International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 05 | May 2019 www.irjet.net p-ISSN: 2395-0072

EMOTION RECOGNITION FROM VOICE

Abhijeeth V madalgi[1], Bhyravi S [1], Hemanthkumar M[1], Jagadish D N[1], Kavya N L[2]

[1]Student, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bengaluru
[2]Professor, Department of Computer Science and Engineering, Sapthagiri College of Engineering, Bengaluru

---

Abstract - *Identification and analysis of human emotions associated with voice is one of the major challenges involved in emotion recognition systems. This project presents development of such a system that involves application of Machine learning classification algorithms over a set of audio features extracted from various speech segments. These segments contain a set of recorded sentences by people across the world who express different emotions.*

Key Words: Classification, Naïve-Bayes, SVM, Random forest, confusion matrix.

1. INTRODUCTION

Typically, speech recognition engines or ASRs simply transcribe audio recordings into textual content. But a lot of meta-information, apart from the text, exists. Examples might be the speaker's intonation, emotion, loudness, shades etc. The emotions may so influence the speech that the entire meaning that the speaker intends to convey is turned into opposite- which is regarded sarcasm or irony. Therefore, there exists a direct co-relation between characteristics of the speech and the emotion. A person angry might have the loudness levels higher than normal while the speech of a person afraid, would have a higher pitch. This project intends to analyze these features and establish a co-relation between the acoustic features and the emotion expressed, thereby allowing the prediction of emotion from a given voice sample.

Emotion recognition is a hot topic that's been on research for decades, which intends to find the nature of audio and the system proposed classifies into the seven basic human emotions, anger, boredom, sadness, neutral, happiness, fear and disgust. Today's advanced technologies consists of tools for efficient retrieval, analysis and classification of auditory data into different classes. In today's digital world lot of data is available for emotional analysis from image, text, audio and video through media networks like YouTube, soundcloud etc., However, they raise concerns on privacy and copyright laws. A well-defined labelled dataset can be obtained through the berlin database of emotional speech, a database purely for research. Supervised machine learning techniques are therefore used to train the data for classification later on.

The nature of human emotions and their influence on speech provides invaluable insights into how the features of emotion and the underlying emotion are related. This relation may not be consistent with people all over the world, since the ethnicity and culture of various groups largely influence how emotions are expressed. Therefore, a need for machine learning algorithms arise since a distinct hypothesis consisting of continuous static values may not be consistent with the emotions expressed by various people.

The project presents the application of supervised machine learning techniques such as Naïve Bayes Classifier, Support Vector Machines, and Random Forest classifiers.

2. BACKGROUND

A. Dataset

The dataset is obtained by extracting features from the berlin audio database. The berlin dataset consists of lines spoken by various professional actors, and grouped into the seven basic classes of emotion- anger, boredom, sadness, neutral, happiness, fear and disgust.

B. Features

The distinct attributes of speech that are a result of the influence of human emotion are generally associated with the frequency, loudness, and the rate of utterance. The human ear does not enable the perception of small differences in the frequencies of the sound, but rather, perceives it over a scale known as Mel scale. The frequencies mapped to the Mel scale, are MFCCs (Mel frequency cepstral coefficients) and prove to be vital in providing insight about how humans perceive frequency. Other attributes such as loudness, amplitude, zero crossing rate, and spectral centroid are used as the features that determine the underlying emotion,

C. Naive Bayes Classifier

Naive Bayes Classifier implements Bayes theorem with a solid independence assumption [1], that the features are independent of each other. Bayes Theorem works on conditional probability which finds out the probability of an event given that some other event has already occurred. It predicts the conditional probability of a class given the set of evidences and finds the most likely class based on the maximum likelihood hypothesis. The naive Bayes classifier is a famous and popular technique because it is much quicker and performs well with less data [2]. The Gaussian naïve Bayes classifier is used in cases where the features are normally distributed and continuous in nature. For example, suppose the training data contains a continuous attribute, x. We first segment the data by the class, and then compute the mean and variance of x in each class. Let μk be the mean of the values in x associated with class Ck, and let σ^{2k} be the

variance of the values x associated with class Ck. Suppose we have collected some observation value v. Then, the probability distribution of v given a class Ck, p(x=v| Ck), can be computed by plugging v into the equation for a Normal distribution parameterized by μk and σ^{2k}. That is,

$$p(x = v \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{(v-\mu_k)^2}{2\sigma_k^2}}$$

D. Support Vector Machine

Support Vector Machine(SVM) is a supervised learning algorithm that analyses the data and recognizes patterns [3]. Given an input set, SVM classifies them as one or the other of two categories. SVM can deal with both linear and non-linear classification.

