The use of Automatic Speech Recognition in education for identifying attitudes of the Speakers

Lomthandazo Matsane
School of Computer Science and Applied Mathematics
DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP)
The University of the Witwatersrand
Johannesburg, South Africa
lomthandazomatsane@gmail.com

Ashwini Jadhav
Faculty of Science
The University of the Witwatersrand
Johannesburg, South Africa
Ashwini.Jadhav@wits.ac.za

Ritesh Ajoodha
School of Computer Science and Applied Mathematics
The University of the Witwatersrand
Johannesburg, South Africa
ritesh.ajoodha@wits.ac.za

***Abstract*—State-of-the-art Automatic Speech Recognition (ASR) systems convert the spoken words into a corresponding text. One of the problems faced in ASR is that speakers have a different way of pronouncing words, and their accents are different from one speaker to another due to age, gender, nationality, rapidity of words, expressive form of the speaker. This paper uses two data sets, Surrey Audio-Visual Expressed Emotion (SAVEE) and The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) data sets to determine the effect of the tone in the learning environment by using the ASR and check which classifier is giving the best result. Feature such energy, Mel filter Central coefficients, energy etc. were extracted using jAudio and Waikato Environment for Knowledge Analysis (WEKA) data mining tools was used for classification. Classifiers called multi-layer Perceptron (MLP) neural network model, Support Vector Machines (SVM), Simple Logistic Regression (SLR), K-Nearest Neighbour (K-NN) and Random Forests (RF) was used to obtain the results of the emotion state for the both data sets. The data sets used to train the classifiers are in ARFF format. The results show that SAVEE data sets overcomes RAVDESS data sets in overall emotion classification performance. The result shows that RF performed better than the other classifier. The performance of classification models is evaluated in WEKA using 10-fold cross validation. The presented study examines seven emotions - anger, happiness, sadness, fear, surprise, disgust and neutral.**

***Index Terms*—Feature extraction, Emotional state, Machine Learning Classification, Automatic Speech Recognition**

I. Introduction

In Higher education Emotional state play crucial role as effective medium to identify psychological of a speaker by expressing emotions, attitudes and behaviour. Emotion also drives attention, which in turn drives learning and memory. But because we don't fully understand our emotional system, we don't know exactly how to regulate it in school, beyond defining too much or too little emotion as misbehaviour Education [1]. Research shown that modelling emotion improve the efficient of learning . We have rarely incorporated emotion comfortably into the curriculum and classroom. Further, researchers has not fully addressed the important relationship between a stimulating and emotionally positive classroom experience and the overall health of both students and staff. Learning is challenging in an environment full of anxiety, tension, and distraction. At some period, entirely learners feel some anxiety which prevents them from learning and partaking in the teaching space. This cause student to lose focus or forget what has been said in the learning environment in a short period of time. This research study address the use of automatic emotional state to recognize the effect of the tones using five supervised machine learning algorithm. Classification such as Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), K-Nearest Neighbor know as IBK in WEKA, Simple Logistic Regression (SLR), and Random Forests we used to train the emotional state of the speakers. The main reason why we are interested in conducting such research study was that speaker (lecture or teacher) must be attentive, vigilant, and not easily distracted in the learning environment. This study elaborate the use automatic speech recognition (ASR) in education by providing model that will be capable to identifying the effect of emotional state in the learning environment. Since, "Emotion is an essential part of any human decision-making and planning, and the famous distinction made between reason and emotion is not as clear as it seems". The questions that we are trying to answer are : (1) What are the best features that can be used in the speech recognition for building up better classification for identifying the effect of the tones in the learning environment? (2) How can ASR technology determine the accuracy model of emotional state to improve the learning environment?.

The main aim of this study is to have Automatic Speech Recognition model capable of monitoring and evaluating the attitudes of the speakers to improve the teaching platform in education. The main contribution of this work from learner perspective. We provide ASR model that will be used to monitor and watching for signs of confusion, curiosity, and enjoyment of the speaker by providing feedback based on their current emotional which can lead to greater success. More specifically, the results of this study will be beneficial to education by providing classification models capable of determining the effect of the tones in the learning environment by giving the correction and feedback. We note that the Random Forests (decision tree) achieves the best performance with a 96%

accuracy over the four classifiers, which is significantly better classifier for emotional classification problem. This document introduces the research field chosen for this research. It then explains the problem statement, the purpose, literature review, methodology, and the contribution of the study. The next section focus on the literature review. Section III explains methodology by describing the datasets, features extraction, feature selection and different classifiers and analysis; section IV outlines the results of the prediction model for emotional state; and section V discusses the contributions of this research work, emphasising on the research questions to answer the objectives; and recommendations for future work.

