Predicting Human-Reported Enjoyment Responses in Happy and Sad Music

1st Benjamin Ma
Signal Analysis and Interpretations Lab
University of Southern California
Los Angeles, USA
benjamjm@usc.edu

2nd Timothy Greer
Signal Analysis and Interpretations Lab
University of Southern California
Los Angeles, USA
timothdg@usc.edu

3rd Matthew Sachs
Brain and Creativity Institute
University of Southern California
Los Angeles, USA
msachs@usc.edu

4th Assal Habibi
Brain and Creativity Institute
University of Southern California
Los Angeles, USA
ahabibi@usc.edu

5th Jonas Kaplan
Brain and Creativity Institute
University of Southern California
Los Angeles, USA
jtkaplan@usc.edu

6th Shrikanth Narayanan
Signal Analysis and Interpretations Lab
University of Southern California
Los Angeles, USA
shri@sipi.usc.edu

***Abstract*—Whether in a happy mood or a sad mood, humans enjoy listening to music. In this paper, we introduce a novel method to identify auditory features that best predict listener-reported enjoyment ratings by splitting the features into qualitative feature groups, then training predictive models on these feature groups and comparing prediction performance. Using audio features that relate to dynamics, timbre, harmony, and rhythm, we predicted continuous enjoyment ratings for a set of happy and sad songs. We found that a distributed lag model with L1 regularization best predicted these responses and that timbre-related features were most relevant for predicting enjoyment ratings in happy music, while harmony-related features were most relevant to predicting enjoyment ratings in sad music. This work adds to our understanding of how music influences affective human experience.**

***Index Terms*—affective computing, neural networks, multivariate time series modeling, music processing**

I. Introduction

No matter our mood, it always seems possible to find a song to suit it. A happy song can enhance our joy, a sad song can help us reminisce and reflect, and a song that is neither happy nor sad can serve as appropriate background music.

Is there a difference in what we enjoy and respond to in happy and sad songs? What are the dynamics of human perception and other subjective psychological responses to music listening? What is the nature of music enjoyment for a given piece of music? Is it a nonlinear response, best modeled by attention models? How can we best predict changes in enjoyment at a fine temporal level?

To answer such open research questions, we apply a set of multivariate time series (MTS) prediction models to continually-rated enjoyment levels using auditory features as predictors. We compare prediction models and quantify the relevance of each auditory feature for predicting enjoyment in sad and happy songs. We hypothesize that attention models perform best on these tasks, as they can nonlinearly synthesize aural information from previous points in a track to predict subjective human responses.

This work aims at identifying what elements of music affect us the most profoundly. It also offers more insight into how we enjoy sad and happy music. This work can be used in applications that include music emotion recognition and music information retrieval, as well as basic research in understanding human subjective experiences to music.

II. Related Work

A. Music Enjoyment

It has long been wondered why humans seem to universally enjoy music [1]. Enjoyment of music has been studied through many scientific perspectives, including evolutionary psychology [2], physiology [3], audiology [4], and otology [5]. Music enjoyment has concrete applications in music recommendation systems; indeed, some music recommendation systems use user-generated enjoyment descriptions [6]. While past studies have investigated which musical features correlate with various human-reported emotions [7]–[9], as far as we know, there are still no studies that explore which musical features correlate to human-reported *enjoyment ratings*. In this paper, we provide insight into which musical features contribute to models that predict enjoyment responses during music listening.

Kawakami, Furukawa, Katahira, and Okanoya reported that sad music can elicit vicarious pleasure in listeners in the same way that happy music can [10]. It has also been shown that sad songs activate different music processing structures of our brains than happy songs, suggesting that signatures of enjoyment may also be different for happy songs and sad songs [11]. In this paper, we extend this work, investigating enjoyment ratings in happy and sad songs at a fine temporal level. We also compare musical features that contribute to enjoyment in happy music to musical features that contribute to enjoyment in sad music.