With kernel trick, it can efficiently perform non-linear classification. It does so by mapping the input set to high dimensional feature space. The types of kernel include polynomial, Gaussian radial basis function (RBF), Laplace RBF kernel, Hyperbolic tangent kernel and Sigmoidal kernel. Construction of hyper plane is employed by SVM for the classification.

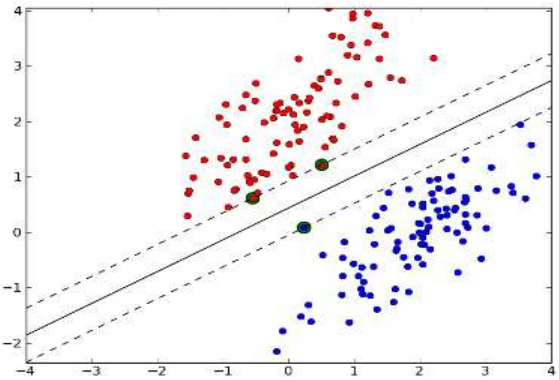

Fig. 1. Support vector classification with linear kernel

E. Random Forest Classifier

Random forest algorithm is a supervised learning algorithm which can be used for both classification and regression. Random forests create decision trees on randomly selected data samples, obtains the predictions from each tree and selects the best solution by means of voting. Random data samples are used to generate decision trees, and based on the classification performance of each of these decision trees, the best sub-decision trees are selected. Random Forest Classifiers are often known to avoid overfitting [4].

I.How random forests work

Most of the options depend on two data objects generated by random forests.

when the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob (out-of-bag) data is used to get a running unbiased estimate of the classification error as trees are added to the forest. It is also used to get estimates of variable importance.

After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data.

II. The out-of-bag (oob) error estimate

In random forests, there is no need for cross-validation or a separate test set to get an unbiased estimate of the test set error. It is estimated internally, during the run, as follows:

Each tree is constructed using a different bootstrap sample from the original data. About one-third of the cases are left out of the bootstrap sample and not used in the construction of the kth tree.

Each case left out in the construction of the kth tree is put down the kth tree to get a classification. In this way, a test set classification is obtained for each case in about one-third of the trees. At the end of the run, take j to be the class that got most of the votes every time case n was oob. The proportion of times that j is not equal to the true class of n averaged over all cases is the oob error estimate. This has proven to be unbiased in many tests [5].

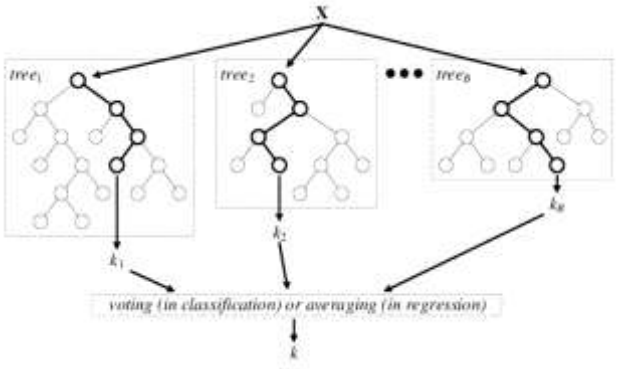

Fig. 2. Architecture of Random Forest model

3. METHODOLOGY

The workflow in this paper is divided into four phases

i. Audio pre-processing
ii. Feature Extraction
iii. Training
iv. Testing

i. Audio pre-processing

An audio signal is a representation of sound, typically using a level of electrical voltage for analog signals, and a series of binary numbers for digital signals. Audio signals have frequencies in the audio frequency range of roughly 20 to 20,000 Hz, which corresponds to the lower and upper limits of human hearing. Audio signals may

be synthesized directly, or may originate at a transducer such as a microphone, musical instrument pickup, phonograph cartridge, or tape head. Loudspeakers or headphones convert an electrical audio signal back into sound [6].

