Application of Ensemble Learning for Speech Emotion Recognition

Shane Auberger
Department of Electrical and Computer Engineering
L3-Norm
Gainesville, FL, USA
shaneauberger@ufl.edu

Wei-Che Huang
Department of Electrical and Computer Engineering
L3-Norm
Gainesville, FL, USA
huangw1@ufl.edu

Bradley Shelley
Department of Electrical and Computer Engineering
L3-Norm
Gainesville, FL, USA
shelleyb@ufl.edu

Jose Blanco
Department of Electrical and Computer Engineering
L3-Norm
Gainesville, FL, USA
joseblanco@ufl.edu

I. Abstract

Emotion recognition can be arduous for machine learning algorithms, especially when a multitude of test samples are input from various people. A way to combat this could be the use of ensemble learning. Ensemble learning allows for a combination of multiple machine learning algorithms to come to the most accurate conclusion based upon multiple predictions. In this paper, we devise a method of emotion recognition using ensemble learning of multiple machine learning algorithms from: k-nearest neighbors (KNN), multilayer perceptron (MLP), and convolutional neural networks (CNN). A combination of these relatively accurate algorithms can establish a versatile model for emotion recognition that classifies a plethora of input data. Using ensemble learning, we were able to create a generalized and accurate model for emotion recognition. Using the collection of emotional speech recordings, following a template like the RAVDESS speech data set, from EEE 4773 recordings samples and the features extracted from these speech samples are the mel-frequency cepstrum (MFC) spectrograms, mel-frequency cepstral coefficients, and the short time Fourier transform chromagram of the sample. The ensemble model then draws classification parameters from the extracted features, which are then evaluated against the true labels for a performance measure. Our hybrid model using ensemble learning was able to achieve accuracy ratings of up to 84.2% on the given data set.

II. Introduction

A. Generalized Overview of Experimentation

To develop a model that could accurately classify emotions, members recorded audio samples of various emotions and assigned labels based on the emotion. Once this data was collected, a variety of features were considered. From these features, multiple machine learning algorithms were used to test these features. Based on individual accuracies of these algorithms, an ensemble learning approach was used to combine the top three performing algorithms to allow for versatility and generalization. Once the ensemble learning model was established, many of the extracted features were tested and the ones that yielded the highest accuracy were used to train and test the final model.

B. Literature Review

Speech is an essential mode of communication for correspondence between people that consists of various complex signal information to reflect a speakers communication message, emotion, pitch, tone, etc. It is becoming more transparent than ever there could be significant use for machines to accurately make assumptions classifying human voices emotion[1]. State of the art speech recognition is only beginning to perfect natural language speech detection without a deep level of understanding the emotional recognition. The added layer emotion to speech signal processing of human voices is proposed to boost the performance of speech recognition systems[2]. The applications of emotion recognition lead to various helpful cases in conjunction with speech recognition within criminal investigations, intelligent assistance[3], and other industries. Emotion recognition in speech is particularly useful for humans interacting with machines[4].

For speech emotion feature extraction a multitude of features can be extracted from the complex speech signal thus it comes down to finding key features to identify the emotion of the speaker's voice. In a multitude of research topics a clear favorite for emotion recognition in speech is the MFCCs [3]-[6]. Bou-Ghazale and Hansen [6] study on proposed features for speech emotion recognition found that features based on cepstral analysis, such as the MFCC will outperform other feature sets comparatively. With speech signals being significantly complex limiting feature extraction to just one feature left a significant amount of information still in the speech signal. While the MFCC is a commonly used feature for emotion extraction there is no clearly defined perfect feature for emotion extraction of a complex speech signal. To help identification of the emotion from the speech signal the additional chromagram of the speech signal was chosen as an extra representation of the signal. The

chromagram is a 12-element vector that represents the short-time energy distribution of the speech signal[7]. The 12 vectors represent the spectral energy mapped to 12 spectral bins corresponding to the chromatic scales. These 12 vectors are then concatenated across time to produce the chromagram feature.

III. Implementation

A. Data Collection

Each member of the project recorded eighty audio samples of two main phrases; each two seconds long sampled at 44kHz. Of these eighty audio samples, there were ten of each emotion (five of phrase one and five of phrase two). These audio samples were then assigned labels: 1 'neutral', 2 'calm', 3 'happy', 4 'sad', 5 'angry', 6 'fearful', 7 'disgust', and 8 'surprise'. Once audio samples were produced, features could be extracted.