978-1-7281-3888-6/19/$31.00 ©2019 IEEE

B. Time Series Modeling

In order to analyze how we experience music in real time, it is necessary to track time-continuous subjective ratings using time series modeling. The most well-known model for linear univariate time series forecasting is the autoregressive integrated moving average (ARIMA) [12], which encompasses other autoregressive time series models, including autoregression (AR), moving average (MA), and autoregressive moving average (ARMA). Another common method for time series prediction, especially in econometrics, is using distributed lag models, which apply ordinary linear regression to MTS forecasting [13]. While effective for some tasks, linear regression models and autoregressive models cannot capture nonlinearity in time series. For this reason, nonlinear models for time-series forecasting based on kernel methods [14], ensembles [15], or Gaussian processes [16] have been introduced. One drawback to these approaches is that they apply predetermined nonlinearities and may fail to recognize different forms of nonlinearity for different MTS. Recently, Long Short-Term Memory systems (LSTMs) [17], variants of Recurrent Neural Networks (RNNs), have also been employed for MTS forecasting. Shih et al. proposed an MTS prediction model called Temporal Pattern Attention-LSTM (TPA-LSTM) which incorporates LSTM, as well as an attention mechanism designed to automatically tune parameters and adapt to nonperiodic and nonlinear datasets [18]. We use this model in our study to determine which auditory features are best correlated with subjective time-continuous ratings of music listening experiences.

III. Data Collection

In order to identify musical stimuli with high affective content, we explored online music streaming sites, such as Spotify and Last.fm, as well as social media sites such as Reddit and Twitter, for songs with social tags containing the words "happy" or "sad" and their synonyms. Social tags provided by users of online music streaming sites have been shown to be an effective method for classifying music based on their emotional content and correlation with acoustic features known to be associated with a particular emotion [19]. To minimize any influence of prior exposure to the songs, we selected songs with "happy" or "sad" tags from the pieces with fewest play counts. This resulted in an initial list of 120 pieces: 60 sad pieces and 60 happy pieces, some with lyrics and some without lyrics. Eight human coders listened to 30-second clips from these pieces and rated whether they conveyed either happiness or sadness. All pieces in which at least 75% of coders agreed on the intended emotion (27 songs total) were then included in an online survey that was completed by 82 adult participants via Amazon's Mechanical Turk. The survey included 60-second clips from these 27 pieces of music and asked participants to rate how much they enjoyed the piece, what emotion they felt in response to the piece (sadness, happiness, calmness, anxiousness, boredom), and how familiar they were with the piece using a 5-point Likert scale. Each participant was presented with only 12 clips of music selected at random to ensure that the Mechanical Turk workers were not overloaded.

Due to the potential confounds associated with the semantic information conveyed through the lyrics of a song [11], we only selected pieces that did not contain lyrics. We additionally excluded pieces that were rated as highly familiar to prevent bias. Based on these criteria from the survey, we selected three pieces of music to be used for this study: (1) a shorter piece that reliably induced sadness (Ólafur Arnalds's "Fyrsta," the "sad short song," 256 seconds); (2) a longer piece that reliably induced sadness (Michael Kamen's "Discovery of the Camps," the "sad long song," 515 seconds); and (3) a piece that reliably induced happiness (Lullatone's "Race Against The Sunset," the "happy song," 169 seconds). The Waveform Audio File Format (WAV) files of these songs were used for playing to participants and extracting auditory features.

A. Participant Selection

Sixty healthy adult participants were recruited from the greater Los Angeles community based on responses to the online survey in which they listened to a 60-second clip of the final three pieces. Only participants who were not familiar with the pieces of music and reported feeling happiness during the clip of the happy song or sadness during the clips of the sad songs were asked to participate. Participants had normal hearing and no history of neurological or psychiatric disorders. All participants were presented with the final three pieces of music in a random order and were instructed to listen attentively to the music and simultaneously report changes in their enjoyment using a fader with a sliding scale. Participants continuously rated their momentary feelings of pleasure from 0 (no pleasure) to 10 (extreme pleasure). The experiment was then repeated with participants asked to rate emotion rather than enjoyment, but only the enjoyment ratings were used in this study. Stimuli were presented using Psychtoolbox for MATLAB [20] and the order of ratings and the order of the pieces were counterbalanced across participants.

B. Auditory Features

Past research suggests that auditory features related to dynamics, timbre, harmony, and rhythm are correlated with affective responses to music [8]. Dynamics refer to "loudness" and change in "loudness" of music, timbre refers to tone quality of music, harmony refers to musical pitches, and rhythm refers to properties of the musical beat.

Seventy-four features that capture dynamics, timbre, harmony, and rhythm were extracted using the MIRtoolbox in Matlab [21]. See Table I. These features were extracted using a sliding window with a duration of 50 ms and a step size of 25 ms, consistent with [22].

