MUSICAL PROSODY-DRIVEN EMOTION CLASSIFICATION: INTERPRETING VOCALISTS PORTRAYAL OF EMOTIONS THROUGH MACHINE LEARNING

Nicholas FARRIS (nicholas.farris@gatech.edu)[1], **Brian MODEL** (bmodel@gatech.edu)[1], **Richard SAVERY** (rsavery3@gatech.edu)[1], and **Gil WEINBERG** (gilw@gatech.edu)[1]

[1]*Robotic Musicianship Lab*, **Georgia Institute of Technology**, Atlanta, GA USA

ABSTRACT

The task of classifying emotions within a musical track has received widespread attention within the Music Information Retrieval (MIR) community. Music emotion recognition has traditionally relied on the use of acoustic features, verbal features, and metadata-based filtering. The role of musical prosody remains under-explored despite several studies demonstrating a strong connection between prosody and emotion. In this study, we restrict the input of traditional machine learning algorithms to the features of musical prosody. Furthermore, our proposed approach builds upon the prior by classifying emotions under an expanded emotional taxonomy, using the Geneva Wheel of Emotion. We utilize a methodology for individual data collection from vocalists, and personal ground truth labeling by the artist themselves. We found that traditional machine learning algorithms when limited to the features of musical prosody (1) achieve high accuracies for a single singer, (2) maintain high accuracy when the dataset is expanded to multiple singers, and (3) achieve high accuracies when trained on a reduced subset of the total features.

1. INTRODUCTION

The work presented in this paper is situated in the intersection between research on emotion for robotics [1] and emotional classification research in Music Information Retrieval [2]. In particular, we focus on the under-explored domain of emotion-driven prosody for human-robot interaction [3]. Verbal prosody is concerned with elements of speech that are not individual phonetic segments but rather pertain to linguistic functions such as intonation, tone, stress, and rhythm. Similarly, musical prosody is defined as the performer's manipulation of music for certain expressive and coordinating functions [4]. It has been hypothesized that these expressive functions serve to communicate emotion [5].

In this paper, we explore the relationship between musical prosody and emotion through three research questions. First, are traditional machine learning algorithms able to accurately classify an individual's emotions when trained on only the features of musical prosody? Next, are these models able to generalize to a larger group of vocalists? Finally, which features of musical prosody contribute the most to the classification of emotion?

The paper is structured as follows, in Section 2, background and motivation are discussed. Section 3 describes the dataset collection, training and testing, the taxonomies used in classification, the feature extraction methodology and analysis of their relevance to emotion, feature aggregation, feature selection, and model generalization. Section 4 presents the experiments: Experiment 1 asks how well can traditional machine learning models classify emotion when limited to inputs of musical prosody, Experiment 2 explores our approach's ability to generalize to a larger population of singers, and Experiment 3 explores the individual contribution to accuracy of each feature via training on reduced subsets of the input vector. Section 5 provides discussion to these results, with particular attention paid to the relationships between emotions and potential future work. Finally, section 6 concludes the paper. A demo via python notebook with audio samples is available online. [1]

2. BACKGROUND

Emotion classification has been a major focus of research in recent years. Ekman created a discrete categorization that consists of fundamental basic emotions which are the root for more complex emotions [6]. Another classification model is the Circumplex model proposed by Posner et al which plots emotions on a continuous, two-dimensional scale of valence and arousal [7]. In this paper, we classify emotions using a model similar to the two-dimensional Circumplex model which is further described in section 3.1.

There has also been much work done in the field of analyzing emotion from text for tasks such as sentiment analysis. Research on classification of emotion in audio has taken many different approaches. Research into classifying emotions in knocking sounds has found that anger, happiness and sadness could be easily classified from audio alone [8]. There have been multimodal approaches which use audio in combination with another feature, namely visual facial features [9] [10] or text lyrics [11]. Furthermore, researchers have performed emotional classification from audio in the context of music by analyzing which musical features best convey emotions [12]. Panda et al. have found a relationship between melodic and

[1] https://github.com/brianmodel/EmotionClassification

dynamic features to a number of specific emotions [13]. Such features that were used to classify emotion in music, however, cannot be easily generalized to other domains. Prosody has been found by linguists to communicate emotion across various cultures, with patterns of pitch and loudness over time representing different emotions [14], and has shown the potential to improve human-robot interaction [15–17]. Our approach aims to bridge this gap by analyzing these prosodic features which are fundamental to everyday speech and explore how they can be used to classify emotional driven prosody.

Koo et al. have done work in speech emotion recognition using a combination of MFCC and prosodic features with a GRU model on the IEMOCAP dataset [18]. We expand upon their work by performing an in-depth analysis of 11 different audio features and their effect on classifying emotion. We also classify emotion beyond spoken language by analyzing prosodic features which better generalize to how humans convey emotion using the new dataset collected, as described in section 3.2.

3. METHODOLOGY

3.1 Taxonomy

One of the main challenges in emotional classification is the derivation of a taxonomy that accurately reflects the problem domain. The two common approaches to address this challenge are 1. Discrete emotional categorization; and 2. Continuous quantitative metrics of Valence and Arousal (sometimes called Control). We use both approaches with a categorical, as opposed to regression, approach to the latter.