II. Literature Review

Automatic Speech Recognition (ASR) defined as the field of artificial intelligence that capable of converting spoken words into the text. ASR has been investigated widely in the fields of human-machine interaction, including education, and other sectors. ASR is manly use in education as a way to improve the efficient of learning and teaching to help both the students, and lectures. The use of ASR in education serves as the medium to identify the lack of communication skills, pronunciation, reading and teaching styles by providing corrections, and feedback. During conservation, people are constantly sending and receiving different nonverbal clues, communicated through speech signals, and facial expressions.

The ASR system allowed researchers to search for ways to extract features from the speech that allow discriminant between different words. The features such as Mel-filter Cepstral coefficients (MFCC), delta, energy component were extracted from the research study done by Kurniawa to determine the factors that influence the speaking skills. Also Choukiker extracted features, such as pitch, energy, duration, and MFCC from samples of emotional datasets [2]. Furthermore Tickle makes use of low-level descriptors such as intensity, Loudness, 12 MFCC, pitch, probability of voicing, and Delta regression coefficients [3]. This study extracted features such as MFCCs, spectral variability, compactness etc. This research study will use some of the features mentioned above.

The study of developing ASR application to identifying emotional states from recorded audio have become one of the trends topic in the field of data science. Different databases are being used to train different classifiers using different emotional datasets such Berlin Database of Emotional speech that was used in research study that was conducted by Elshaw [3], Emotion Database [2], and SAVEE database that contain 480 samples of emotional. Many application of ASR systems integrate the tones of the speaker by extracting the emotional states of a spoken words [4]. Several researchers have recently tackled the use of speech recognition systems by making the use of machine learning models. Datcu perform research study based on develop a bimodal system for emotion recognition that uses Hidden Markov Models to learn the speech audio to recognize the emotional states model [5]. Hossain makes use of the Convolutional Neural Network, Hidden Markov Model, ELM-based fusion and Support Vector Machine (SVM) classifier, and the trained models were converted into function using statistical analysis to obtain the results [6]. The classification performed to determine the emotional states of ASR are HMM, SVM, Gaussian Mixture Model(GMM), K-nearest neighbours (KNN), and types of Artificial Neural Network (ANN) such as Convolutional Neural Network(CNN), Multilayer Perceptron.

III. Methodology

In this paper we used supervised machine learning algorithm to automatically learn the attitude of the speaker based on their tones in order to improve the learning environment. Two emotion datasets are collected in the form of speech audio as secondary data from Kaggle. Five classification algorithm we used to train each of the emotion datasets. To get the speech data to be ready for training and evaluation, we first extracted different features using jAudio, follow by labelling datasets, feature selection using Information Gain Ranking (IGR) and train the classifiers. Analysis of the final results a interpreted using cross-validation.

A. Datasets

Two datasets used in this study was obtained from Kaggle namely, SAVEE and RAVDESS datasets. The SAVEE and RAVDESS datasets composed of English speech audio files of different speakers. This step answers to the 1st objective of our research: To collect emotional datasets from the online platform (Kaggle). To accomplish this objective, we collected two datasets, SAVEE and RAVDESS datasets from kaggle as free open source. These datasets are described in the subsections below.

1) Surrey Audio-Visual Expressed Emotion (SAVEE): is the emotional recognition dataset. It consists of recordings from 4 male actors altering 7 different emotions, 480 British English utterances in total. Emotion has been described psychologically in discrete categories: anger, disgust, fear, happiness, sadness and surprise [**?**].

2) Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset: The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 1440 speech audio files. Speech samples which consist of Labels: total = 1440, neutral, calm , happy , sad , angry, fearful , disgust and surprised contain 192 classes [7].