Mel frequency cepstral coefficients (MFCCs) were calculated using a Hamming window, pre-emphasis coefficient of .97, frequency range of 100-6400 Hz, 20 filterbank channels, and 22 liftering parameters. Linear prediction filter coefficients (LPCs) were computed using Matlab's lpc function. Compression ratio was computed by taking the ratio between the file

TABLE I
AUDITORY FEATURES USED

Feature	Type	Feature Group
MFCCs 1-12	Timbre	MFCCs
ΔMFCCs 0-12	Timbre	MFCCs
ΔΔMFCCs 0-12	Timbre	MFCCs
HCDF	Timbre	Spectral
Spectral Flux	Timbre	Spectral
Centroid	Timbre	Spectral
Skewness	Timbre	Spectral
Kurtosis	Timbre	Spectral
Spread	Timbre	Spectral
Brightness	Timbre	Spectral
LPCs 0-10	Timbre	LPCs
Chroma 1-12 (C, Db, D, ..., B)	Harmony	Harmony
Key Strength	Harmony	Harmony
Key Mode	Harmony	Harmony
Compressibility	Dynamics	Dynamics
MFCC 0	Dynamics	Dynamics, MFCCs
RMS	Dynamics	Dynamics
Pulse Clarity [24]	Rhythm	Rhythm

size of each window's WAV format and that same window's Free Lossless Audio Codec (FLAC) format, after conversion with ffmpeg [23][1]. We chose to use FLAC instead of MP3 because using MP3 compressibility resulted in higher error for the majority of tests. Key strength was taken as the maximum value of the 24-dimensional output vector from MIR Toolbox's key_strength function. This function outputs a vector containing the probability that a musical segment is in each major or minor key. All other features were extracted using MIR Toolbox's eponymous functions with default parameters.

IV. METHODS

A. Self-Reported Enjoyment Ratings

We averaged the enjoyment annotations at each time step across participants for each song. By inspection, we found that removing the first 30 seconds of annotations (<18% of each song) allowed the annotations to stabilize, as the fader started on the lowest annotation value at the start of each song. We also removed the last 10 seconds of annotations so that the varying fade-out times of each song would not affect the results. We then resampled this signal to 40 Hz to match the sampling frequency of the auditory features.

B. Problem Formulation

Following the example of previous work in MTS prediction, we formulated our problem as such: given a series of auditory feature vectors $X_1, X_2, ..., X_t$ and enjoyment values $y_1, y_2, ..., y_t$, an "attention length" a, and a "horizon" h, we predict $\hat{y}_{t+h}$ using $X_{t-a+1}, X_{t-a+2}, ..., X_t$ [18], [25]. In this task, we set the attention length to be 40 samples (1 second) because it captured a meaningful amount of data without being too computationally expensive for TPA-LSTM. We performed all of our tests with horizons of 40 and 80 samples (1 and 2 seconds), because we empirically concluded this was an appropriate delay between auditory changes and affective reaction in our data.

[1]A higher compressibility indicates that a window's FLAC file is much smaller than that window's WAV file. A lower compressibility indicates that the window's FLAC file is close in size to that window's WAV file.

V. RESULTS

We used six different models to predict enjoyment ratings. The first three were univariate prediction models, and the latter three were multivariate prediction models.

1) Baseline: Autoregression with an attention length of 1 sample (this model predicted $\hat{y}_{t+h}$ using y_t as its only feature).
2) AR: Standard autoregression on the enjoyment rating.
3) ARIMA: Autoregressive integrated moving average model [12].
4) Lasso: A distributed lag model with L1 regularization. The distributed lag model predicted $\hat{y}_{t+h}$ as a linear combination of all feature vectors X for time steps in the attention window:
   $\hat{y}_{t+h} = c + A_1 X_{t-a+1} + A_2 X_{t-a+2} + ... + A_a X_t$.
5) Ridge: A distributed lag model with L2 regularization.
6) TPA-LSTM: Temporal Pattern Attention-LSTM [18].

We conducted a grid search over tunable parameters for TPA. We tried a range of values for the number of hidden units: [4, 16, 32, 64, 256]. We found that using 32 hidden units resulted in good performance and relatively fast training times when predicting rated emotion in the happy song. Following the example of [18], we normalized each time series by the maximum value in itself and used the absolute loss function, an Adam optimizer, and a 10^{-3} learning rate. We used 3 layers for all RNNs, as done in [26]. TPA-LSTM was implemented in Tensorflow 1.9.0, as in [18]. The AR, Baseline, Lasso, and Ridge models were used out of the box from scikit-learn's linear_model package [27]. We conducted a grid search on [1e-8, 3e-8, 6e-8, 1e-7, 3e-7, 6e-7, ..., 1, 3, 6] to find the optimal regularization parameter for Lasso and Ridge in every test. The ARIMA model was used as implemented in statsmodels' arima_model module with lag orders [1, 5, 10], degree of differencing 1, and order of moving average 1.