Our models classify emotion under two taxonomies: first we categorize each data point as belonging to one of the twenty emotions located around the Geneva Wheel of Emotion. Then we categorize each data point as belonging to one of the quadrants depicted by the intersection of valance and control by assigning each emotion from the Geneva Wheel of Emotion to its respective quadrant. We abbreviate each of these quadrants as follows: "High Control Negative Valance": "HCN", "High Control Positive Valance": "HCP", "Low Control Negative Valance": "LCN", and "Low Control Positive Valance": "LCP". See Table 1 and Figure 1 for a visualization of the domain's taxonomy.

HCN	HCP	LCN	LCP
Anger	Amusement	Disappointment	Admiration
Contempt	Interest	Fear	Compassion
Disgust	Joy	Guilt	Contentment
Hate	Pleasure	Sadness	Love
Regret	Pride	Shame	Relief

Table 1. Selected emotional taxonomy for training

3.2 Data Collection

Due to a lack of data labeled with the appropriate taxonomy, we decided to collect and annotate a new dataset. To achieve this goal, we asked professional singers to consciously sing each emotion. To generate our dataset, three professional singers were tasked to improvise as many phrases as possible for each emotion in the Geneva Wheel of Emotion. The singers were instructed to sing each phrase between 1 and 20 seconds, and to spend approximately 15 minutes on each emotion, resulting in 4 to 6 hours of recordings per singer annotated with ground-truth labels.

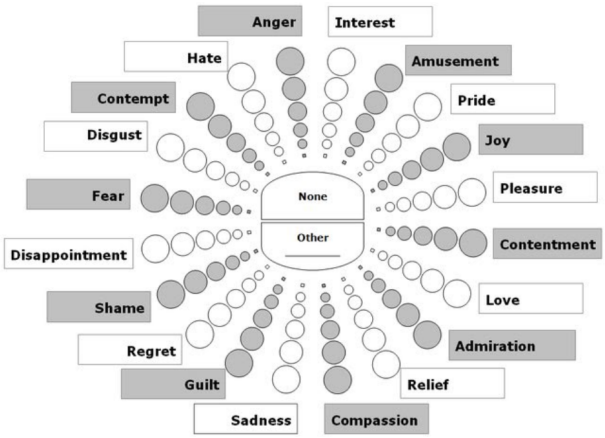

Figure 1. The Geneva Wheel of Emotion

Additionally, the singers were given the following instructions during their recording session:

1. Do not attempt to control for different intensities for each emotion
2. Sing anything for each phrase that you believe matches the emotion except use words.
3. After recording, mark any phrase that you believe did not capture the intended emotion and it will be deleted

3.3 Feature Extraction

In the following section, we define the features selected for extraction from our dataset prior to model training. Furthermore, we discuss each feature's relevance to emotional classification through an analysis of prior works.

3.3.1 Zero Crossing Rate

Zero Crossing Rate, the rate of sign-changes across a signal, is key in classifying percussive sounds. Unvoiced regions of audio are known to have higher Zero Crossing Rates [19]. One study analyzed ZCR for Anger, Fear, Neutral, and Happy signals and noted that higher peaks were found for Happy and Anger emotions [20].

3.3.2 Energy

Energy, the area under the squared magnitude of the considered signal, relates to the amount of spectral information in a signal [21] and previous studies have found energy is essential in distinguishing stressed and neutral speech [22].

3.3.3 *Entropy of Energy*

Entropy of Energy, the average level of "information" or "uncertainty" inherent within a signal's energy, has been shown in one study to have similar values for disgust and boredom [23]. To accurately measure the entropy of the different emotions, we must make sure we are not including parts of the signal where the individual is not speaking.

3.3.4 *Spectral Centroid*

Spectral Centroid, the power spectrum's center of mass, perceptually has a connection with a sound's brightness. It follows, that this parameter serves as an indicator of musical timbre [24]. Previous studies have shown spectral centroid is a significant component in music emotion [25].

3.3.5 *Spectral Spread*

Spectral Spread, the second central moment of the power spectrum, has shown to help the listener to differentiate noise-like and tone-like portions of a signal [26].

3.3.6 *Spectral Entropy*

Spectral Entropy, the entropy of the power spectrum, when used with MFCC features has shown an improvement in speech recognition accuracy [27]. Another study found spectral entropy to have the highest correlation to emotional valence of all features tested [28].

3.3.7 *Spectral Flux*

Spectral Flux, a measure of the rate of change of the power spectrum calculated as the Euclidean distance between sequential frames, relates to how fast the pitch changes in time and has been shown to be dominant in cross-domain emotion recognition from speech and sound and from sound and music [29].

3.3.8 *Spectral Rolloff*

Spectral Rolloff, the frequency under which some percentage of the total energy of the spectrum is contained, helps differentiate between harmonic content, characterized below the roll-off, and noisy sounds, characterized above the roll-off. Spectral rolloff has been shown to be one of the most important prosodic features in classifying emotion [28].

3.3.9 *MFCCs*

Mel-Frequency Cepstral Coefficients (MFCCs), a representation of the short-term power spectrum based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency, are used in speech recognition with their ability to represent the speech amplitude spectrum in a compact form [30]. Many studies have linked the importance of MFCC analysis to emotion recognition [20] [31] [32] .

3.3.10 *Chroma Vector and Deviation*

Chroma Vector, an approximation of the pitch class profiles present within a given frame and often used as the twelve tones, allows for the capture of harmonic and melodic characteristics while remaining robust toward changes in timbre and instrumentation. Previous studies have shown increases in emotional classification accuracy with chroma vector and its standard deviation [33, 34].