Audio files are susceptible to noise hence noise removal and enhancement of speech signal is necessary for the retrieval of desired features, which are helpful in creating an efficient model. The logMMSE algorithm explained in [7] is used for the noise reduction and enhancement of speech signal.

ii. Feature Extraction

The human ear is largely sensitive to changes in frequency of sounds having less frequency i.e. the human ear is capable of distinguishing low frequency sounds and this capability decreases as the frequency range increases. This necessitates the formulation of the Mel scale of frequency which is analogous to what the human ear can hear. The quantity MFCC represents the human hearing equivalent of frequency and pitch, the features which contain maximum emotional data. The steps to compute the MFCCs of a given audio sample are as follows:

1 Application of Fourier transform over the obtained power spectrum,

2 Application of Mel filter banks to the transformed power spectra, summing up the energies,

3 Taking logarithms of the energies computed.

4 Taking the discrete cosine transform of the logarithms of the filter bank energies.

5 Retention of coefficients 1-13 and discarding the rest

Zero crossing rate is another quantity that represents the frequency of the sound. Zero crossing rate is a measure of the number of times the amplitude of the sound waves crosses zero, which essentially gives a measure of the frequency of the wave.

Spectral centroid characterizes the spectrum. It gives a measure of the location of the "centre of mass" of the spectrum. It is mathematically computed as the weighted mean of frequencies determined by Fourier transforms.

The MFCCs, zero crossing rate, and the spectral centroid, after scaling and transformation, form the features that are to be used as attributes to train a machine learning model.

iii. Training

Three Predictive models are created using naïve Bayes, support vector machine and Random Forest algorithms which were discussed afore. The data file would contain seventeen columns with thirteen columns representing thirteen MFCC coefficients necessary for speech analysis, and three other columns representing zero-crossing rate, spectral centroid and standard deviation of MFCC coefficients respectively and final column containing the values from one to seven representing the seven emotions. The model is trained on the basis of these seventeen features.

iv. Testing

In testing phase, the 30% of data which was split randomly from the dataset is tested on the predictive model. The test data is pre-processed and classified into one of the seven emotions Anger, Happiness, Disgust, Sad, Fear, Boredom and Neutral.

4. RESULTS

The results are evaluated on the basis of accuracy, precision and recall.

TABLE I. CONFUSION MATRIX PARAMETERS

		Predicted class	
Actual Class		Class = Yes	Class = No
	Class = Yes	True Positive	False Negative
	Class = No	False Positive	True Negative

True Positives (TP) - These are the correctly predicted positive values which means that the value of actual class is yes and the value of predicted class is also yes.

True Negatives (TN) - These are the correctly predicted negative values which means that the value of actual class is no and value of predicted class is also no

False Positives (FP) – When actual class is no and predicted class is yes.

False Negatives (FN) – When actual class is yes but predicted class in no.

Accuracy - Accuracy is simply a ratio of correctly predicted observation to the total observations.
The formula to find accuracy is given by,

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+FN+TN}$$

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.
Precision is given by,

$$\text{Precision} = \frac{TP}{TP+FP}$$

Recall - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes.
Recall is given by,

$$\text{Recall} = \frac{TP}{TP+FN}$$

Gaussian Naïve Bayes algorithm gave an output of about 0.8571 accuracy, 0.8567 recall, 0.8563 precision and the confusion matrix is as shown below