For all tests, we ran two cross-validation folds: one with the test set as the first 20% of each time series, and the second with the test set as the last 20% of each time series. In each case, the training set comprised of the remaining 80%. The root mean squared error (RMSE) values reported are the mean of the two cross-validation RMSEs.

A. TPA-LSTM

First, we ran a preliminary test comparing TPA-LSTM to the baseline and AR models on all three songs (See Table II). TPA-LSTM used all auditory features as well as previous enjoyment values, making it a model with an autoregressive component. TPA-LSTM failed to complete training on the sad long song at horizon 80, likely due to vanishing or exploding gradients. We used an AR model instead of an ARIMA model for comparison in this experiment because TPA-LSTM does not have a moving average part in its autoregressive component. We found that TPA-LSTM underperformed even the naive baseline in these tests and it was computationally expensive to

TABLE II
VALIDATION RMSE, AFFECTIVE DESCRIPTIONS
(*RMSE values* $\times 10^2$)

Horizon 40	Happy	Sad Short	Sad Long
Baseline	3.026	2.050	1.565
AR	2.406	1.173	1.242
TPA	5.742	2.853	2.968
Horizon 80	Happy	Sad Short	Sad Long
Baseline	4.319	2.240	2.307
AR	3.584	1.790	2.030
TPA	7.235	5.751	N/A

TABLE III
BASELINE & ARIMA TEST RMSE
(*RMSE values* $\times 10^2$)

Horizon 40	Lag Order	Happy	Sad Short	Sad Long
Baseline	1	5.835	7.421	5.500
AR	1	5.84	7.421	5.507
	5	5.846	7.404	5.509
	10	5.829	N/A	5.505
Horizon 80	Lag Order	Happy	Sad Short	Sad Long
Baseline	1	5.961	7.513	5.530
AR	1	5.967	7.513	5.536
	5	5.973	7.495	5.539
	10	5.956	N/A	5.535

run, so we chose not to include it in the rest of the experiments. TPA-LSTM may have overfitted due to the small number of samples in our training set compared to the datasets it was designed for (3,325 samples vs. 31,536 for the solar energy dataset in [18]). Furthermore, the lack of periodicity in the enjoyment signal may have hampered performance.

B. ARIMA

We ran an ARIMA model on each of the three songs. We tested ARIMA at horizon 40 and 80 on lag orders [1, 5, 10] (Table III). ARIMA did not converge on lag order 10 for the sad short song. ARIMA generally performed best with lag order 10, followed by 1 and 5, though in all cases the differences were minimal between ARIMA and the naive baseline ($<$0.5%). Due to this negligible improvement and the rapidly increasing computational requirements of ARIMA at larger lag values, we did not test any higher lag orders. Fig. 1 shows that ARIMA overfitted terrifically during training, leading to poor out-of-sample predictions in the test set. Since the subjects' enjoyment ratings were dependent on an external stimulus (the music) rather than previous ratings, the data seems unfit for an autoregressive model like ARIMA, and we excluded it from the remainder of the experiments.

C. Distributed Lag Models

After running prediction experiments with the purely-autoregressive Baseline, AR, and ARIMA models, we introduced auditory features for the distributed lag models, Lasso and Ridge, to use. In order to identify which features were

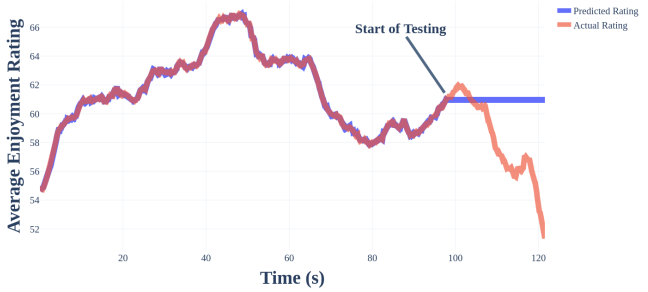

Fig. 1. Average enjoyment rating and prediction in the happy song. The signal is nonstationary, but ARIMA models overrely on previous enjoyment values for prediction, suggesting that autoregressive models may not be appropriate for modeling enjoyment ratings.

