WHISKER PLOT OF AVERAGE GROUPINGS/TASK Figure 9



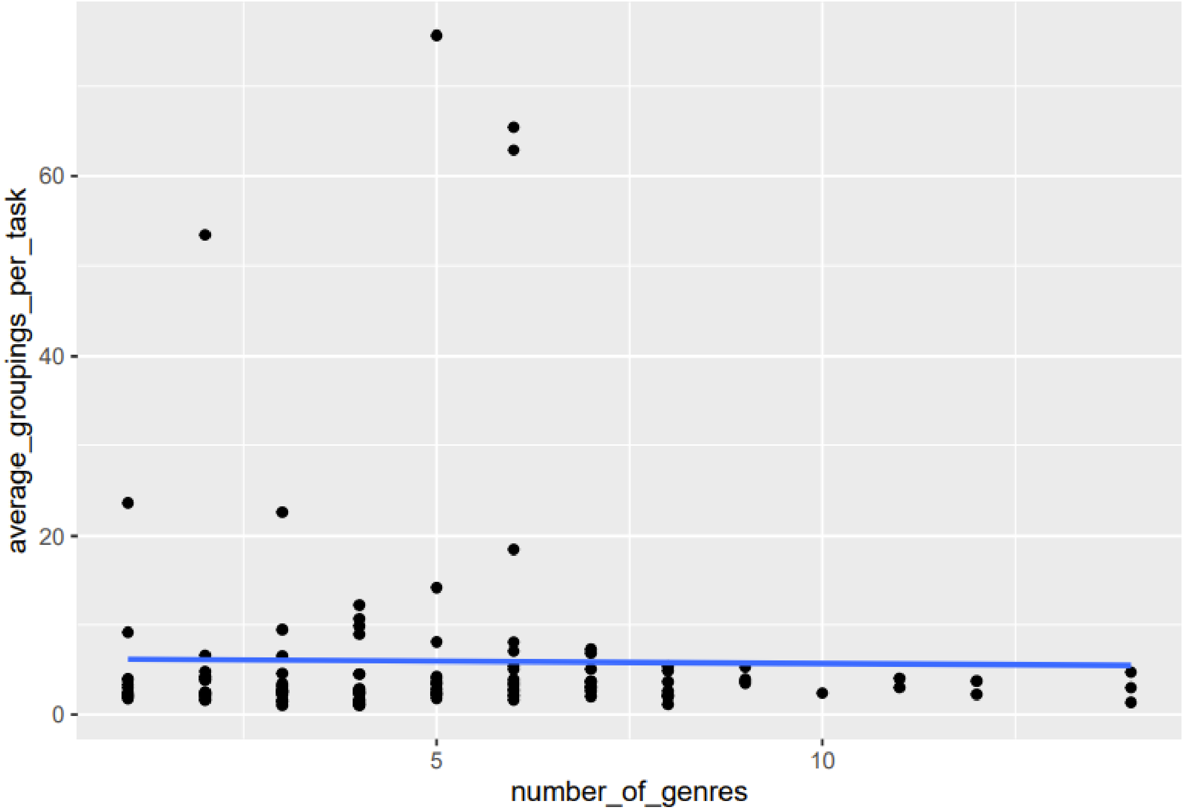
The logarithmically-adjusted Box-and-Whisker plot of Figure 9 clearly indicates the several outliers are present in the results from the 254 survey participants, and that such outlier data values may need to be considered for validity prior to inclusion of the outlier data values in the data set for the machine learning algorithm.

Correlating Demographic Data with Grouping Data

At Figure 10, we provide a scatter plot comparing the average number of groupings per task with the number of preferred musical genres. Additionally, Figure 10 includes a solid “line-of-correlation” to graphically indicate the degree of correlation (depicted by the slope of the “line-of-correlation”) between the average number of groupings per task with the number of preferred musical genres indicated by the subjects. We hypothesized that subjects with broader musical experience would indicate more groupings per task.

Due to the larger quantity of data values obtained from a subject’s selection of number of preferred musical genres, as compared to other provided demographic data values, we chose to plot the number of preferred musical genres along with the average number of groupings per task. The four uppermost data points shown on the graph could be considered outliers, but not statistical test to determine the impact of outliers on the analysis was conducted.