B. Feature Extraction

The second objective of this study was: to extract different features of each of the speech audio of emotional state. Feature extraction is one of the aspect used to find the most compacted and informative set of features (distinct patterns) to enhance the efficiency of the classifier. Feature extraction is accomplished by changing the speech waveform to a form of parametric representation at a relatively lesser data rate for subsequent processing and analysis. We extract different features using jAudio. jAudio is a new framework for feature

extraction designed to eliminate the duplication of effort in calculating features from an audio signal [8].

jAudio provides three basic metafeature classes (mean, standard deviation, and derivative), which are also combined to produce two more metafeature (derivative of the mean and derivative of the standard deviation). There are 27 distinct features implemented in jAudio. Based on the list, MFCCs are popular audio features extracted from speech signals for use in recognition tasks and widely used for speaker and speech recognition [9]. The complete lists of all the features used in this study are listed below:

a) Spectral Rolloff Point: Spectral Rolloff Point is the fraction of bins in the power spectrum at which 85% of the power is at lower frequencies.

b) Spectral Flux: Spectral Flux is a measure of how quickly the power spectrum of a signal is changing, calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame. The spectral flux feature is useful for the distinction of speech signals, since speech has a higher rate of change.

c) Compactness: Good measure of how important a role regular beats play in a piece of speech audio [9].

d) Spectral Variability: Statistical variability measures dispersion in data, i.e. how closely or spread-out the signal is clustered. We can achieve this by measuring the standard deviation of the magnitude spectrum of the signal [10] .

e) Zero Crossing Rate: The rate of sign-changes of the signal during the duration of particular frame [9].

f) Mel Filter Cepstral Coefficients (MFCCs): The MFCC feature extraction technique basically includes windowing the signal, applying the rule, taking the log of the magnitude, and then wrapping the frequencies on a Mel scale, followed by applying the inverse discrete cosine transform.

g) Strongest Frequency Via Spectral Centroid: An estimate of the strongest frequency component of a signal found via the spectral centroid.

h) Strongest Frequency Via Zero Crossings: An estimate of the strongest frequency component of a signal found via the number zero-crossings.

i) Linear Predictive Coding (LPC): is to minimize the sum of the squared differences between the original speech signal and the estimated speech signal over a finite duration.

j) Root Mean Square: is the averaged signal magnitude of each window.

k) Energy: Energy is measured by calculating the RMS of a discrete-time signal.

l) Method of Moments: feature that consists of the first five statistical moments of the spectrograph.

C. Feature Selection

This is the process of selecting the features that are relevant to improve the accuracy of the classifiers. Information Gain Ranking algorithm called filter methods was used in Weka data mining for selecting the relevant features and eliminating irrelevant and redundant features. Information Gain algorithm is also being used to evaluate the features that are worth to improve the accuracy and reduce the training time. Table I shown features for SAVEE data sets that are extracted using jAudio and features that are eliminated using IGR. The features that had the highest rank was selected to model our five classifier. Extracted 127 features and manage to select 100 attributes for SAVEE datasets and for RAVDESS datasets we eliminated 29 features. The highlighted rows represent the features that are being eliminated from both the datasets.

Table I: Extracted Features vs Eliminated features using Information Gain Ranking Filter

Extracted Features	**Rep.**	**Dim.127**
Spectral Rolloff Point	Mean + SD	2
Spectral Flux	Mean + SD	2
Compactness	Mean + SD	2
Spectral Variability	Mean + SD	2
MFCCs	Mean + SD	26
LPC	Mean + SD	20
Methods of Moments	Mean + SD	10
Strongest Frequency vs Spectral Centroid	Mean + SD	2
Strongest Frequency vs Zero Crossing	Mean + SD	2
Peak Based Spectral Smoothness	Mean + SD	2
Area Moment of Method	Mean + SD	20
Root Mean Square	Mean + SD	2
Relative Different Function	Mean + SD	2
Strongest Frequency vs FFT Maximum	Mean + SD	2
Zero Crossings	Mean + SD	2
Beat Sum	Mean + SD	2
Strongest Beat	Mean + SD	2
Fraction of low energy	Mean + SD	2
Strenght of Strongest Beat	Mean + SD	2
Area Method of Memonts of MFCCs	Mean + SD	20
Elimanted Features for RAVDESS	**Rep.**	**Dim.27**
MFCCs	Mean + SD	9
Strongest Beat	Mean + SD	2
LPC	Mean + SD	4
Beat Sum	Mean + SD	2
Compactness	SD	1
Fraction of low energy	Mean + SD	2
Strongest Beat	Mean + SD	2
Strongest Beat	Mean + SD	2
Area Method of Memonts of MFCCs	SD	1
Method of Memonts	SD	3
Area Method of Memonts	SD	1
Elimanted Features from RAVDESS datasets	**Rep.**	**Dim.29**
Spectral Rolloff Point	SD	1
Strongest Frequency vs Spectral Centroid	SD	1
LPc	Mean +SD	7
Method of Memonts	Mean	1
Area Method of Memonts	SD	6
MFCC	Mean	1
Partial Based Spectral Centroid	Mean + SD	2
Partial Based Spectral Flux	Mean + SD	2
Strongest Beat	Mean + SD	2
Strenght of Strongest Beat	Mean + SD	2
Beat Sum	Mean + SD	2
Area Method of Memonts of MFCCs	SD	1
Spectral Rolloff Point	SD	1