TABLE IV
LASSO VALIDATION RMSE, AUDITORY FEATURES ONLY
Underline indicates the feature group with lowest RMSE for each song.
RMSE values $\times 10^2$.

Horizon 40	Happy	Sad Short	Sad Long
All	5.563	16.757	8.128
MFCCs	5.563	17.627	9.631
Spectral	<u>4.625</u>	18.263	9.456
Harmony	5.488	<u>16.571</u>	<u>7.698</u>
Dynamics	4.900	17.671	9.813
Rhythm	5.563	18.882	10.934
LPCs	5.211	18.862	10.567
Horizon 80	Happy	Sad Short	Sad Long
All	5.437	16.501	7.875
MFCCs	5.437	17.507	9.348
Spectral	<u>4.555</u>	18.139	9.196
Harmony	5.360	<u>16.333</u>	<u>7.496</u>
Dynamics	4.824	17.258	9.538
Rhythm	5.437	18.661	10.744
LPCs	5.191	18.643	10.292

most useful in prediction and eliminate irrelevant ones, we qualitatively split the features into different "feature groups" (Table I). In addition to the feature groups listed in Table I, we tested an "All" feature group which had access to all 74 auditory features. We ran Lasso and Ridge on each feature group with the task of predicting enjoyment ratings. Table IV shows the cross-validated RMSE results for each horizon, feature group, and song. The Lasso model performed equal or better than Ridge in every case, likely by setting the weights of irrelevant features to zero, so only the Lasso values are included in the table.

Table IV reveals that the Spectral feature group performed the best for the happy song, followed by Dynamics and LPCs. The two feature groups with lowest RMSE for the sad short song were Harmony and All, with third place taken by MFCCs for horizon 40 and Dynamics for horizon 80. Finally, for the sad long song, the top three feature groups in order were Harmony, All, and Spectral. For both time horizons tested, the Spectral feature group performed best on the happy song, while the Harmony feature group yielded the

best results for both sad songs. This suggests that features related to tonal quality were more important for predicting enjoyment of happy music, whereas interacting melodies and tones were more relevant in sad music. Happy songs may tend to use chord progressions that stay within the (major) key of the song, whereas sad songs may typically employ more complex harmonies with more dissonance. The happy, sad short, and sad long songs used in this study seemed to follow this paradigm. It is feasible that enjoyment in happy songs is therefore most correlated with changes in timbre (since the harmonies were mostly in the key of the song), whereas enjoyment in sad songs may correlate more with harmonic dissonance and resolution.

Another notable observation is that Lasso actually *improved* its predictions in the horizon 80 test compared to the horizon 40 test. This suggests that human affective reaction time to a change in audio features is closer to 2 seconds (80 samples) than 1 second (40 samples). In future research, we will test more horizon values in order to determine a more specific affective delay time.

D. Extracting Relevant Features

To glean a more detailed picture of which features within each feature group contributed to that group's success, we designed a scoring system to extract the relative importance of each auditory feature from the trained Lasso models. Each feature i's score s_i was calculated as the sum of each weight $A_{i,j}$ over every time step j in the attention window: $s_i = \sum_{j=1}^{a} A_{i,j}$. We then summed this score s_i over both horizons tested to determine the final score for that feature.

We calculated the relevance score two times: once with the absolute value of each weight considered (absolute score), and once with the signed value. The two scores differ if that feature had both positive and negative weights at different time steps within the attention window.

In Table V, we provide the scores of the top five features for each feature group. The features contributing to models predicting affective rating varied based on the musical properties of each song. Since each feature was normalized by the maximum value in itself, the relevance scores are comparable across and within feature groups. However, larger weights in a feature group did not necessarily lead to lower RMSE for that group, because the larger weights sometimes led to overfitting.

1) All: The feature group using all features was the second-best group behind Harmony in both sad songs, but it performed poorly on the happy song. In the happy song, the model assigned a weight of zero to every feature, relying solely on a bias term to make its predictions.

In the sad short song, the model positively correlated LPCs 6, 8, and 4 and negatively correlated skewness and kurtosis with enjoyment ratings. As seen in the Spectral group analysis below, skewness and kurtosis were important features for spectral prediction in the happy and sad long songs as well. Interestingly, despite having LPC features as the three most relevant features in the All group, the LPCs model for the sad short song failed to learn weights: possibly, the LPC values were only useful in combination with other features like skewness and kurtosis.