B. Feature Extraction

The data would then go through two different feature extraction processes. The first process extracts MFCC features and chromagram and stacks them together creating a NumPy feature containing 20 features from the Mel-spectral coefficient of the signals matrix and 12 features from the chroma coefficient matrix concatenated together and saved as **Data1**. The second process extracts Mel-frequency spectrograms for each speech sample and saves the NumPy feature matrix as images in **Data2**. Extraction of these feature matrices was done through the Librosa packages. For the Mel-frequency spectrograms multiple different iterations of images were tested using various different values for computing the spectrogram. The convolutional neural network saw best results when the Mel-frequency spectrogram used a 1024 point fast Fourier transform, a 512 hamming windowing filter, a time shift or hop length of 256 points and a total of 128 Mel spectra frequency bins. The spits out a (128,391) feature matrix in the form of:

$$(Mel\ Frequency\ Bins, \frac{Speech\ Signal\ Sample\ Overall\ Length}{Hop\ Size\ of\ Mel\ Spectrogram})$$

The spectrogram is then normalized and given to the CNN as a single channel image in grayscale to achieve the desired normalization. Each of the samples is pre-processed in the above manners to feed the overall ensemble learning model. The use of two different feature matrices is the key attribute of our ensemble learning model. The two separate feature matrices feed our 3 separate classification models. The *Data1* feature matrix is the input to the k-NN and MLP algorithms in the model. The *Data2* matrix of Mel-frequency spectrograms is then given as input to our CNN.

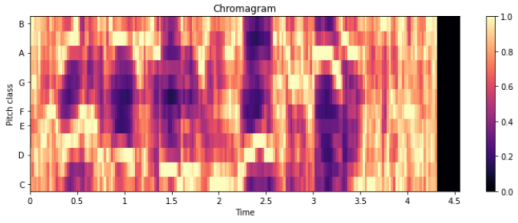

Figure 1: Chroma feature extraction visualized as the Chromogram (2-dimensional image not used. Coefficient vectors were used for features). (Neutral (1) example)

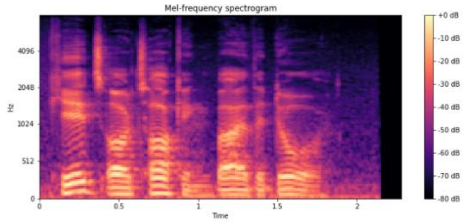

Figure 2: Mel-spectrogram feature extraction. (Neutral (1) example)

C. Model Creation, Parameter Selection, and Training

Data1 and *Data2* are both randomly split between eighty percent training data and twenty percent testing data before being fed into algorithms for parameter selection.

For the k-NN algorithm, *Data1* is used, and cross validation was performed to determine the number of neighbors "k" that would produce the most accurate predictions. Then the model would be looped through different Principal Component Analysis (PCA) transformations of *Data1* to determine the dimension that yields the best accuracy. The second algorithm, MLP, also uses *Data1*, and is looped through different numbers of dimensions using PCA components until the best accuracy is achieved. Finally, the third algorithm, CNN, utilizes *Data2* directly as interpreting the data as a 2-D image.

Figure 3: Hierarchy of the overall model.

Once each of the three algorithms return their predictions and probability of their predictions, the result is sent to an algorithm that performs hard voting when two or three of the classifiers agree with a prediction. The decision is made based upon a class that receives the most votes. When none of the classifiers agree with a certain prediction, soft voting is performed to combine the probabilities of different classes, then the algorithm would decide to predict the class that has the highest probability.

D. Validation

Multiple test sets are input to model, and final predictions are compared to the true labels to output the individual accuracies of each classifier, the confusion matrix, and the overall accuracy score of the hybrid model.

IV. Experiments

After the creation of the initial hybrid model, many tests and variations were made to allow for an increase in accuracy of the final model. Below are multiple experiments performed to develop the model that we achieved in the end.

A. Alternative Feature Extraction

Speech signals contain a numerous amount of information that can be extracted for features for use in emotion recognition of the speaker. The initial features chosen for extraction were the spectral centroids of each frequency bin of the speech signal and the tonal centroid of the feature known as a "Tonnetz" that projects chroma features onto a 6-dimensional basis for representation of certain octaves. Through testing, the features provided significantly worse predictions for features comparatively to the Mel frequency spectrogram, MFCC's and chroma features. Given the Tonnetz and tonal centroids the average result for the ensemble learning model performed significantly worse compared to other features. The MFCC was a clear favorite for extraction of emotion from speech signals and considerable research has shown that given an appropriate classification model the feature performs at sufficient accuracy ratings for recognition of emotion in speech.

B. Dimensionality Reduction Methods

For *Data1*, there were initially thirty-two dimensions. Due to the curse of dimensionality, it was crucial to lower the number of dimensions to allow for faster execution time, minimize error. Dimensionality reduction was performed using multiple methods such as: Multi-Dimensional Scaling (MDS), Isometric Mapping (ISOMAP), Locally Linear Embedding (LLE), and Principal Component Analysis (PCA). When determining the optimal dimension to reduce to, the k-NN and MLP algorithms were designed to loop through all possible dimensions and select the dimension that yielded the best results. All methods listed above were experimented with, and the results varied significantly. MDS had a sluggish execution rate, LLE and ISOMAP resulted in very poor accuracy. However, PCA had a rapid execution rate and produced the highest accuracy for the k-NN and MLP algorithm. From this, PCA was used in the final model to reduce the dimensionality of *Data1*. For *Data2*, Tensorflow Keras layers allowed for a much easier way to reduce the feature matrix in the CNN to a one-dimensional set of features in a flatten layer. The flatten layer is then followed by two dense layers to fit into a final prediction feature set to classify each of the emotion labels.