Parameter	Value
Mid-term Window Step	1.0 seconds
Mid-term Window Size	1.0 seconds
Short-term Window Step	0.05 seconds
Short-term Window Size	0.05 seconds

Table 2. Feature Aggregation Parameters

3.4 Feature Aggregation

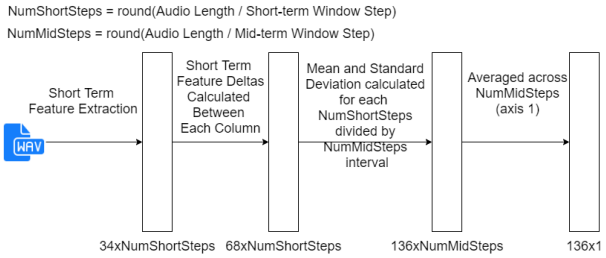

Figure 2. Model of Feature Aggregation

In this section, we define the aggregation pipeline from feature extraction to feature vector for each audio file. Figure 2 provides a visual modeling of our feature aggregation pipeline. Table 2 delineates the feature aggregation hyper parameters used in this study.

3.4.1 *Short-term Aggregation*

The short-term aggregation of a 5-second clip, using a Short-term Window Step of .05 seconds and a Short-term Window Size of .05 seconds is defined as follows: Each of the 34 features discussed above are extracted for every 50ms, resulting in 100 feature vectors of size 34x1, represented as a 34x100 matrix. Next, the deltas between each time step are calculated according to the equation $delta = feature\_vector - feature\_vector\_prev$. The first time stamp has all deltas set to 0. Each delta vector is concatenated onto its respective feature vector resulting in a size of 68x1, represented as a 68x100 matrix for the entire 5 second audio clip.

3.4.2 *Mid-term Aggregation*

Next, mid-term aggregation occurs with a Mid-term Window Size of 1.0 seconds and Mid-term Window Step of 1.0 seconds. The 68x100 matrix of Short-term features is split according to the ratio between the Mid-term and Short-term window size and step, resulting in 5 matrices of size 68x20. For each matrix, we calculate and flatten the mean and standard deviation for each row, resulting in 5 136x1 mid-term feature vectors, represented as a 136x5 matrix. Finally, we take the mean across the first axis resulting in a 136x1 feature vector representing our 5 second audio clip.

3.5 Classification

Prior work focused on musical classification has primarily found success in the implementation of k-nearest neighbor (K-NN) and support vector machines (SVM), finding the highest accuracies using SVMs [35]. In exploration of the relationship between musical prosody and emotion, we will implement a variety of machine learning models, namely we will train and evaluate KNNs, Linear SVMs, Random Forests, Extra Trees, Gradient Boosting, and Feed Forward Neural Networks (FFNN). FFNNs are used in experiment 3 only.

Experiment 1: we explore the base line accuracies, F-scores, and confusion matrices achieved by training each model with identical training, validation, and testing data from a single singer.

Experiment 2: we explore our model architecture's ability to generalize by expanding the dataset to include all 3 singers from data collection.

Experiment 3: we explore model performance on a reduced subset of the training feature, utilizing additive feature selection to compile a ranking of features.

4. RESULTS

4.1 Experiment 1

In experiment 1, we analyze the baseline accuracies, F-scores, and confusion matrices achieve by training KNNs, linear SVMs, Random Forests, Extra Trees, Gradient Boosting models on a single singer utilizing only the prosodic features outlined in the previous section. All models were trained with features extracted according to the parameters outlined in Table 2. Additionally, each model is optimized with respect to its associated hyper parameter. We optimize KNN for the number nearest neighbors, SVM for the soft margin, random forest for number of trees, gradient boosting for the number of boosting stages, and extra trees for the number of trees. [2]

Table 3 provides the best accuracy, F1-score, and selected hyper-parameter for each of our models trained on a Big 4 taxonomy for a single singer. All models perform better than twice the accuracy of random guessing, with the linear SVM and Gradient Boosting models achieving the highest accuracies. Further analysis of the confusion matrix of the Gradient Boosting model, shown in Figure 4, provides information about the classes that are most often confused for one another. The model struggles in distinguishing between Low Control Positive Valance and High Control Positive Valance. This is to say the model can tell that an individual is in a positive mood, but has difficulties distinguishing the Control or Arousal of the emotion.

Next, we examine classification under a single emotion taxonomy for a single singer. Table 4 shows the best accuracy, F1-score, and selected hyper-parameter for each of our models. Each model significantly outperforms random guessing. Even the worst model, the KNN, performs 6.5 times better than random chance (20 possible categories = 5% chance random guessing). Our best model, the linear SVM, performs approximately 10 times better than random guessing with an accuracy of 49.1%. The confusion matrix for the single emotion taxonomy has been included in Figure 3. Analysis of this confusion matrix yields a few observations: Disgust is rarely confused with other emotions, having the highest individual accuracy of 81.4%. Fear and Guilt are the two most common pair of emotions to be confused for one another. Pleasure is the most difficult emotion for the model to classify correctly, having the lowest individual accuracy of 18.6%.

Finally, our models perform extremely well when tasked with categorizing between two emotions, achieving accuracies as high as 98.9% with a f1 of 98.9 in the distinction between Love and Disgust using a SVM. This reinforces the intuition that by reducing the number of emotional categories we can achieve higher accuracies for identification.