Finally, in the sad long song, the model using all features synthesized features of all four types examined in this study: timbre, harmony, dynamics, and rhythm. Each of the top-five-scoring features besides chroma Db were also among the top scoring features in their relevant feature group. However, the fact that the All feature group had a higher RMSE than the Harmony feature group in the sad long song shows that simply combining the top features from each other feature group was not enough to overcome overfitting from the other features.

2) Spectral: The Spectral feature group performed exceptionally well on the happy song, fairly well on the sad long song, and poorly on the sad short song. We found that kurtosis and skewness were the two most relevant features to predicting enjoyment in these songs. In the happy song, kurtosis was positively correlated while skewness was negatively correlated. In the sad long song, the reverse was true: kurtosis correlated negatively with enjoyment ratings while skewness had a positive correlation with enjoyment ratings. It is no surprise that kurtosis and skewness were correlated with human-reported enjoyment, as they are commonly used in music emotion recognition tasks [28]. A positive kurtosis indicates a larger proportion of outlying (very low or very high) frequencies, and positive skew indicates a preponderance of low frequencies, while negative skew shows a larger portion of high frequencies. In the happy song, positive kurtosis correlation and negative skewness correlation suggest that instruments with very high frequency overtones, such as the bell playing the melody, were associated with higher enjoyment ratings. Conversely, in the sad long song, negative kurtosis and positive skewness correlation indicate that middle and low frequencies corresponded to enjoyment.

3) Harmony: Harmony was the best feature group for predicting enjoyment in both sad songs. Most of the features selected were chroma features, which correspond to the strength of a given note in the chromagram. In the sad short song, chroma Db was most relevant and negatively correlated with enjoyment ratings. This note is the "minor 6th" in the key of the song, and is heavily featured in the song's bridge. The chromas of Eb and Gb, respectively the minor 7th and minor 2nd, were positively correlated. Unexpectedly, chroma B–the dissonant diminished 5th–was positively correlated with enjoyment in this song. In the sad long song, Gb, Ab, E, and A were the most relevant chroma notes. Gb, a dissonant note in the song, was positively correlated to enjoyment ratings. Key mode was negatively correlated, meaning that a minor modality was associated with enjoyment more than a major modality in this song. This indicates that the "sadder-sounding" the sad long song was, the higher the enjoyment ratings tended to be.

4) Dynamics: Dynamics performed well on the happy song, moderately well on the sad short song, and poorly on the sad long song. Compressibility was the highest-scoring feature for the happy and sad long songs, but RMS was the highest-scoring for the sad short song. Compressibility was positively

TABLE V
MOST RELEVANT FEATURES, AUDIO ONLY

	Happy			Sad Short			Sad Long		
	Feature	Score (Abs)	Score	Feature	Score (Abs)	Score	Feature	Score (Abs)	Score
All	MFCC 0	0.000	0.000	LPC 6	14.984	14.333	P. Clarity	2.260	-2.260
	MFCC 1	0.000	0.000	LPC 8	14.874	14.874	Compress.	0.953	0.953
	MFCC 2	0.000	0.000	LPC 4	8.328	8.328	Flux	0.664	0.664
	MFCC 3	0.000	0.000	Skewness	8.017	-0.423	Chroma Db	0.299	-0.299
	MFCC 4	0.000	0.000	Kurtosis	5.094	-0.953	Chroma E	0.189	0.189
Spectral	Kurtosis	3.781	3.781	Centroid	5.538	-5.538	Kurtosis	1.480	-1.480
	Skewness	3.198	-3.189	Kurtosis	2.326	2.014	Flux	1.146	1.146
	Centroid	1.079	-1.023	Skewness	2.223	-2.152	Skewness	1.071	1.019
	Spread	1.007	0.889	Brightness	1.471	1.471	Spread	0.388	-0.380
	Flux	0.493	0.493	Spread	1.324	1.249	Centroid	0.217	-0.041
Harmony	Chroma Gb	0.137	-0.137	Chroma Db	0.548	-0.548	Chroma Gb	0.577	0.577
	Key Strength	0.078	-0.078	Chroma Eb	0.325	0.322	Chroma Ab	0.550	0.550
	Chroma E	0.028	-0.028	Chroma Gb	0.304	0.302	Chroma E	0.460	0.460
	Chroma Ab	0.028	0.028	Chroma D	0.274	0.274	Key Mode	0.454	-0.454
	Chroma D	0.023	0.023	Chroma B	0.266	0.266	Chroma A	0.445	-0.445
Dynamics	Compress.	1.792	1.617	RMS	1.014	0.530	Compress.	1.388	1.388
	RMS	0.802	0.637	Compress.	0.473	-0.454	MFCC 0	0.267	0.267
	MFCC 0	0.802	-0.784	MFCC 0	0.191	0.003	RMS	0.094	-0.094
Rhythm	P. Clarity	0.000	0.000	P. Clarity	0.000	0.000	P. Clarity	6.903	-6.903
LPCs	LPC 6	7.294	-7.294	LPC 1	0.031	0.031	LPC 2	0.921	0.921
	LPC 8	4.984	-4.984	LPC 0	0.000	0.000	LPC 9	0.291	-0.291
	LPC 7	1.780	-1.780	LPC 2	0.000	0.000	LPC 10	0.198	0.198
	LPC 9	1.617	-1.617	LPC 3	0.000	0.000	LPC 7	0.138	0.138
	LPC 2	1.031	1.015	LPC 4	0.000	0.000	LPC 1	0.118	0.118
MFCCs	MFCC 0	0.000	0.000	MFCC 0	0.136	0.136	ΔMFCC 0	1.744	1.700
	MFCC 1	0.000	0.000	MFCC 10	0.113	0.113	ΔMFCC 1	0.621	-0.449
	MFCC 2	0.000	0.000	MFCC 9	0.075	0.075	MFCC 0	0.465	0.465
	MFCC 3	0.000	0.000	MFCC 12	0.065	-0.065	ΔMFCC 4	0.401	0.312
	MFCC 4	0.000	0.000	MFCC 1	0.060	0.060	ΔMFCC 12	0.389	0.259

