Speech Emotion Recognition using Support Vector Machine (SVM), Decision Tree, Linear Discriminant Analysis (LDA)

Rahul Yerramsetti
Texas A&M University, Corpus Christi
rahulyerramsetti@islander.tamucc.edu

***Abstract*— Speech is the most popular method of communication. Recognizing human emotions from speech signals became a challenging research topic known as Speech Emotion Recognition (SER).Capturing the emotion from only speech is a difficult challenge. The proper selection of features in relation with both the time and frequency domains together is necessary to produce optimized results. In this paper, we concentrate on recognizing four emotional states: Happy, Sad, Anger and Neutral from speech. For this we explore features like Energy, Pitch, Zero-crossing rate and Mel-Frequency Cepstrum Coefficients (MFCC). Three different classification algorithms Support vector machine (SVMs), Decision Tree, Linear Discriminant Analysis (LDA) are compared for recognizing emotions from speech. A Poland Corpus (Database of Polish Emotional Speech) is used for training and testing the classifiers used.**

Keywords - Support Vector Machine; Decision Tree; Linear Discriminant Analysis; Speech Emotion Recognition;

I. Introduction

Speech is an important way of communication. People express their emotions either through speech or actions. In-order to make robots and softwares interactive with their users, Emotion recognition from speech is very important.

Speech Emotion Recognition is a very active research topic in the Human Computer Interaction (HCI) field and has a wide range of applications. For distance learning, identifying students' emotion timely and making appropriate treatment can enhance the quality of teaching. In automatic remote call center, it is used to timely detect customer's dissatisfaction. It is also used to aid clinical diagnosis or to play video games. The research of automatic speech emotion recognition, not only can promote the further development of computer technology, but also will greatly enhance the efficiency of people's work and study, and help people solve their problems more efficiently. It will also further enrich our lives and improve the quality of life.

In recent years, a great deal of research has been done to recognize human emotion using speech information. Many speech databases have been built for speech emotion research, such as Polish Emotional Dataset [5], BDES (Berlin Database of Emotional Speech) that is German Corpus and established by Department of acoustic technology of Berlin Technical University, DES (Danish Emotional Speech) [7] that is Danish Corpus and established by Aalborg University, Denmark , SES (Spanish Emotional Speech) that is Spanish Corpus etc.

In this paper we used Polish Emotional Dataset and 3 different classifiers (support vector machine, linear discriminant analysis and decision tree) to classify emotions from speech.

The paper is organized as follows. Section II describes the Background of Research, Section III discusses about the Research Statement. Section VI describes the Technical Approach used for this system. Section V discusses about the Results obtained, section VI concludes the paper and Section VII gives detail on Further Research.

II. Background

In this paper we want to work on features like MFFC and compare the accuracy results from three different algorithms.

Many researchers have proposed important speech features which contain emotion information, such as energy, pitch frequency [7], formant frequency [8], Mel-Frequency Cepstrum Coefficients (MFCC) [6] and its first derivative [3]. Several classification techniques have been used by many researchers such as Neural Networks (NN) [2], Gaussian Mixture Model (GMM), Hidden Markov model (HMM) [4], Maximum Likelihood Bayesian classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN) and Support vector machines (SVM) [1].

In this paper, we use the Polish emotional database to train and test our automatic speech emotion recognition system. We concentrate on obtaining features based on both the time (Energy, Zero-crossing rate etc.) and frequency (Pitch, Mel-frequency cepstral coefficients, etc.) domains together(which improves accuracy) and use this to train and test on our set of classifiers for results. Instead of using a single classifier we are using three different classifiers (Support vector machine, Decision Tree, Linear Discriminant Analysis) so as to compare the results on a particular type of emotion state and use this data to make a comparative study so that we can draw conclusions and further improve the system for accuracy.

III. Research Statement

Speech Emotion recognition involves recognition of emotion from speech signal. This can improve the communication between humans and robots. Applications of emotion classification based on speech have already been used to facilitate interactions in our daily lives. For example, call centers apply emotion classification to prioritize impatient

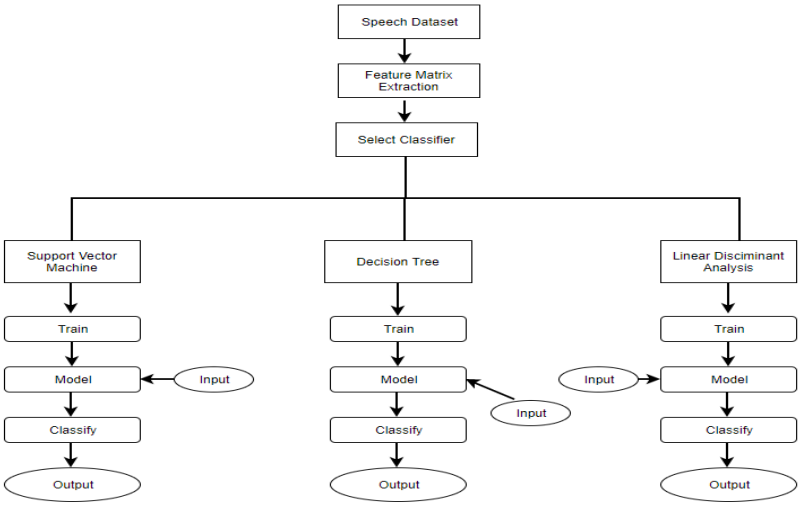

Fig 1: Workflow Diagram

customers and a warning system can be been developed to detect if a driver exhibits anger or aggressive emotions.