Model	Accuracy	F1	Hyperparam
KNN	56.1	56.2	C=11
SVM	66.5	65.3	C=1.0
Extra Trees	64.6	64.3	C=100
Gradient Boosting	67.0	66.7	C=500
Random Forest	63.5	63.2	C=200

Table 3. Big 4 Taxonomy, 1 Singer Classification Results

Model	Accuracy	F1	Hyperparam
KNN	33.8	32.1	C=15
SVM	49.1	48.1	C=5.0
Extra Trees	44.3	42.8	C=500
Gradient Boosting	47.2	46.6	C=200
Random Forest	43.8	42.3	C=200

Table 4. Single Taxonomy, 1 Singer Classification Results

4.2 Experiment 2

Within machine learning, model generalization poses many challenges as models tend to memorize data and perform worse when exposed to new datasets. In experiment 2, we generalized our model by training on 3 different singers as opposed to training on one singer. Tables 5 and 6 compare the accuracies achieved by the various model architectures for 3 singers vs 1 singer.

With the exception of linear SVM, all model architectures maintain similar accuracies when trained on the 3 singer datasets. This maintenance of accuracy demonstrates the ability for traditional machine learning models to generalize well to a larger population when trained on only the features of musical prosody. We are unsure of why linear SVMs perform worse during generalization as compared to other models, seeing a drop of 6% in Big 4 taxonomy and a drop of 13% in single emotion taxonomy. This drop could potentially be a limitation in our methodology of only applying a linear kernel to SVM training, as perhaps an RBF or polynomial kernel would be better able to generalize to a larger population.

The results of this experiment are encouraging to the development of a general model of emotional classification based on musical prosody as accuracy is maintained when

[2] https://scikit-learn.org/

												Prediction											
				HCN					HCP					LCN					LCP				
			Ang	Contem	Disg	Hate	Reg	Amu	Int	Joy	Ple	Pri	Disa	Fear	Gui	Sad	Sha	Adm	Com	Conten	Love	Rel	Total	Acc
		Ang	2.94	0.41	0.16	0.41	0	0.08	0	0.08	0.08	0.08	0	0.16	0.08	0	0.16	0	0	0.08	0.08	0.08	4.88	60.2%
		Contem	0.65	1.88	0.16	0.33	0.41	0.08	0.08	0.08	0.16	0	0.16	0.49	0.33	0	0.08	0.33	0	0	0	0.08	5.3	35.5%
	HCN	Disg	0.24	0.08	5.31	0.08	0	0.33	0	0	0	0	0.16	0.16	0	0	0	0.08	0	0	0	0.08	6.52	81.4%
		Hate	0.57	0.33	0.24	3.35	0	0.16	0.08	0.33	0	0	0	0.16	0	0.16	0	0	0	0	0.24	0.08	5.7	58.8%
		Reg	0.24	0.08	0.08	0.16	1.47	0	0.08	0	0.41	0.16	0.16	0.08	0.49	0.08	0.16	0.08	0.41	0.16	0	0.16	4.46	33.0%
		Amu	0	0	0.16	0.08	0	3.18	0.33	0.33	0	0.08	0	0.08	0	0	0	0.16	0	0	0	0.08	4.48	71.0%
		Int	0	0.16	0	0	0.08	0.82	2.12	0	0.24	0.16	0.16	0	0	0.16	0.08	0.24	0.24	0.65	0.08	0.08	5.27	40.2%
	HCP	Joy	0.08	0.16	0	0.33	0.08	0.57	0.08	2.53	0.41	0.08	0.08	0.08	0	0	0	0.49	0.16	0	0.33	0.24	5.7	44.4%
		Ple	0	0.16	0.08	0	0.24	0	0.41	0	1.06	0.16	0.24	0	0.41	0.08	0.24	0.41	0.73	0.49	0.82	0.16	5.69	18.6%
		Pri	0.16	0.08	0	0	0	0	0	0.08	0.08	2.37	0.08	0.08	0.08	0	0	0.24	0.16	0.16	0.33	0.16	4.06	58.4%
Truth		Disa	0.16	0	0.16	0	0	0	0.08	0.08	0.08	0.08	2.45	0.08	0.33	0	0.16	0.08	0.16	0	0	0.16	4.06	60.3%
		Fear	0.24	0.33	0.08	0.16	0.08	0	0	0	0.24	0.08	0.08	2.53	0.49	0.08	0.16	0.08	0.08	0	0.08	0.08	4.87	52.0%
	LCN	Gui	0.08	0.08	0	0	0.24	0.08	0	0	0.24	0.24	0.08	1.14	2.45	0.08	0.16	0.08	0.08	0.24	0	0	5.27	46.5%
		Sad	0.33	0	0.16	0.24	0	0	0.24	0	0	0	0.24	0	0	1.71	0.16	0	0.08	0.16	0.08	0.24	3.64	47.0%
		Sha	0.08	0.16	0.08	0.16	0.33	0	0	0	0.08	0.33	0.24	0.65	0.33	0.08	1.55	0.08	0.16	0.33	0.16	0.08	4.88	31.8%
		Adm	0.08	0.08	0	0.16	0	0.41	0.24	0.24	0.16	0	0.24	0	0	0	0	2.29	0	0	0.08	0.08	4.06	56.4%
		Com	0	0.08	0	0	0	0	0.16	0	0	0.41	0.49	0.16	0.16	0	0.16	0.08	2.2	0.41	0.16	0	4.47	49.2%
	LCP	Conten	0.08	0.08	0	0	0.16	0.16	0.08	0	0.33	0.08	0.08	0.08	0.16	0.24	0.08	0.49	0.98	1.71	0.08	0	4.87	35.1%
		Love	0.08	0.08	0	0.08	0	0.16	0.24	0.16	0.33	0.33	0	0.16	0.24	0.08	0	0.16	0.82	0.41	1.96	0	5.29	37.1%
		Rel	0.08	0.16	0.33	0.24	0	0	0.16	0.16	0.08	0	0.16	0	0.08	0.16	0	0.16	0.16	0.08	0.08	4	6.09	65.7%
		Total	6.09	4.39	7	5.78	3.09	6.03	4.38	4.07	3.98	4.64	5.1	6.09	5.63	2.91	3.15	5.53	6.42	4.88	4.56	5.84	100	49.1%