D. Classification

This study adopted the supervised method for building the classifiers. WEKA data mining tool was used to train and validating the different models. WEKA is a collection of machine learning algorithms for data mining tasks. At the start, the classifiers learn the data in a supervised manner. A model is built from the training data set which emotion of the class label is known. The training dataset is made up of emotional occurrences with the labels specified for them once the model

is built. For emotional classification, mostly Artificial Neutral, SVM, and K-NN have been used and perform well in emotional classification. For our method, we used IBK for K-Nearest Neighbour (K-NN), Multi layer perceptron (MLP), Random Forest (RF), Simple Logistic Regression (SLR), and Support Vector Machines (SVM) to build our classifiers since they are good and perform better in the classification problem. Table II show how each of the classifiers are implemented in Weka.

Table II: The summary of the implementation of the five classifiers .

CLASSIFIERS	DESCRIPTION
	use 1000 iteration to build decision tree
Random Forests K-Nearest Neighbor	used linear search algorithm and using cross-validation to select the best k value
Multi Layer Perceptron	Training time = 500, validation threshold = 20, learning rate = 0.3 and batch size = 100
Support Vector Machine	used kernel called poly kernel
Simple Logistic Regression	Random seed = 1 is used for cross validation.

E. Performance Evaluation

Confusion matrix was used for evaluation in Weka data mining tools. Confusion Matrix (CM) is the machine learning classifier that measure the performance of the Emotional datasets. It use a table to analysis the performance of the different classifiers on each of the datasets. CM visualizes the accuracy of a classifier by comparing the actual and predicted classes of the emotional state. Measure F-Measure, Precision, Recall, ROC curve to validity of the classifier model and provide kappa statistics and Mean Error Rate (MAE). The following are the list of rates that are often computed from a confusion matrix for the five classifiers:

- **Accuracy** is how close a measured value is to the actual (true) value. Accuracy retrieves the percentage of correctly classified instances. The formula for accuracy is defined by:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

- **Precision** Precision tells us how many of the correctly predicted emotional classes actually turned out to be positive. The formula for precision is depicted by:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- **Recall** tells us how many of the actual positive emotional classes we were able to predict correctly with our model. The formula for recall is denoted by:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- **F-Measure** it signifies a joined classification performance of both precision and recall. F-Measure is given by:

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{4}$$

- **Kappa Statistics** measure the agreement between two emotional classes, Kappa statistics formula defined by:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \tag{5}$$

Where Where P (A) = percentage agreement and P (E) is the chance agreement. Kappa of 0 is doing better than chance hence kappa of 1 is a perfect agreement. Where TP = True Positive, FP = False Positive, TN = True Negative and FN = False Negative.

IV. RESULTS AND DISCUSSION

This section elaborate the results of the prediction model for emotional state and compared our results with the related work done in speech emotion recognition analysis. Two datasets were used for classification by SLR, RF, K-NN, SVM and MLP to predict which emotions are being incorrectly or correctly classified. This section discuss the accuracy and performance of the five classifiers in order to determine which model achieved highest accuracy. During implementations, we trained and validated our classifiers using 10-fold cross validation. The experiment were done using Weka data mining for training the classifiers.