C. Implementation of Various Algorithms

Once features were extracted from the data, determining a route to take for the model creation was rather ambiguous. There are a multitude of algorithms that could be implemented to develop a model for emotion classification. However, finding the most accurately performing algorithm was a challenge. Many algorithms from k-NN, Support Vector Machine (SVM), MLP, CNN, and Recurrent Neural Network (RNN) were used in an attempt to create a model for emotion recognition. Throughout testing, many algorithms performed well on select test sets, and poorly on others. In order to combat the possibility of select algorithms performing poorly, the ensemble learning route was pursued. From this, three of the top performing algorithms were selected and used as inputs to the voting system to structure the ensemble learning model. Three algorithms were chosen specifically to allow for the majority vote system to be used and function as intended.

D. Voting System Design

At first, two designs were considered: hard voting and soft voting. The hard voting design is based on the concept of majority voting, and when none of the classifiers agree (or tie), the algorithm would choose the prediction from one of the classifiers randomly. The soft voting design is to simply add the probabilities of all classes, and to predict the class that has the largest probability. Several rounds of testing show that the hard voting design has an accuracy of around ~0.80, while the soft voting design scores around ~0.71. The weakness of the hard voting is that the accuracy is not consistent when there is a three-way tie since it uses the "rolling dice" approach. To combat this problem, the soft voting approach was adapted, since it can use probabilities to break the tie. Upon testing, the combined design can reach up to ~84% accuracy, and it consistently betters the results from all three of the classifiers.

E. Class Separation

63	8	3	5	2	0	3	2
12	63	0	2	0	0	3	0
0	2	56	1	1	1	1	1
0	1	4	66	0	4	0	0
1	0	2	1	63	0	3	4
0	0	5	2	2	72	2	3
2	0	0	0	5	1	61	1
0	0	3	0	2	1	1	59

Figure 4: Confusion Matrix of a random test set input to the Final Model

We observed that our model struggles to classify emotion neutral and calm in the confusion matrix above. This issue stems from the similar tone that the two emotions have. Therefore, we attempted to utilize two One-class SVM with RBF kernel to single out class neutral and class calm, and to add them to the majority vote system to improve the robustness of the model. However, the one-class SVMs struggled significantly to single out the class calm or neutral as outliers, and we were forced to abandon the idea and dedicate time to improve the existing classifiers.

V. Conclusions

Overall, tackling the concept of using machines to classify emotions was strenuous. Once the features were extracted, various algorithms were applied to come to a variety of classification accuracies. Throughout testing, it was apparent to apply ensemble learning due to the variation of select algorithm performance when different test sets were applied. Once this hybrid approach was established, a consistent accuracy above eighty percent was achieved, which allowed for a relatively accurate emotion recognition model.

Algorithm/Model	Accuracy
k-NN	81.2 %
MLP	74.3 %
CNN	71.5 %
Final Hybrid Model	**84.2** %

Figure 5: Chart of accuracies of each individual algorithm and the final hybrid model using a 75/25 split.

In this paper, we discovered the power of using ensemble learning for emotion recognition. Due to the variation of input samples whether being from different people, different ways of expressing emotions, etc., select algorithms can have varying performances due to the variation in the input space. Ensemble learning allowed the model to be versatile when given a range of input data, which allowed for the creation of a consistently accurate model to predict emotions.

VI. References

[1] J. Nicholson, K. Takahashi and R. Nakatsu, "Emotion recognition in speech using neural networks," ICONIP'99. ANZIIS'99 & ANNES'99 & ACNN'99. 6th International Conference on Neural Information Processing. Proceedings (Cat. No.99EX378), Perth, WA, Australia, 1999, pp. 495-501 vol.2, doi: 10.1109/ICONIP.1999.845644.

[2] W. Minker, J. Pittermann, A. Pittermann, P. Strauss, and D. Buhler, "Challenges in speech-based human–computer interfaces," Int. J. Speech Technol., vol. 10, no. 2–3, pp. 109–119, 2007.

[3] K. Wang, N. An, B. N. Li, Y. Zhang and L. Li, "Speech Emotion Recognition Using Fourier Parameters," in IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 69-75, 1 Jan.-March 2015, doi: 10.1109/TAFFC.2015.2392101.

[4] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recog., vol. 44, no. 3, pp. 572–587, 2011.

[5] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2257-2260, doi: 10.1109/WiSPNET.2017.8300161.

[6] S. E. Bou-Ghazale and J. H. L. Hansen, "A comparative study of traditional and newly proposed features for recognition of speech under stress," in IEEE Transactions on Speech and Audio Processing, vol. 8, no. 4, pp. 429-442, July 2000, doi: 10.1109/89.848224.

[7] B. G. and P. M., "Speech/music classification using visual and spectral chromagram features," Journal of Ambient Intelligence and HumanizedComputing, pp. 1–19, Apr. 2019.