Figure 3. SVM, Individual Taxonomy, 1 Singers Confusion Matrix

				Prediction			
		HCN	HCP	LCN	LCP	Total	Acc
	HCN	20.82	2.38	2.46	1.39	27.05	77.0%
	HCP	2.21	15.16	2.38	5.66	25.41	59.7%
Truth	LCN	2.87	2.13	15.66	2.3	22.96	68.2%
	LCP	1.39	5.16	2.62	15.41	24.58	62.7%
	Total	27.29	24.83	23.12	24.76	100	67.0%

Figure 4. Gradient Boosting, Big 4 Taxonomy, 1 Singer Confusion Matrix

Model	1-S Accuracy	3-S Accuracy
KNN	56.1	57.9
SVM	66.5	60.6
Extra Trees	64.6	63.5
Gradient Boosting	67.0	68.8
Random Forest	63.5	65.1

Table 5. Big 4 Taxonomy, 1 Singer vs 3 Singer Accuracy

the dataset is expanded to a larger portion of the overall population.

4.3 Experiment 3

Experiment 3 analyzes model performance on a reduced subset of the feature vector for our single emotion taxonomy. Our implementation of Feature Selection follows an additive approach. We start with an empty permanent feature set and each feature is trained on its own. The feature with the highest f1 score is selected and added to our permanent feature set. This process is repeated until all features have been added to the permanent feature set. Finally, we plot the f1 score vs features used in model training.

Model	1-S Accuracy	3-S Accuracy
KNN	33.8	32.5
SVM	49.1	36.9
Extra Trees	44.3	42.7
Gradient Boosting	47.2	43.8
Random Forest	43.8	43.8

Table 6. Single Emotion Taxonomy, 1 Singer vs 3 Singer Accuracy

For 136 features, an additive feature selection training loop requires the training and f1 validation of 9316 models. Our initial training and validation was based on implementations using the python library sklearn. Unfortunately, sklearn does not provide native GPU training support and thus performing an additive feature selection using sklearn is not feasible with respect to training time. Our solution is to continue to use the feature selection and aggregation outlined above, and to replace the sklearn models with a Tensorflow feed forward neural net. All of these models look for statistical correlations between our features and the emotional classification. Thus the particular model should have minimal affect on the analysis of feature importance performed by additive feature selection. Training was done sequentially on a RTX 3090 using CUDA v11 and took just under 24 hours to train and validate all 9316 models.

Our feed forward neural net contained the input layer, two dense layers of 136 nodes with relu activation functions, and a dense 20 node output layer. We trained using a Sparse Categorical Cross entropy loss function optimized using an Adam optimizer with 5 epochs per model.

Figure 7 shows the F1 score achieved vs the Feature included in the model pipeline. All feature on and to the right of any point in the x axis are included in training. An F1 of 45 is achieved within the first 25 features. Furthermore, the addition of the remaining 111 features only increases our F1 score to 52. This graph emphasizes the importance of spectral roll-off and MFCC 7 in the classification of emotion, as aggregations of these two features allow for an F1 score just below 20 with 4 total features.

5. DISCUSSION

5.1 Analysis

We demonstrate that prosodic features can be used to classify human emotions, achieving high accuracies on classifying emotions for a single singer dataset as seen in tables 3 and 4. Furthermore, we obtained encouraging results

