Recent Trends in Intensive Computing
M. Rajesh et al. (Eds.)

© 2021 The authors and IOS Press.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/APC210225

A Study on Speech Emotion Recognitions on Machine Learning Algorithms

Shanmuk Srinivas Amiripalli [a 1], Potnuru Likhitha [a], Sisankita Patnaik [a], Suresh Babu K [a], Rampay Venkatarao [b]
[a] *Dept of CSE, GIT, GITAM University,Visakhapatnam,Andhrapradesh,India*
[b] *Dept of CSE, Wollega University, Nekemte, Eithopia*

Abstract. Speech emotion detection has been extremely relevant in today's digital culture in recent years. RAVDESS, TESS, and SAVEE Datasets were used to train the model in our project. To determine the precision of each algorithm with each dataset, we looked at ten separate Machine Learning Algorithms. Following that, we cleaned the datasets by using the mask feature to eliminate unnecessary background noise, and then we applied all 10 algorithms to this clean speech dataset to improve accuracy. Then we look at the accuracies of all ten algorithms and see which one is the greatest. Finally, by using the algorithm, we could calculate the number of sound files correlated with each of the emotions described in those datasets.

Keywords. Scikit-learn, MLPClassifier, Logistic Regression, Naïve Bayes, XGB, LightGBM, Stochastic Gradient Descent, Support Vector Machine.

1. Introduction

In classification, a set of data is categorized into classes, and it can be performed not only on structured data but also on unstructured data. Each data point of datasets is predicted into which class it falls under. These classes can be referred as targets, labels or categories. The task of the Classification predictive model is to approximate the mapping function from input variables to discrete output variables. Example for Binary Classification: While flipping a coin, the chances of getting head or tail can be categorized into two classes. Example for Multiclass Classification: There may have a 3-class classification problem of an animal set to classify as lion, tiger or leopard with a total of 100 instances. The classifier, in this case, needs training data to understand how the given input variables are related to the class. Once the classifier is trained accurately, it can detect the output of the particular testing data. This concept was used to get the accuracy of the dataset. Speech is one of the most natural means for us to communicate ourselves as humans. We depend on it so often that we can deduce its significance from other forms of contact, such as emails and instant messaging, in which we often use sentiment to convey the messages' contents. Since feelings are so central in conversation, sensing and interpreting them is crucial in today's world of remote communication.

[1] Shanmuk Srinivas Amiripalli, Dept of CSE, GIT, GITAM University,Visakhapatnam, India

Email: shanmuk39@gmail.com

It is tough to detect feelings because they are emotional. There is no agreed-upon procedure for quantifying or categorizing them. An SER paradigm is a collection of methodologies for interpreting and categorizing speech signals to detect emotions. This type of interface can be used for some items, including interactive voice-based assistants and caller-agent contact research. We examine audio files' acoustic properties in this study to detect real feelings in recorded expressions. The method of recognizing human emotion through expression and affective states is known as Speech Emotion Recognition or SER. This takes advantage of the idea that voice always reflects natural feeling by tone and pitch. It involves defining people's feelings dependent on the tone of their voice in their expressions. People cannot all talk at the same volume. It varies based on their moods and circumstances. This is the same phenomenon that allows creatures like dogs and horses to understand human feelings. It's popular in call centres. If you note, call centre workers do not communicate to each customer in the same way; their method of speaking to them differs based on the customer. Speech emotion recognition systems enable workers to understand consumer emotions through speech. In order for them to strengthen and deliver services to their customers.