correlated with enjoyment in the happy and sad long songs and negatively correlated in the sad short song. This would indicate that simpler musical passages in the happy and sad long song were rated highest for enjoyment, while the highlights of the sad songs were generally the more musically complex. The lack of consistently strong performance of the features across the three songs implies that the relative importance of each feature varies strongly by song.

5) Rhythm: The Rhythm feature group was not effective at predicting enjoyment in any song. The Rhythm models set all weights to zero for the happy and sad short songs, relying solely on a bias term for predictions. Pulse clarity was negatively correlated with enjoyment ratings in the sad long song, suggesting that complex rhythms stimulated enjoyment more than simple or clear beats.

6) LPCs: The LPCs feature group performed third-best in the happy song, but predicted poorly for both sad songs. In the happy song, LPCs 6 and 8 were the most relevant features for predicting enjoyment ratings and negatively correlated. LPCs 7 and 9 were also negatively correlated, while LPC 2 was positively correlated.

7) MFCCs: The MFCCs feature group was tied for worst-performing in the happy song and showed average utility for modeling enjoyment ratings in both sad songs. The sad short song model scored MFCC 0, 10, 9, 12, and 1 highly, while the sad long song model picked ΔMFCC 0, 1, 4, and 12, along with MFCC 0. The presence of MFCC 0 and ΔMFCC 0, positively correlated, indicates that increased energy across the entire frequency spectrum was associated with enjoyment. MFCCs 10, 9, and 12, which correspond to timbre, scored highly in the sad short song. In the sad long song, ΔMFCCs contributed most to models predicting enjoyment, suggesting that *change* in timbre–rather than just presence or absence of certain tonal qualities–correlated with enjoyment.

VI. CONCLUSION

Music is a universally appreciated form of art, but not all compositions are enjoyed in the same way. In this study, we investigated and quantified auditory features associated with listener-reported enjoyment ratings. We applied a set of multivariate time series prediction models to predict enjoyment using auditory features in sad and happy musical pieces as predictors. We compared performance of various prediction models and commented on what auditory features are useful for predicting affective responses. We hypothesized that attention models would perform best on these tasks, but found that this was not the case, likely due to overfitting and a lack of periodicity in the enjoyment signal. In the future, this work will be connected with subjective experience models with other time series, such as physiological and neural behavior. This work can be used to identify what aspects of songs affect us most profoundly on a fine timescale. It also offers more insight into how we respond to sad and happy music. Results of this study can be used to inform music emotion recognition, music information retrieval, and studies on subjective experience.

REFERENCES

[1] H. Kohut and S. Levarie, "On the enjoyment of listening to music," *The psychoanalytic quarterly*, vol. 19, no. 1, pp. 64–87, 1950.