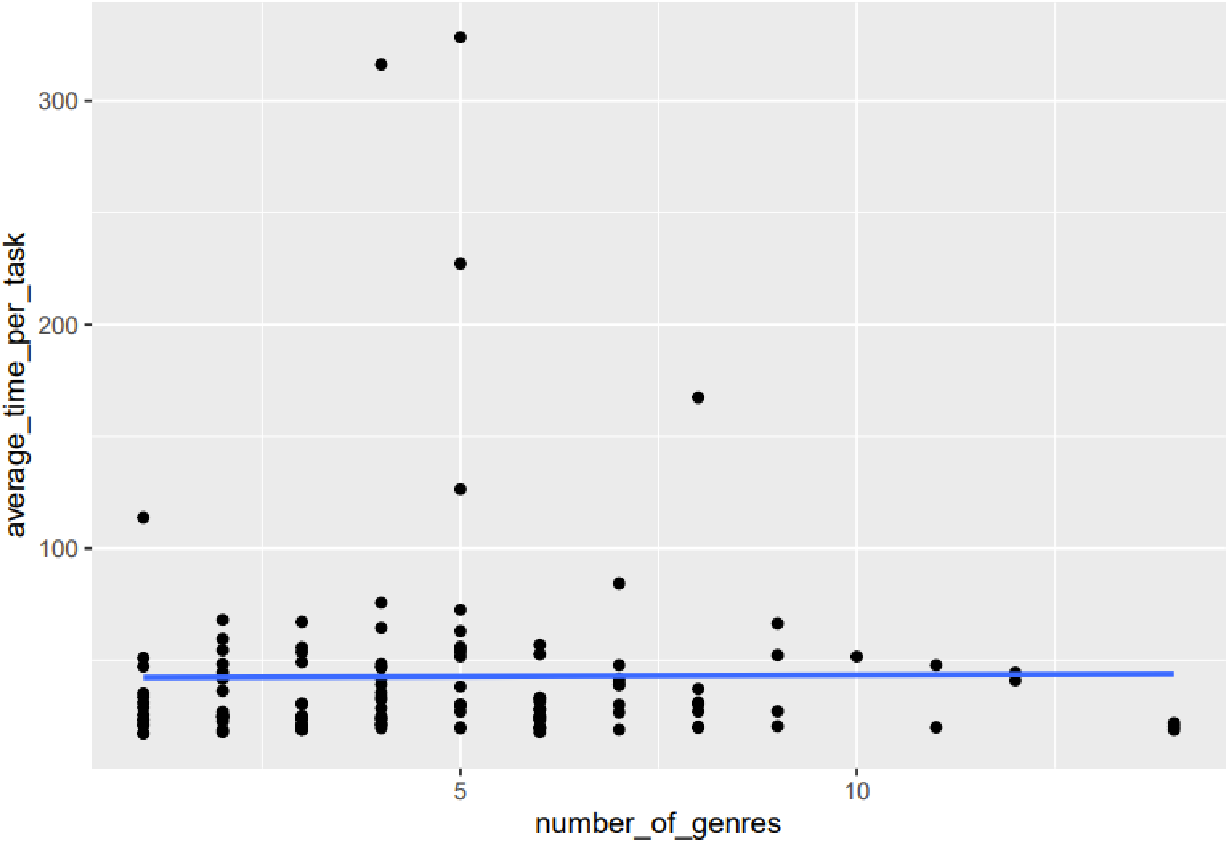
SCATTER PLOT OF AVERAGE GROUPINGS/TASK AND
NUMBER OF PREFERRED MUSICAL GENRES (with LINE OF CORRELATION) Figure 10



Our evaluation of the data yielded a Pearson's product-moment correlation of 0.112093 when comparing the average number of groupings per task with the number of preferred musical genres. There were 137 subjects, and the p-value is .07512. Thus, we found a slight correlation between these two variables. Since the p-value is only slightly above .05, this indicates that there is a small positive correlation between the average groupings placed per task, and the number of musical genres selected.

At Figure 11, we provide a scatter plot comparing the average time per task with the number of preferred musical genres. Additionally, Figure 11 includes a solid “line-of-correlation” to graphically indicate the degree of correlation (depicted by the slope of the “line-of-correlation”) between the average time per task with the number of preferred musical genres indicated by the subjects. We hypothesized that subjects with broader musical experience would indicate spending more time on the task.