2. Literature Survey

In [1-3] compares the speech emotion classification accuracy of speaker-based and the time to construct the model between Support Vector Machine(SVM) and Multi-Layer Perceptron(MLP) classifiers. The classification was performed with the WEKA unit, and the features were extracted with PRAAT. A basic SER module structure was chosen to compare the described classifiers. Confusion matrix, classification precision, and construct time are used to test supervised learning algorithms' efficiency. Despite the fact that MLP outperforms SVM in total emotion classification, SVM's preparation was quicker.MLP and SVM had acceptance rates of 78.69 and 76.82, respectively. In MLP, the highest emotion identification was for depression (89%), with pleasure and anxiety being the most perplexing emotions, while in SVM, the highest emotion recognition was for indignation (87.4%), with disgust and fear being the most perplexing emotions. In research [4-7], wavelet packet techniques were used to recognize speech sentiment. The wavelet packet coefficients were examined at five decomposition stages, analyzed, and used as inputs to Support Vector Machine (SVM) classifiers. The findings showed that using these features on seven emotional states in two languages, German and Chinese, this wavelet packet strategy increased efficiency by 4.5 percent and 16.9 percent, respectively, as opposed to a single one without these features. These two datasets have a final success average of 61.9 percent and 62.2 percent, respectively. As a result, it was concluded that wavelet packet coefficient features outperform Mel-Frequency Cepstral Coefficient (MFCC) features. The ANN-based decision fusion for SER was introduced in [7-9]. SVM, k-NN, Gaussian Mixture Model, HMM, ANN, and other sequence classification methods were used to determine which was the most efficient tool for classifying speech emotions. SVM was said to have the highest results out of all of them. Some suspected that the ANN could achieve ideal results, but they didn't know which kind was best for SER. To identify various feelings, four separate ANNs were used: Probabilistic Neural Network (PNN), Radial Base Function (RBF) network, Back Propagation (BP) network, and Elman Network. At the judgment stage, voting systems were used to fuse the recognitions utilizing Statistical and Spectral characteristics. Principal Factor Analysis decreases the

dimensionality of super vectors built from spectral features (PCA). However, since PCA was used as a dimensionality reduction method rather than a pattern recognition method, it caused some issues. Proposed decision fusion was discussed as a way to escape them. The proposed decision fusion was successful, and the dimensionality reduction was probable, according to the results. In research [10-12] proposes a speaker-independent approach for categorizing emotional vocal sounds. The treatment divided the mechanism of recognizing emotions into two sections. The first phase entails a coarse encoding and grouping of six emotional states to determine which pair of emotions has the highest likelihood. Low-level encoding procedures were suggested at this time, and the extracted features were combined to produce the best emotional state descriptive acoustic vectors. Second, utilizing the Sequential Floating Forward Selection (SFFS) algorithm, modern encoding strategies were used to define a special collection of acoustic features for each pair of emotions that can be used to discriminate between them. There are a total of 72 high-level acoustic features.

3. Proposed Classification Algorithm

The classification is a supervised learning principle of Machine Learning that separates a dataset into groups. Speech Expression Recognition, Face Identification, Handwriting Recognition, Text Classification, and other classification problems are some of the most important. It may also be a multiclass problem or a binary classification problem. In Machine Learning, there are many classification algorithms. On the RAVDESS, TESS, and SAVEE datasets, the following algorithms were used. First of all, we give some audio datasets as input. Extracted features from those speech files. Declared one dictionary for emotions in the dataset and another dictionary for emotions that we want to observe. Loaded the dataset and split into two subsets i.e. 75% of data for training and 25% of the data for testing. Initialized a classifier and trained the model using the dataset to predict the emotions of each of the speech files. Finally, it gives the emotion as output. We took 3 different datasets, namely RAVDESS, SAVEE and TESS, which consist of different emotions. We found the accuracies of the ten different classifiers Logistic Regression, Naïve Baye's, Stochastic Gradient Descent, KNN, Decision Tree, Random Forest, Support Vector Machine, MLPC, XG Boost and Light GBM, for each dataset and compared them to know which classifier have more accuracy [13-16]. The accuracy was calculated before and after masking of the datasets. Depending on the accuracy, we got to know that MLP Classifier has more accuracy compared to others. Before masking, the accuracy for RAVDESS, SAVEE, and TESS in MLPC were 70%, 100% and 80%, respectively and after masking, the accuracy for RAVDESS, SAVEE, and TESS in MLPC were 75.60%, 100% and 84%, respectively [17-21].