Table III summaries the accuracy, precision, recall, f-measure, kappa statistics and MAE derived from two datasets, namely SAVEE and RAVDESS. Generally, from the SAVEE datasets we observed that RF performed better than all the classifiers with the highest accuracy of 96.04%, followed by MLP with 88.54%, K-NN with 80.00%,SLR with 77.50%and SVM had lowest accuracy of 75.42%. Furthermore, Random Forests yield highest precision of 0.96, recall and f-measure of 0.96, and kappa agreement of 0.95. MLP had an accuracy of 88.54%, K-NN of 80.00% and SLR had accuracy of 77.50% lower than Random Forests at 96.04%, but the mean error of MLP, K-NN, and SLR was better with 0.046, 0.06, and 0.09 respectively than that of Random Forests with 0.010 and SVM with 0.21. Accordingly to the experiment performed on RAVDESS datasets, We observed that RF had the highest accuracy of 66.04%, followed by K-NN with accuracy of 65.79%, MLP with 62.71%, SMV with 56.39% and SLR had the lowest accuracy of 56.04%. Altogether RF algorithm performed way better than all the other classifiers with the highest precision of 0.67, recall of 0.66, and kappa statistics of 0.61. We observed that K-NN accuracy is lower than RF at 65.79%, but the f-measure, MAE of K-NN was better with 0.66 and 0.09 respectively than RF with 0.66. In this section we will present the results of the classification algorithms.

The results presented in Table III display that in overall the RF algorithm has the highest accuracy rate for each datasets. We used the confusion matrix to further recognize our emotional classes. A confusion matrix contains data about actual and predicted classifications for classification algorithms. From the above experiments done on WEKA data mining

Table III: The summary of the different classifiers showing accuracy, precision, recall, f-measure, kappa statistics and MAE: SAVEE and RAVDESS datasets

Classifier	**SAVEE Datasets**					
	Accuracy	***Precision***	***Recall***	***F-Measure***	***Kappa Statistics***	***Mean Absolute Error***
MLP	88.54%	0.88	0.89	0.88	0.86	0.05
SVM	75.42%	0.76	0.75	0.75	0.71	0.21
Random Forests	96.04%	0.96	0.96	0.96	0.95	0.10
Simple Logistics regression	77.50%	0.77	0.78	0.77	0.73	0.09
K-NN	80.00%	0.80	0.80	0.80	0.76	0.06
Classifier	**RAVDESS Datasets**					
	Accuracy	***Precision***	***Recall***	***F-Measure***	***Kappa Statistics***	***Mean Absolute Error***
MLP	62.71%	0.63	0.63	0.63	0.57	0.10
SVM	56.39%	0.58	0.56	0.57	0.50	0.20
Random Forests	66.04%	0.67	0.66	0.66	0.61	0.16
Simple Logistics regression	56.04%	0.56	0.56	0.56	0.49	0.14
K-NN	65.80%	0.66	0.66	0.66	0.61	0.09

Table IV: Table of Confusion Matrix generated using SAVEE datasets

Predicted Classes for RF

	a	b	c	d	e	f	g
angry	58	1	0	1	0	0	0
disgust	0	52	5	0	2	1	0
fear	0	2	54	3	0	0	1
happy	2	0	3	55	0	0	0
neutral	0	0	0	0	120	0	0
sadness	0	1	0	0	1	58	0
surprise	0	0	0	1	0	0	59

Table V: Table of Confusion Matrix generated using RAVDESS datasets

Predicted Classes for SVM

	a	b	c	d	e	f	g	h
neutral	41	22	5	18	0	0	5	5
calm	13	144	0	17	0	4	12	2
happy	8	7	103	19	14	17	7	17
sad	11	37	15	91	0	16	12	10
angry	1	3	21	2	116	5	31	13
fearful	1	8	15	19	3	122	6	18
disgust	3	19	11	22	4	4	121	8

tool, RF came out as the best model to predict emotional for each datasets and SVM yield lowest accuracy based on RAVDESS; therefore we used the confusion matrix of RF and SVM to see how best it can predict each of the emotion classes. We decided to select the classifier that yield the highest accuracy. Because accordingly to the implementation performed on the SAVEE emotional datasets, RF achieved an highest accuracy of 96.04% than SVM, MLP, SLR and K-NN. Based on research study conducted by Gaurav [11] shown that RF performed better than SVM, K-NN, MLP and SLR, that yield accuracy of 97.00% for emotional prediction using Deap datasets. RF is one of the learning algorithm that produces highly accurate and perform better than other classifiers. This make RF the best fit for emotional analysis. The experiment done on the RAVDESS datasets shows that RF is also the highest algorithm that performed better than the other four classifiers. Study done by Muhammad et al [12] for evaluating the performance of RF with respect to emotion classification achieved accuracy of 64.02% as the second best classifier compared to K-NN, SLR and SVM. This good performance is prevalent even on our study. RF performed better than all the other because it work best when there are large number of training datasets with less features. In our study, trained the five classifiers using 480 instances(SAVEE datasets) and 1440 instances for RAVDESS emotion speech separately together 100 features. Accordingly to our results, we establish that K-NN provide good results than SVM and SLR with accuracy of 65.79%. The reason is that K-NN perform better in dataset with low dimensional space and when the data set was not easily separable using the decision planes. This proves that RAVDESS emotion datasets used in this study contain low dimensional and classifier had some difficulties in identifying which emotion class belong to certain classes.