SCATTER PLOT OF AVERAGE TIME/TASK AND
NUMBER OF PREFERRED MUSICAL GENRES (with LINE OF CORRELATION) Figure 11



Our evaluation of the data yielded a Pearson's product-moment correlation of -0.202196 when comparing the average time per task with the number of preferred musical genres. There were 137 subjects, and the the p-value is .9331. Thus, we found little correlation between the two variables of average time per task and the number of preferred musical genres, but did find a slight positive correlation between the number of groupings placed, and the number of genres selected by subjects.

Conclusion

We conducted a study wherein the participants were primarily members of the Amazon MTURK pool, with a minority from the Worcester Polytechnic Institute (WPI) student body. Some of the survey participants were incentivized. The survey participants were provided 100 10 second audio clips and were requested to perform auditory grouping and segmentation of each of the provided audio clips. The survey participants were also asked, on a voluntary basis, to answer questions related to their demographic characteristics. We obtained 254 valid survey responses, and of the 254 valid survey results, 87 of the survey participants also voluntarily answered the questions related to their demographic characteristics. Significantly less subjects answered the musical experience questions, and as such no attempt was made to correlate those data with the grouping variables.

We used the results of the auditory groupings performed by the survey participants, and the voluntarily provided demographic data, to generate input data for a machine learning algorithm (Jan and Chen, 2022). Specifically, data collected from the subjects' grouping/segmentation of the audio clips was utilized by our machine learning algorithm to enhance the algorithm's ability to emulate human auditory grouping and segmentation. Our objective was to demonstrate that it is possible to improve the accuracy of current computer performed audio grouping models with such a hybridized approach (i.e., using human-performed auditory grouping data and demographic data in combination with a machine learning algorithm).

Overall, we found it was difficult to establish a direct correlation between the demographic data and the human-performed auditory grouping of the audio clips. However, there was a slight positive correlation found between the average number of groupings placed on clips, and the number of musical genres selected by subjects. As a result, the data set we generated for the machine learning algorithm was primarily comprised of the human-performed auditory groupings which we received from our human subjects.

Further, due to our inability to establish a direct correlation between the demographic data and the human-performed auditory grouping of the audio clips, we were unable to generate any supportable prediction of human-performed auditory grouping based upon our received demographic data. This may be due in part to the limited number of subjects that answered the demographic questions, or to the influence of outliers in the data.

Suggestions for Future Research

We suggest using a larger survey sample size in the future. We believe that a much larger survey sample would further enhance the capability to determine whether or not a direct correlation exists between the demographic data and the human-performed auditory grouping.

In addition to increasing the size of the survey sample, a more diverse survey sample may also improve the ability to confidently determine whether or not a direct correlation exists between the demographic data and the human-performed auditory grouping. In particular a study comparing musical novices with experts would be very informative.

We believe also that it may be possible to increase the accuracy of the grouping data if the survey participants were first trained in the process of listening to an audio clip and immediately responding upon perceiving a grouping location within the audio clip. The responses collected from the open-ended question regarding grouping task difficulty indicates that the fast pace of the tasks and the esotericism of some of the clips presented a significant challenge for subjects.

We are also confident that the accuracy and precision of survey results could be increased by providing survey participants with an improved user-interface for indicating their intended groupings. The primary route for improving the data collection interface would have been to eliminate the need for separate programs that require the subject to navigate. Because subjects had to navigate between browser tabs, data collection was significantly hindered, especially the questions later in the survey.

References

Cambouropoulos, E. (2006). Musical Parallelism and Melodic Segmentation: : A Computational Approach. *Music Perception: An Interdisciplinary Journal, 23*(3), 249-268. doi:10.1525/mp.2006.23.3.249

Jackendoff, Ray, and Fred Lerdahl. “Generative Music Theory and Its Relation to Psychology.” *Journal of Music Theory* 25, no. 1 (1981): 45–90. https://doi.org/10.2307/843466.

Jan, Cheng-Hsuan, and Chen, Y. (2022). *Auditory Grouping: Using Machine Learning to Predict Locations of Groups in Audio Clips*. : Worcester Polytechnic Institute.

Moss, A. (2020, August 10). *Demographics of people on Amazon Mechanical Turk*. Demographics of People on Amazon Mechanical Turk. Retrieved April 27, 2022, from https://www.cloudresearch.com/resources/blog/who-uses-amazon-mturk-2020-demographics/

Schlüter, Jan, and Grill, Thomas. (2015) “Exploring Data Augmentation for Improved Singing Voice Detection with Neural Networks.” *ISMIR*, 2015, doi:10.5072/ZENODO.243314.