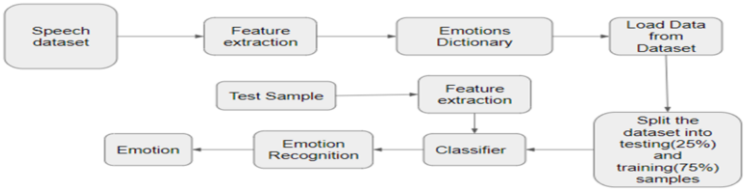


Figure 1. Block Diagram of SER

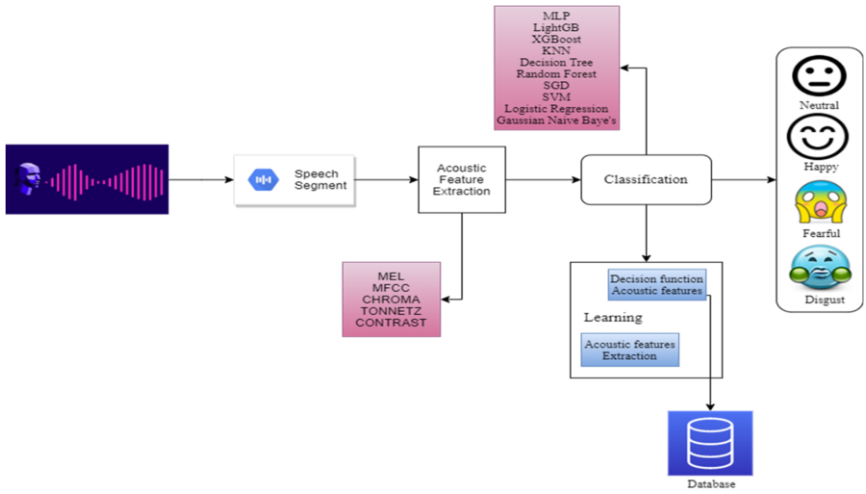
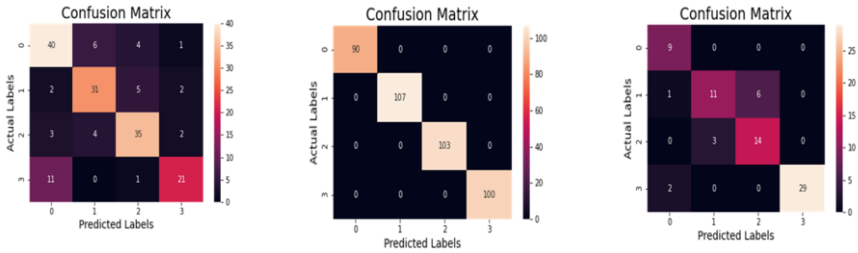


Figure 2. Architecture of SER

Algorithm
Step1: *Start* ***Step2:*** *Imported all required packages, libraries and modules.* ***Step3:*** *Considering five standard features of any audio file and declared a dictionary that contains the emotions in the dataset and a list with the emotions observed.* ***Step 4:*** *Input datasets are considered in this research is RAVDESS, TESS, SAVEE.* ***Step 5:*** *Configuring the experiment with considered 75% of data for training and 25% of data for testing.* ***Step 6:*** *Initialized a classifier model and fit the model.* ***Step 7:*** *Accuracy for ten different classifiers and picked one classifier with more accuracy.* ***Step 8:*** *Using that classifier, we found a count of files of each of the emotions in each dataset.* ***Step 9:*** *Classification reports and confusion matrix for each of them are drawn.* ***Step10:*** *Stop*

4. Results and Discussion

After considering four emotions only as the observed emotions increases, there may be a chance to decrease accuracy, so we did not consider all emotions. RAVDESS contains eight emotions, whereas TESS and SAVEE contain seven emotions, so we took four common emotions from these three datasets. We took emotions happiness, fear, disgust and neutral in our observation. The below are the results for the count of files of each emotion we considered in each dataset.



(a) (b) (c)

Figure 3. Confusion Matrix of MLP Classifier using (a) RAVDESS Dataset (b) TESS Dataset (c) SAVEE Datasets

Table 1. Comparison of Modulation schemes

Emotion	Number of files in RAVDESS	Number of files in TESS	Number of files in SAVEE
HAPPY	45	103	20
FEARFUL	41	107	14
DISGUST	56	90	12
NEUTRAL	26	100	29

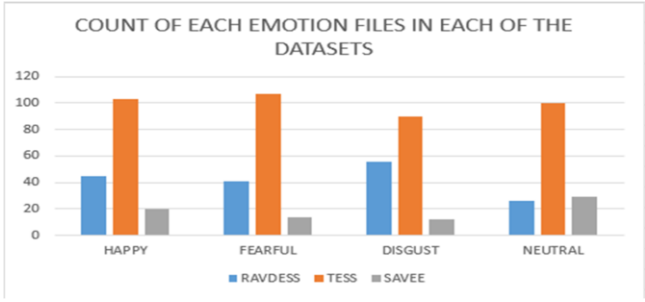


Figure 4. Count of files of each emotion in each dataset

5. Conclusion

In the above research, the major observation is that MLP Classifier is the best classifier compared to any other classifier in Machine Learning. However, to improve this model's accuracy, we need to clean the noise in our dataset. We can improve the efficiency of the model from 71% to 76% for RAVDESS, accuracy for TESS before, and masking remained constant, i.e. 100%, 80% to 84% for SAVEE by considering four emotions such as neutral, happy, fearful and disgust. We can use the model to predict the emotions of the audio files in the datasets and any other sample audio files with the extension .wav.