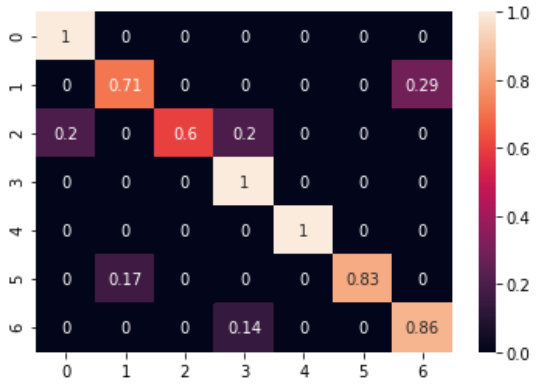

Fig. 3. Confusion Matrix for Gaussian Naïve Bayes algorithm

Support vector machine algorithm gave an output of about 0.8809 accuracy, 0.8832 recall, 0.8821 precision and the confusion matrix is as shown below

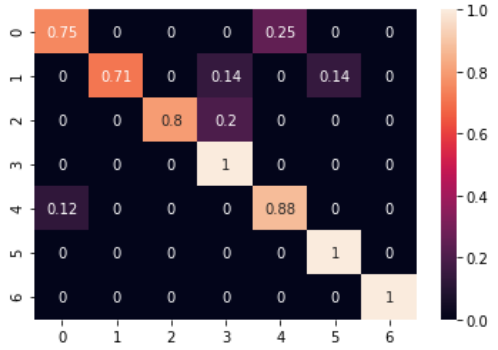

Fig. 4. Confusion Matrix for support vector machine algorithm

Random Forest algorithm gave an output of about 0.7857 accuracy, 0.7753 recall, 0.7748 precision and the confusion matrix is as shown below

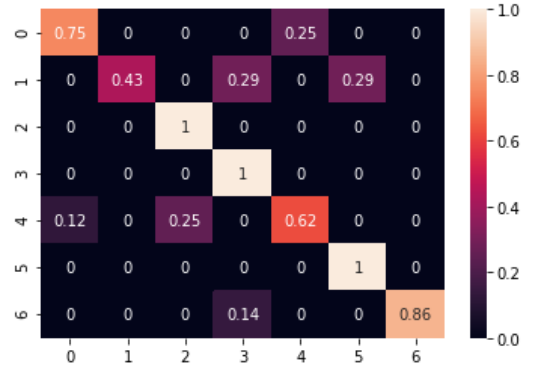

Fig. 5. Confusion Matrix for Random Forest algorithm

TABLE II. RESULTS

PERFORMANCE METRICS			
ALGORITHM	**ACCURACY**	**PRECISION**	**RECALL**
GAUSSIAN NAÏVE BAYES	0.8571	0.8567	0.8563
SUPPORT VECTOR MACHINE	0.8809	0.8832	0.8821
RANDOM FOREST	0.7857	0.7753	0.7748

5. CONCLUSION

Based on the above observations, we can conclude that for the context of emotion recognition SVM has performed well since the accuracy is higher for SVM. Even though human level accuracy is not achieved, the result obtained is good enough and can be used for further research. The proposed system may be assimilated with existing systems that detect emotion based on visual data, wherein the detection is carried out based on both auditory and visual features.

ACKNOWLEDGEMENT

Thanks to Dr. Kamalakshi Naganna for their valuable suggestions and thanks to everyone who has given valuable information about technologies through their research papers, journals and conference papers and finally thanks to all my friends and family for their moral support.

REFERENCES

[1] A. N Hasan, B. Twala, and T. Marwala,Moving Towards Accurate Monitoring and Prediction of Gold Mine Underground Dam Levels, IEEE IJCNN (WCCI) proceedings, Beijing, China, 2014.

[2] T.Mitchell, H.McGraw, "Machine Learning, Second Edition, Chapter One", January 2010.

[3] Xu, Shuo & Li, Yan & Zheng, Wang. (2017). Bayesian Gaussian Naïve Bayes Classifier to Text Classification. 347-352. 10.1007/978-981-10-5041-1_57.

[4] https://www.datacamp.com/community/tutorials/random-forests-classifier-python

[5] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm

[6] Hodgson, Jay (2010). *Understanding Records*, p.1. ISBN 978-1-4411-5607-5.

[7] Ephraim, Y. and Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust., Speech, Signal Process., ASSP-23(2), 443-445.