Table IV shows the confusion matrix results of two the classifiers, RF from SAVEE and SVM train using RAVDESS datasets. The classifier RF recognized neutral with high priority, and very clear without being confused with other classes. RF manage to recognize happy, fear, disgust emotions were recognized with lower priority.To see how well each classifiers performed, We calculated the percentages of the confusion matrix for RF and SVM. We achieved this by dividing the total number of correctly classified instances for class ‘a’ with the total number of all instances. We then multiplied this number by 100 to come up with a percentage. We repeated the same steps for classes’ b, c, d, e, f and g. For SAVEE datasets a = angry, b = disgust, c=fear, d=happy, e=neutral, f=sadness and g=surprise. RF classifier neutral seems to have the highest percentage of correctly classified instances by 100%, followed by surprise with 98.33%, angry and sadness with 96.66%, happy with 91.66% and fear were predicted by 90.00%. The

emotional with the lowest correctly classified was disgust with 86.67%. Table IV represent the confusion matrix of the classifiers that performed better and one that yield the lowest accuracy. SVM yield the lowest accuracy, because manage to correctly classified Instances of 830 and Incorrectly Classified Instances of 610 with calm correctly classified by 75.00%, followed by fearful with 63.54%, disgust with 63.02%, angry with 62.42%, happy with 53.65%. The emotional with the lowest correctly classified was sad with 47.40% and neutral by 42.71%.

V. CONCLUSION AND RECOMMENDATION

Our research study was inspired by a research gap specified by A. Iqbal [13] which stated that these a need for further study for considering more emotions such as fear, surprise and happy in order to recognize a real-time emotion recognition from speech. This study make use of seven emotional state such as angry, happy, fear, sadness, surprise, neutral and calm. The main objective of this research study makes the use of a supervised machine learning algorithm to automatically learn a model that will be used to identify which Emotional state affect the attitude of the speakers in order to improve the learning environment. Our aim was to develop machines that can detect user's (speakers) emotions and express different kinds of emotion in the learning environment. The aim for analysing emotion is to understand how tone of the speaker affect the learning environment. The results obtained can be used to monitor and evaluate the teaching platform by providing feedback and watching for signs of confusion, curiosity, and enjoyment of the speaker, which can lead to greater success of improving the learning environment. The main objective was achieved by following the steps: data collection, feature extraction, labelling of data, training and testing the model. The classifiers was evaluated using cross-validation.

This paper addresses the ASR of emotional problem by investigating the attitudes of the speakers in education by identifying which emotional state affect the learning environment by checking the accuracy of the models. The proposed study compare two datasets of emotional state. There are four main steps in the presented approach: first feature extraction is applied, then labelling the emotional state carried out, feature selection and finally classification is employed. In this paper, five classification models were addressed for speech emotion recognition task. We first extracted different feature sets for each dataset using jAudio and trained, feature selection and analysed its performance on identifying the emotion of the speech using the five model. Our experimental results reveal that using SAVEE datasets gives us the best prediction accuracy for our Emotional recognition model. Then data set used to train the classifiers consists of attributes with category values only and with a considerably larger amount of data.

Results depends how the datasets are in terms of robustness, how the speakers alter the speech, background of the recording such as noisy environment. This are the main problem faced by ASR system. Our study make a conclusion that using SAVEE emotion recognition datasets for understanding how tones of the speakers affect learning environment produces better emotional recognition model. Give us the full credit to make a conclusion that RF classifier is the best, and accurate model that can be used in education to identify the attitudes of the speakers through their tones. This kind of the research study can help in education for providing corrections and feedback based using real-time speech recording.