Spierings, M., Hubert, J. & ten Cate, C. Selective auditory grouping by zebra finches: testing the iambic–trochaic law. *Anim Cogn* **20,** 665–675 (2017). https://doi.org/10.1007/s10071-017-1089-3

Szabó, B. T., Denham, S. L., & Winkler, I. (2016, November 15). *Computational models of Auditory Scene Analysis: A Review*. Frontiers In Neuroscience. Retrieved April 26, 2022, from https://www.frontiersin.org/articles/10.3389/fnins.2016.00524/full

Velik, R. (2008, April 01). A bionic model for human-like machine perception. Retrieved September 4, 2021, from https://repositum.tuwien.at/handle/20.500.12708/11039

Zhuang, Yingying, et al. (2020) “Music Genre Classification with TRANSFORMER CLASSIFIER.” *Proceedings of the 2020 4th International Conference on Digital Signal Processing*, 2020, doi:10.1145/3408127.3408137.

Appendix A: Informed Consent

Informed Consent:

Investigators: Zachary Wagner, William McDonald, Yang Chen, Cheng-Hsuan Jan Contact Information: zwagner@wpi.edu, ychen18@wpi.edu, cjan@wpi.edu, and wbmcdonald@wpi.edu

Title of Research Study: Machine Perception Auditory Grouping MQP Advisors: Professor Scott Barton (sdbarton@wpi.edu), Professor James Doyle (doyle@wpi.edu), Professor Gillian Smith (gmsmith@wpi.edu)

You are being asked to participate in a research study. Before you agree, however, you must be fully informed about the purpose of the study, the procedures to be followed, and any benefits, risks or discomfort that you may experience as a result of your participation. This form presents information about the study so that you may make a fully informed decision regarding your participation. The purpose of our study is to collect data on how human beings group pieces of music. To fulfill this goal, you will be listening to 20 different audio clips from various songs, with each recurring 5 times as to allow certainty with the groupings. Each audio clip is ten seconds long and the participant has full discretion to make as many or as few groupings as they want. The volume can be controlled on your device and there will be no risks to the participant.

By participating in this research, you will be aiding in the scientific understanding of music perception, as well as emerging technologies that could benefit from the data. Your responses will be completely confidential, and no data about you is collected other than that which is provided through the music grouping and survey tasks. Records of your participation in this study will be held confidential so far as permitted by law. However, the study investigators, the sponsor or it's designee and, under certain circumstances, the Worcester Polytechnic Institute Institutional Review Board (WPI IRB) will be able to inspect and have access to confidential data that identify you by name. Any publication or presentation of the data will not identify you. If you are participating through WPI's SONA Systems, the appropriate study credit will be applied to you account a few days after study completion. If you are participating through Amazon MTURK, you will receive monetary compensation of 3$.

For more information about this research or about the rights of research participants, or in case of research-related injury, contact: Professor Scott Barton (sdbarton@wpi.edu), Professor James Doyle (doyle@wpi.edu), Zachary Wagner (zwagner@wpi.edu), Yang Chen (ychen18@wpi.edu), Cheng-Hsuan Jan (cjan@wpi.edu), and William McDonald (wbmcdonald@wpi.edu) Also, please feel free to reach out to the IRB Manager (Ruth McKeogh, Tel. 508 831- 6699, Email: irb@wpi.edu) and the Human Protection Administrator (Gabriel Johnson, Tel. 508-831-4989, Email: gjohnson@wpi.edu). Your participation in this research is voluntary. Your refusal to participate will not result in any penalty to you or any loss of benefits to which you may otherwise be entitled. You may decide to stop participating in the research at any time without penalty or loss of other benefits. The project investigators retain the right to cancel or postpone the experimental procedures at any time they see fit. By clicking “continue,” you agree to understanding all of the above information.

Q2 Do you agree to participate in the survey?

○ Yes (1)

○ No (2)

Appendix B: Survey Flow and Questions

Q3 In order to listen to and group the songs, we need to redirect you to a separate link for the first part of our study.

Please open this link in a **new, seperate tab**. (Link on the next page).

At the end of the grouping task, you will return to this study portal. You will be asked to copy and paste a unique code before continuing in this main portion of the study.

Please be prepared to copy the unique code.

Page Break

Q4 Please click HERE to be taken to the grouping task.