[2] E. Brattico, P. Brattico, and T. Jacobsen, "The origins of the aesthetic enjoyment of musica review of the literature," *Musicae Scientiae*, vol. 13, no. 2_suppl, pp. 15–39, 2009.

[3] H. P. Weld, "An experimental study of musical enjoyment.," *The American Journal of Psychology*, 1912.

[4] M. R. Leek, M. R. Molis, L. R. Kubli, and J. B. Tufts, "Enjoyment of music by elderly hearing-impaired listeners," *Journal of the American Academy of Audiology*, vol. 19, no. 6, pp. 519–526, 2008.

[5] L. Migirov, J. Kronenberg, and Y. Henkin, "Self-reported listening habits and enjoyment of music among adult cochlear implant recipients," *Annals of Otology, Rhinology & Laryngology*, vol. 118, no. 5, pp. 350–355, 2009.

[6] K. Yoshii, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "An efficient hybrid music recommender system using an incrementally trainable probabilistic generative model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 435–447, 2008.

[7] C. Laurier, O. Lartillot, T. Eerola, and P. Toiviainen, "Exploring relationships between audio features and emotion in music," pp. 260–264, 01 2009.

[8] Y. E. Kim, E. M. Schmidt, R. Migneco, B. G. Morton, P. Richardson, J. Scott, J. A. Speck, and D. Turnbull, "Music emotion recognition: A state of the art review," in *Proc. ISMIR*, pp. 255–266, Citeseer, 2010.

[9] T. Greer, K. Singla, B. Ma, and S. Narayanan, "Learning shared vector representations of lyrics and chords in music," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3951–3955, IEEE, 2019.

[10] A. Kawakami, K. Furukawa, K. Katahira, and K. Okanoya, "Sad music induces pleasant emotion," *Frontiers in psychology*, vol. 4, p. 311, 2013.

[11] E. Brattico, V. Alluri, B. Bogert, T. Jacobsen, N. Vartiainen, S. K. Nieminen, and M. Tervaniemi, "A functional mri study of happy and sad emotions in music with and without lyrics," *Frontiers in psychology*, vol. 2, p. 308, 2011.

[12] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[13] G. G. Judge, R. C. Hill, W. Griffiths, H. Lutkepohl, and T. C. Lee, "Introduction to the theory and practice of econometrics.," 1982.

[14] S. Chen, X. Wang, and C. J. Harris, "Narx-based nonlinear system identification using orthogonal least squares basis hunting," *IEEE Transactions on Control Systems Technology*, vol. 16, no. 1, pp. 78–84, 2008.

[15] A. Bouchachia and S. Bouchachia, *Ensemble learning for time series prediction*. na, 2008.

[16] R. Frigola, Y. Chen, and C. E. Rasmussen, "Variational gaussian process state-space models," in *Advances in neural information processing systems*, pp. 3680–3688, 2014.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[18] S.-Y. Shih, F.-K. Sun, and H.-y. Lee, "Temporal pattern attention for multivariate time series forecasting," *arXiv preprint arXiv:1809.04206*, 2018.

[19] Y. Song, S. Dixon, and M. Pearce, "A survey of music recommendation systems and future perspectives," in *9th International Symposium on Computer Music Modeling and Retrieval*, vol. 4, 2012.

[20] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, C. Broussard, *et al.*, "Whats new in psychtoolbox-3," *Perception*, vol. 36, no. 14, p. 1, 2007.

[21] O. Lartillot, P. Toiviainen, and T. Eerola, "A matlab toolbox for music information retrieval," in *Data analysis, machine learning and applications*, pp. 261–268, Springer, 2008.

[22] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 5–18, 2006.

[23] F. Developers, "ffmpeg tool." http://ffmpeg.org/, 2019.

[24] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-feature modeling of pulse clarity: Design, validation and optimization.," in *ISMIR*, pp. 521–526, Citeseer, 2008.

[25] G. Lai, W.-C. Chang, Y. Yang, and H. Liu, "Modeling long-and short-term temporal patterns with deep neural networks," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 95–104, ACM, 2018.

[26] C.-H. Chuan and D. Herremans, "Modeling temporal tonal relations in polyphonic music through deep networks with a novel image-based representation," in *Thirty-second AAAI conference on artificial intelligence*, 2018.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[28] R. Panda, R. Malheiro, B. Rocha, A. Oliveira, and R. P. Paiva, "Multimodal music emotion recognition: A new dataset, methodology and comparative analysis," in *International Symposium on Computer Music Multidisciplinary Research*, 2013.