References

[1]. Idris, I., & Salam, M. S. H. (2014, December). Emotion detection with hybrid voice quality and prosodic features using neural network. In 2014 4th World Congress on Information and Communication Technologies (WICT 2014) (pp. 205-210). IEEE.

[2]. Wang, K., An, N., & Li, L. (2014, September). Speech emotion recognition based on wavelet packet coefficient model. In The 9th International Symposium on Chinese Spoken Language Processing (pp. 478-482). IEEE.

[3]. Xu, L., Xu, M., & Yang, D. (2009). ANN based decision fusion for speech emotion recognition. In Tenth Annual Conference of the International Speech Communication Association.

[4]. Atassi, H., & Esposito, A. (2008, November). A speaker independent approach to the classification of emotional vocal expressions. In 2008 20th IEEE international conference on tools with artificial intelligence (Vol. 2, pp. 147-152). IEEE.

[5]. Huang, C., Jin, Y., Zhao, Y., Yu, Y., & Zhao, L. (2009, September). Speech emotion recognition based on re-composition of two-class classifiers. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops (pp. 1-3). IEEE.

[6]. Zhang, J., Yin, Z., Chen, P., & Nichele, S. (2020). Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion, 59, 103-126.

[7]. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., & Mahjoub, M. A. (2018, January). Speech Emotion Recognition: Methods and Cases Study. In ICAART (2) (pp. 175-182).

[8]. Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017, March). A review on emotion recognition using speech. In 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT) (pp. 109-114). IEEE.

[9]. Tarunika, K., Pradeeba, R. B., & Aruna, P. (2018, July). Applying machine learning techniques for speech emotion recognition. In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

[10]. Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic speech emotion recognition using machine learning. In Social media and machine learning. IntechOpen.

[11]. Amiripalli, S. S., Bobba, V., & Potharaju, S. P.: A novel trimet graph optimization (TGO) topology for wireless networks, (2019) doi:10.1007/978-981-13-0617-4_8.

[12]. Potharaju, S. P., & Sreedevi, M. (2018). A novel cluster of quarter feature selection based on symmetrical uncertainty. Gazi University Journal of Science, 31(2), 456-470.

[13]. Amiripalli, S. S., & Bobba, V.: Research on network design and analysis of TGO topology. International Journal of Networking and Virtual Organisations, 19(1), 72-86, (2018).

[14]. Potharaju, S. P., & Sreedevi, M. (2017). A Novel Clustering Based Candidate Feature Selection Framework Using Correlation Coefficient for Improving Classification Performance. Journal of Engineering Science & Technology Review, 10(6).

[15]. Amiripalli, S. S., & Bobba, V. (2019). An Optimal TGO Topology Method for a Scalable and Survivable Network in IOT Communication Technology. Wireless Personal Communications, 107(2), 1019-1040.z

[16]. Potharaju, S. P., Sreedevi, M., & Amiripalli, S. S. (2019). An Ensemble Feature Selection Framework of Sonar Targets Using Symmetrical Uncertainty and Multi-Layer Perceptron (SU-MLP). In Cognitive Informatics and Soft Computing (pp. 247-256). Springer, Singapore.

[17]. Amiripalli, S. S, V. Bobba, "A Fibonacci based TGO methodology for survivability in ZigBee topologies". INTERNATIONAL JOURNAL OF SCIENTIFIC &TECHNOLOGY RESEARCH, 9(2), pp. 878-881. (2020).

[18]. Ramiah Chowdary, P., Challa, Y., Jitendra, M.S.N.V.: "Identification of MITM Attack by Utilizing Artificial Intelligence Mechanism in Cloud Environments" Journal of Physics: Conference Series, 1228 (1),012044, (2019).

[19]. Thota, J.R., Kothuru, M., Shanmuk Srinivas, A., Jitendra M, S.N.V.: Monitoring diabetes occurrence probability using classification technique with a UI, International Journal of Scientific and Technology Research, 9 (4), pp. 38-41, (2020).

[20]. Potharaju, S. P., Sreedevi, M., Ande, V. K., & Tirandasu, R. K. (2019). Data mining approach for accelerating the classification accuracy of cardiotocography. Clinical Epidemiology and Global Health, 7(2), 160-164.

[21]. Jitendra, M.S.N.V., Radhika, Y.: A review: Music feature extraction from an audio signal, International Journal of Advanced Trends in Computer Science and Engineering, 9 (2), pp. 973-980, (2020).