For future work, this study only uses two emotional datasets that are speech audio only. This provides limitation because audio speech does not give much information about the face expression of the speakers, only expressed the attitudes of the tones. Using emotional datasets that contain video and audio would help in improving the accuracy and performance of the different models. Our results shown different predictive models that are trained using two datasets, SAVEE and RAVDESS datasets. For future purpose, believe that using more than two emotion datasets could help in improving the performance and provides the better classifiers that manage to classify the emotion analysis. In this study, feature selection was done using information Gain. For future work, we consider applying different feature selection methods to the different datasets and compare their accuracy. This study used two alter aggregator, mean and standard deviation for feature extraction using jAudio. Future work includes the computation of more features representation such as MFCC aggregation, the feature histogram and Area moments on both the SAVEE and RAVDESS datasets for aiming to capture additional features of emotions and comparing the accurate of the aggregators added. It would be also interesting to ask human listeners to annotate the datasets and then compare the classifier confusion matrices to the human ones. This paper uses two data sets to determine the effect of the tone in the learning environment by using the ASR and check which classifier is giving the best result. we can see RF outperformed. However, author/s may kind to another classifier. We are sure if SVM parameters tuning may place then SVM can be a competitor of RF.

VI. ACKNOWLEDGEMENT

This work is based on the research supported in part by the National Research foundation of South Africa (Grant number: 121835). I would like to take this opportunity to thank DSI-NICIS National e-Science Postgraduate Teaching and Training Platform for giving me the opportunity to sponsor my masters degree.

REFERENCES

[1] R. Sylwester, "How emotions affect learning," vol. 52, no. 2, 1994, pp. 66–65.

[2] L. Choukiker, A. Chaudhary, A. Sharma, and J. Dalal, "Speech emotion recognition," *Journal of emerging technologies and innovative research*, 2015.

[3] A. Tickle, S. Raghu, and M. Elshaw, "Emotional recognition from the speech signal for a virtual education agent," *Journal of Physics: Conference Series*, vol. 450, p. 012053, jun 2013. [Online]. Available: https://doi.org/10.1088%2F1742-6596%2F450%2F1%2F012053

[4] T. S. D. Kami´nska and A. Pelikant, "Review of emotion recognition from speech," *Review of Scientific Instruments*, vol. 4, no. 6, pp. 2–4, 9 2013. [Online]. Available: https://www.researchgate.net/publication/266968747

[5] D. Datcu and L. J. M. Rothkrantz, "Emotion recognition using bimodal data fusion," in *Proceedings of the 12th International Conference on Computer Systems and Technologies*, ser. CompSysTech '11. New York, NY, USA: Association for Computing Machinery, 2011, p. 122–128. [Online]. Available: https://doi.org/10.1145/2023607.2023629

[6] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Information Fusion*, vol. 49, pp. 69 – 78, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517307066

[7] F. A. R. S. R. Livingstone, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english." *PLoS ONE*.

[8] R. D. Peacocke and D. H. Graf, "An introduction to speech and speaker recognition," in *Readings in Human–Computer Interaction*, ser. Interactive Technologies, R. M. BAECKER, J. GRUDIN, W. A. BUXTON, and S. GREENBERG, Eds. Morgan Kaufmann, 1995, pp. 546 – 553. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780080515748500571

[9] T. B. Mokgonyane, T. J. Sefara, T. I. Modipa, M. M. Mogale, M. J. Manamela, and P. J. Manamela, "Automatic speaker recognition system based on machine learning algorithms," in *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019, pp. 141–146.

[10] R. Ajoodha, R. Klein, and B. Rosman, "Single-labelled music genre classification using content-based features," in *2015 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, 2015, pp. 66–71.

[11] D. Kamińska, T. Sapiński, and A. Pelikant, "Review of emotion recognition from speech," 09 2013.

[12] M. Zubair Asghar, F. Subhan, M. Imran, F. Masud Kundi, S. Shamshirband, A. Mosavi, P. Csiba, and A. R. Varkonyi-Koczy, "Performance Evaluation of Supervised Machine Learning Techniques for Efficient Detection of Emotions from Online Content," *arXiv e-prints*, p. arXiv:1908.01587, Aug. 2019.

[13] A. Iqbal and K. Barua, "A real-time emotion recognition from speech using gradient boosting," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2019, pp. 1–5.