			Prediction																				
			HCN					HCP				LCN					LCP						
			Ang	Contem	Disg	Hate	Reg	Amu	Int	Joy	Ple	Pri	Disa	Fear	Gui	Sad	Sha	Adm	Com	Conten	Love	Rel	Total	Acc
		Ang	1.74	0.34	0.59	0.5	0.07	0.32	0.14	0.07	0.02	0.34	0.09	0.02	0.09	0.02	0.07	0.07	0.05	0	0.02	0.09	4.65	37.4%
		Contem	0.27	2.06	0.25	0.41	0.25	0.09	0.18	0.25	0.18	0.02	0.2	0.2	0.11	0.09	0.25	0.2	0.07	0.02	0.05	0.27	5.42	38.0%
	HCN	Disg	0.45	0.5	2.58	0.68	0.02	0.45	0.07	0.07	0.02	0.16	0.02	0.05	0.07	0.02	0.05	0.11	0	0.02	0	0.2	5.54	46.6%
		Hate	0.38	0.38	0.57	2.47	0.05	0.11	0.07	0.29	0.16	0.27	0.09	0.05	0.02	0.05	0.16	0.07	0.05	0.09	0.14	0.2	5.67	43.6%
		Reg	0.02	0.34	0.02	0.11	2.26	0	0.07	0	0.23	0.07	0.07	0.18	0.18	0.14	0.11	0.11	0.34	0.05	0.23	0.34	4.87	46.4%
		Amu	0.09	0.02	0.34	0.11	0	2.38	0.29	0.38	0.2	0.45	0	0.02	0	0	0	0.18	0.05	0.25	0.05	0.16	4.97	47.9%
		Int	0	0.14	0.11	0.05	0.07	0.2	2.56	0.34	0.05	0.05	0.09	0.29	0.14	0.25	0.25	0.02	0.07	0.14	0.07	0.11	5	51.2%
	HCP	Joy	0.07	0.14	0.07	0.16	0.09	0.45	0.11	2.87	0.09	0.07	0.05	0.23	0.07	0.18	0.07	0.14	0.05	0	0	0.09	5	57.4%
		Ple	0.05	0.16	0.07	0.16	0.07	0.16	0.02	0.07	1.92	0.45	0.11	0.07	0.02	0	0.11	0.29	0.25	0.75	0.27	0.2	5.2	36.9%
		Pri	0.09	0.09	0.05	0.16	0.07	0.34	0.14	0.27	0.38	2.35	0	0.14	0	0.02	0.18	0.23	0.07	0.34	0.16	0.36	5.44	43.2%
Truth		Disa	0.14	0.11	0.07	0.09	0.11	0	0.11	0	0.25	0.09	1.83	0.25	0.25	0.2	0.25	0.05	0.2	0.07	0.07	0.16	4.3	42.6%
		Fear	0.09	0.27	0.05	0.09	0.07	0.05	0.25	0.11	0.23	0.05	0.16	2.85	0.27	0.18	0.29	0.05	0.11	0	0.11	0.05	5.33	53.5%
	LCN	Gui	0.07	0.11	0.07	0.02	0.09	0	0.11	0.07	0.14	0.02	0.32	0.54	1.67	0.07	0.23	0.09	0.11	0.02	0.14	0.07	3.96	42.2%
		Sad	0	0.05	0.02	0.09	0.09	0	0.2	0.09	0.11	0.02	0.07	0.48	0.14	2.53	0.29	0.02	0.16	0.05	0.11	0.11	4.63	54.6%
		Sha	0.14	0.16	0.14	0.14	0.14	0	0.16	0.07	0.07	0.07	0.25	0.45	0.25	0.34	1.97	0.05	0.07	0.16	0.14	0.14	4.91	40.1%
		Adm	0.07	0.11	0.07	0.23	0.27	0.18	0.07	0.18	0.36	0.25	0.09	0.07	0.02	0.18	0.09	1.24	0.27	0.11	0.11	0.32	4.29	28.9%
		Com	0	0.2	0.09	0.14	0.2	0	0.02	0	0.27	0.14	0.05	0.07	0.07	0.07	0.11	0.14	1.76	0.34	0.5	0.36	4.53	38.9%
	LCP	Conten	0	0.09	0.02	0.11	0.11	0.11	0.07	0.07	0.61	0.36	0.14	0	0.05	0.02	0.14	0.16	0.32	1.74	0.34	0.41	4.87	35.7%
		Love	0	0.07	0	0.14	0.2	0.05	0.07	0.11	0.57	0.16	0.14	0.11	0.05	0.05	0.02	0.41	0.48	0.38	1.61	0.5	5.12	31.4%
		Rel	0.11	0.07	0.29	0.2	0.32	0.16	0.14	0.07	0.16	0.23	0.09	0	0	0.07	0	0.29	0.18	0.23	0.5	3.35	6.46	51.9%
		Total	3.78	5.41	5.47	6.06	4.55	5.05	4.85	5.38	6.02	5.62	3.86	6.07	3.47	4.48	4.64	3.92	4.66	4.76	4.62	7.49	100	43.8%

Figure 5. Gradient Boosting, Individual Taxonomy, 3 Singers Confusion Matrix

		Prediction					
		HCN	HCP	LCN	LCP	Total	Acc
	HCN	17.81	2.73	2.51	3.05	26.1	68.2%
	HCP	2.42	16.45	2.44	4.23	25.54	64.4%
Truth	LCN	2.08	1.76	17.4	1.81	23.05	75.5%
	LCP	2.46	3.98	1.74	17.13	25.31	67.7%
	Total	24.77	24.92	24.09	26.22	100	68.8%

Figure 6. Gradient Boosting, Big 4 Taxonomy, 3 Singer Confusion Matrix

regarding the model's generalization between singers as demonstrated by tables 5 and 6. However, given our limited dataset, more research is needed to study how the models generalize for additional singers with different voices.

Our feature selection aligns with prior research indicating that energy and MFCC were the most useful features for classifying emotion [9]. However, we have been able to show that the results holds true not just for phonological speech, but in the more specific domain of musical prosody.

5.2 Relationships between Emotions

The classification results give us new insights into the uniqueness and relationships between emotions. Looking at the individual classification data between all the singers in Figure 3, we can see how the model was best able to classify fear, joy and relief. This is in contrast to emotions such pleasure or admiration which showed the lowest classification accuracy. These results demonstrate the manner in which different humans convey emotions, and what emotions are similarly expressed by different individuals. When conveying relief, all three singers expressed a diminuendo and exhale. Similarly, when conveying fear all three singers expressed a crescendo and more accented tones. On the other hand, there was a high level of variation when conveying pleasure, with many different tone ranges, mouth shapes, etc. being present in the data.