(Note: Make sure to do this in a new tab)

End of Block: Block 2

Start of Block: Block 3

Q5 Please input the unique ID you received at the end of the grouping tasks.

____

End of Block: Block 3

Start of Block: Block 4

Q6 Did you find anything particularly difficult about the grouping tasks?

○ Yes (1)

○ No (2)

Q7 If so, please specify.

____

Q8 What type of device did you use for the survey and grouping tasks?

○ Computer (1)

○ Phone/Tablet (2)

○ Other (3)

Q9 Please take a moment to describe in more detail how you decided to group each music clip.

(percussion, melody, instrumentals, etc.)

____

Q10 What is your gender?

○ Man (1)

○ Woman (2)

○ Non-Binary (3)

○ Prefer not to disclose (4)

○ Prefer to describe (5)

Display This Question:

If What is your gender? = Prefer to describe

Q11 Please describe.

____

Q12 What is your race?

○ White/Caucasian (1)

○ Black/African American (2)

○ Asian (3)

○ Hispanic (4)

○ Pacific Islander (5)

○ Other (6)

○ Multiracial (7)

○ Prefer Not to Say (8)

Q13 What is your age category?

○ Under 18 (4)

○ 18-25 (5)

○ 26-33 (6)

○ 34-41 (7)

○ 42-49 (8)

○ 50-59 (9)

○ 60-69 (10)

○ 70+ (11)

Q14 Are you currently a college student?

○ Yes (10)

○ No (11)

○ Graduated College (12)

○ Currently In Graduate School (13)

○ On Leave from College (14)

○ Past Enrollment in College (15)

○ Other (16)

Q15 Have you ever played an instrument for a period longer than 1 continuous year?

○ Yes (1)

○ No (2)

Q16 Do you currently play an instrument?

○ Yes (1)

○ No (2)

Display This Question:

If Do you currently play an instrument? = Yes

Q17 If so, how many years of experience do you have?

____

Q18 On a scale from 1-7, how much would you say LISTENING to music plays a role in your life?

	None at all	A little		A moderate amount	A lot		A great deal
	1	2	3	4	5	6	7
Drag slider ()							

Q19 On a scale from 1-7, how much would you say PLAYING music plays a role in your life.

	1	2	3	4	5	6	7
Drag Slider ()							

Q20 What is your favorite genre of music?

____

Q21 Do you generally listen to music from outside your own culture?

○ Yes (1)

○ No (2)

Q22 Do you generally listen to instrumental music?

○ Yes (1)

○ No (2)

Q23 What genres are common to your favorite playlists? (select all that apply)

□ Latin (1)

□ Reggae (2)

□ American Pop (3)

□ American Rap (4)

□ Europop (5)

□ Country (6)

□ Pop Country (7)

□ Classic Rock (8)

□ Hard Rock (9)

□ Glam Rock (10)

□ Swing (11)

□ Big Band (12)

□ Classical (13)

□ Rap (Non-American) (14)

□ EDM (15)

□ Dubstep (16)

□ Folk Music (Any Culture) (17)

□ K-Pop (18)

□ C-Pop (19)

□ J-Pop (20)

Q29 What genres are common to your favorite playlists? (select all that apply)

□ Latin (1)

□ Reggae (2)

□ American Pop (3)

□ American Rap (4)

□ Europop (5)

□ Country (6)

□ Pop Country (7)

□ Classic Rock (8)

□ Hard Rock (9)

□ Glam Rock (10)

□ Swing (11)

□ Big Band (12)

□ Classical (13)

□ Rap (Non-American) (14)

□ EDM (15)

□ Dubstep (16)

□ Folk Music (Any Culture) (17)

□ K-Pop (18)

□ C-Pop (19)

□ J-Pop (20)

End of Block: Block 4

Start of Block: Block 4

Q24 **MTURK CODE: 98765432**

Debriefing Statement: Thank you so much for completing our survey today. The true purpose of the survey is to gather data on how people group audio clips, to then feed this data to a deep learning algorithm. Our survey was a part of an MQP that is seeking to determine if there are patterns in how human beings group music, that can then be applied to machine learning models to perform the same groupings. We hope that the data we collected will be sufficient to train an algorithm to group audio in the same way humans do. If you have any questions regarding your participation today in this survey, please contact the researchers at zwagner@wpi.edu, ychen18@wpi.edu, cjan@wpi.edu, and wbmcdonald@wpi.edu

End of Block: Block 4