Furthermore, from the confusion matrix in Figure 5, we can see that the emotion pairs of Hate and Disgust as well as Pleasure and Contentment are the most common emotions to be misclassified as one another. We suggest that this is due to these emotions representing similar meanings, thus they would be conveyed using similar features. For instance, Hate and Disgust both tend to consist of lower tones while Pleasure and Contentment have higher tones.

5.3 Future Work

One of the major challenges we faced was the limited amount of data that was collected. We plan on expanding this dataset to a larger variety of singers and other instrumentalists so that we can better understand how the models can generalize to different sounds. Additional future work includes developing a more sophisticated deep-learning based model on the raw audio data for classifying emotion using the expanded dataset we will collect. This will allow the model to make predictions beyond what could be possible using the features we chose in our feature selection. It would open up the potential to achieve much higher accuracy and better model generalization.

6. CONCLUSIONS

Our novel dataset using an expanded emotion taxonomy provides opportunity for the development of a more articulate understanding of emotions. Previous attempts to correlate emotion to audio or music are based on fewer emotions, and often rely on lyrics or song metadata for classification. Our algorithms demonstrate a high level of accuracy on a 20 category taxonomy for emotions, utilizing only prosodic features. By restricting the type of input data to prosodic features and expanding the number of classified emotions, our models can be used for a wide range of research challenges within the domain of emotional classification. Furthermore, we have demonstrated that our approach is able to generalize to a larger subset of the overall population. Finally, the restriction of our feature vector via additive feature selection demonstrates the ability for prosodic features to achieve a high-level accuracy for emotional classification for a relatively small number of features.

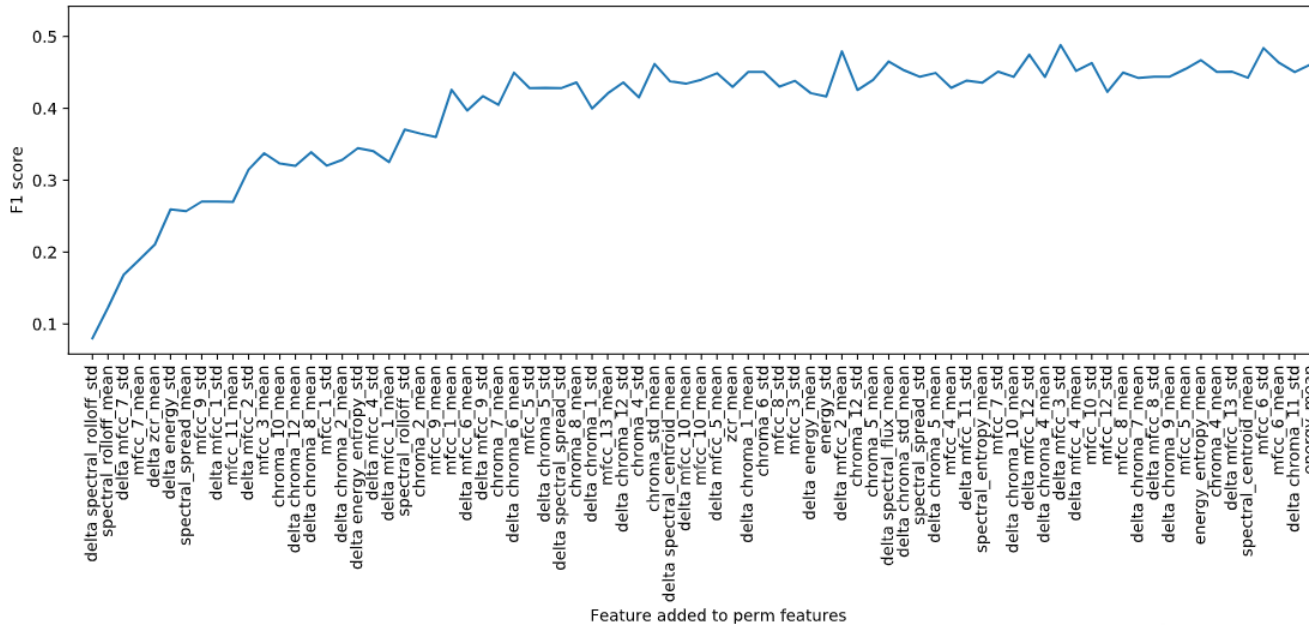

Figure 7. F1 score vs Features included in model pipeline

7. REFERENCES

[1] R. Savery and G. Weinberg, "A survey of robotics and emotion: Classifications and models of emotional interaction," 07 2020.

[2] P. Cano, M. Koppenberger, and N. Wack, "Content-based music audio recommendation," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 211–212.

[3] R. Savery, R. Rose, and G. Weinberg, "Establishing human-robot trust through music-driven robotic emotion prosody and gesture," in *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2019, pp. 1–7.

[4] C. Palmer and S. Hutchins, "What is musical prosody?" *Psychology of Learning and Motivation*, vol. 46, 12 2006.

[5] P. N. Juslin and J. A. Sloboda, *Music and emotion: Theory and research.* Oxford University Press, 2001.

[6] P. Ekman, "Basic emotions," *Handbook of cognition and emotion*, vol. 98, no. 45-60, p. 16, 1999.

[7] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 3, pp. 715–734, 2005.

[8] M. Houel, A. Arun, A. Berg, A. Iop, A. Barahona-Rios, and S. Pauletto, "Perception of emotions in knocking sounds : an evaluation study," in *∴*, 2020, qC 20200722. [Online]. Available: https://smc2020torino.it/adminupload/file/SMCCIM_2020_paper_95.pdf

[9] S. ul haq, P. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," *The proceedings of international conference on auditory-visual speech processing*, pp. 185–190, 09 2008.

[10] L. C. De Silva and Pei Chi Ng, "Bimodal emotion recognition," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, 2000, pp. 332–335.

[11] A. Jamdar, J. Abraham, K. Khanna, and R. Dubey, "Emotion analysis of songs based on lyrical and audio features," *CoRR*, vol. abs/1506.05012, 2015. [Online]. Available: http://arxiv.org/abs/1506.05012

[12] Y. Song, S. Dixon, and M. Pearce, "Evaluation of musical features for emotion classification," 10 2012.

[13] R. Panda, R. M. Malheiro, and R. P. Paiva, "Audio features for music emotion recognition: a survey," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[14] R. W. Frick, "Communicating emotion: The role of prosodic features." *Psychological Bulletin*, vol. 97, no. 3, p. 412–429, 1985.

[15] R. Savery, R. Rose, and G. Weinberg, "Finding shimi's voice: fostering human-robot communication with music and a nvidia jetson tx2," in *Proceedings of the 17th Linux Audio Conference*, 2019, p. 5.

[16] R. Savery, L. Zahray, and G. Weinberg, "Emotional musical prosody for the enhancement of trust in robotic arm communication," in *Trust, Acceptance and Social Cues in Human-Robot Interaction: 29th IEEE International Conference on Robot & Human Interactive Communication*, 2020.

[17] ——, "Before, between, and after: Enriching robot communication surrounding collaborative creative activities," *Frontiers in Robotics and AI*, vol. 8, p. 116, 2021. [Online]. Available: https://www.frontiersin.org/article/10.3389/frobt.2021.662355

[18] H. Koo, S. Jeong, S. Yoon, and W. Kim, "Development of speech emotion recognition algorithm using mfcc and prosody," in *2020 International Conference on Electronics, Information, and Communication (ICEIC)*, 2020, pp. 1–4.

[19] Wai Pang Ng, J. M. H. Elmirghani, R. A. Cryan, Yoong Choon Chang, and S. Broom, "Divergence detection in a speech-excited in-service non-intrusive measurement device," in *2000 IEEE International Conference on Communications. ICC 2000. Global Convergence Through Communications. Conference Record*, vol. 2, 2000, pp. 944–948 vol.2.

[20] E. Ramdinmawii, A. Mohanta, and V. K. Mittal, "Emotion recognition from speech signal," in *TENCON 2017 - 2017 IEEE Region 10 Conference*, 2017, pp. 1562–1567.

[21] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project," *CUIDADO Ist Project Report*, vol. 54, no. 0, pp. 1–25, 2004.

[22] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[23] S. Lalitha, A. Mudupu, B. V. Nandyala, and R. Munagala, "Speech emotion recognition using dwt," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*. IEEE, 2015, pp. 1–4.

[24] R. Kendall and E. Carterette, "Difference thresholds for timbre related to spectral centroid," in *Proceedings of the 4th International Conference on Music Perception and Cognition, Montreal, Canada*, 1996, pp. 91–95.

[25] B. Wu, A. Horner, and C. Lee, "Musical timbre and emotion: The identification of salient timbral features in sustained musical instrument tones equalized in attack time and spectral centroid," in *ICMC*, 2014.

[26] U. Jain, K. Nathani, N. Ruban, A. N. J. Raj, Z. Zhuang, and V. G. Mahesh, "Cubic svm classifier based feature extraction and emotion detection from speech signals," in *2018 International Conference on Sensor Networks and Signal Processing (SNSP)*. IEEE, 2018, pp. 386–391.

[27] A. Toh, R. Togneri, and S. Nordholm, "Spectral entropy as speech features for speech recognition," *Proceedings of PEECS*, 01 2005.

[28] S. Cunningham, J. Weinel, and R. Picking, "High-level analysis of audio features for identifying emotional valence in human singing," in *Proceedings of the Audio Mostly 2018 on Sound in Immersion and Emotion*, ser. AM'18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3243274.3243313

[29] F. Weninger, F. Eyben, B. W. Schuller, M. Mortillaro, and K. R. Scherer, "On the acoustics of emotion in audio: what speech, music, and sound have in common," *Frontiers in psychology*, vol. 4, p. 292, 2013.

[30] B. Logan *et al.*, "Mel frequency cepstral coefficients for music modeling." in *Ismir*, vol. 270. Citeseer, 2000, pp. 1–11.

[31] K. V. Krishna Kishore and P. Krishna Satish, "Emotion recognition in speech using mfcc and wavelet features," in *2013 3rd IEEE International Advance Computing Conference (IACC)*, 2013, pp. 842–847.

[32] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using gmms," in *Ninth international conference on spoken language processing*, 2006.

[33] 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." 2010, pp. 255–266.

[34] E. M. Schmidt and Y. E. Kim, "Learning emotion-based acoustic features with deep belief networks," in *2011 IEEE workshop on applications of signal processing to audio and acoustics (Waspaa)*. IEEE, 2011, pp. 65–68.

[35] K. Bischoff, S. Claudiu, R. Paiu, W. Nejdl, C. Laurier, and M. Sordo, "Music mood and theme classification - a hybrid approach." 01 2009, pp. 657–662.