Capturing the emotion from only speech is a difficult challenge. The selection of proper features in relation with both the time and frequency domains together is one major challenge in this process. To overcome this problem we concentrate on extracting 4 different features namely Zero-crossing rate (time-domain), Energy (time-domain), Pitch (frequency-domain) and Mel-Frequency Cepstrum Coefficients - MFCC [6] (frequency-domain). Another major challenge in this project is to understand the stress in the speech. We are going to extract a feature called Zero-crossing rate to solve this problem. A good dataset is required to get proper results. For this we are using the Polish emotional database.

IV. Technical Approach

The technical approach involves the details about the dataset chosen, features being extracted and the classifiers being used.

A. *Workflow:*

This section gives an overlay of how the system functions as depicted in the above Fig 1. First an appropriate dataset is selected, it is processed to extract the desired features (MFCC, Energy, Pitch, Zero-crossing rate) from it, and a Matrix of the extracted features is built. In the next step we select the required classifier amongst the three (SVM, LDA, Decision tree) to train a model out of it. In the next step the input data is given to the system to predict its emotion. The input data is evaluated with the model and the emotion is predicted which is the output. The process is repeated to test it on another classifier.

The Generated results are analyzed against each of the algorithms used upon each emotion in particular for accuracy with the help of a cost matrix.

B. *Dataset*

The database used in this paper is Polish emotional speech database [5], which is a simulated speech database. It is an open source speech database and easy to access, and it is frequently used in fields of speech emotion recognition. This database contains six basic emotions: joy, boredom, fear, anger, sadness and neutral. All of the speech samples are simulated by 8 professional actors (4 females and 4 males). Each uttered 5 different sentences. There are totally about 240 speech samples in this database. Speech was recorded through a condenser microphone and stored at a sampling rate of 44,1 kHz, 16 bits. The length of the speech samples varies from 2 seconds to 8 seconds. Out of these speech samples we did our Speech Emotion Recognition on 4 different emotions:anger, sadness, joy and neutral to achieve good accuracy results. A total of 140 records (35*4) of speech samples are used to train the models of different classifiers. 20 speech samples (5*4) are used to test the models generated.

C. *Features*

In this paper we use four features namely Energy (time-domain), Pitch (frequency-domain), Mel-Frequency Cepstrum Coefficients - MFCC [6] (frequency-domain) and Zero-crossing rate(time-domain) to understand the emotion in speech. Proper selection of features from time domain in appropriate combination with frequency domain is required for extraction of emotion from the speech files. To solve this we have identified a frequency domain feature called the Mel-Frequency Cepstrum Coefficients (MFCC) which helps in improving the accuracy of the system. This is widely used in speech recognition and speech emotion recognition studies, and it obtains a good recognition rate. MFCC is based on the characteristics of the human ear's hearing, which uses a nonlinear frequency unit to simulate the human auditory system. Mel frequency scale is the most widely used

feature of the speech, with simple calculation, good ability of the distinction, anti-noise and other advantages. MFCC in the low frequency region has a good frequency resolution, and the robustness to noise is also very good, but the high frequency coefficient of accuracy is not satisfactory. So we give up the high-level order of the MFCC and use only low-level order as audio feature parameters.

The zero-crossing rate is the rate of sign-changes along a signal, i.e., the rate at which the signal changes from positive to negative or back. This feature has been used heavily in both speech recognition and music information retrieval, being a key feature to classify percussive sounds. A total of 10 MFCC features, Energy, Pitch, Zero-crossing rate features are extracted from the Polish audio speech dataset to build this system.

D. Data Preprocessing

Various Data Preprocessing techniques were applied on the dataset to increase the accuracy of the system on all the classifiers. Normalization on the values of all features have been applied to obtain the values between 1 and 0. In this the original tuple value is subtracted from the Minimum value among the feature list and is divided with the value that is obtained by subtracting the maximum value from the minimum value of the particular feature list.

Dimensional Reduction on the number and type of features that are to be used for classification, had been applied to identify those unique features that give us acceptable results with decent computational load.

E. Classifiers

Classifiers or Learning classifier systems, or LCS, are a paradigm of rule-based machine learning methods that combine a discovery component (e.g. typically a genetic algorithm) with a learning component (performing either supervised learning, reinforcement learning, or unsupervised learning).For this system, the classifier algorithms based on Supervised learning are used. The following describes about the classifiers used:

1) Support Vector Machine (SVM): A Support Vector Machine (SVM) is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples.

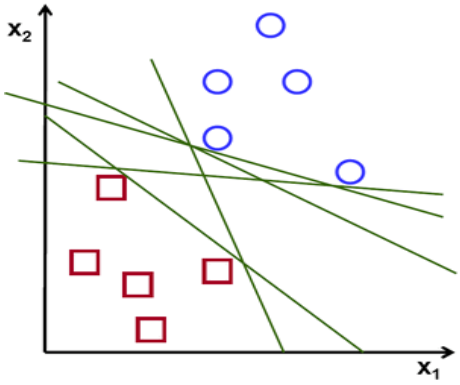

Fig 2: Initial Seperating Lines

To define an optimal hyperplane let us consider the problem : For a linearly separable set of 2D-points which belong to one of two classes, find a separating straight line.

In the above picture we can see that there exists multiple lines that offer a solution to the problem. But, we need to know which line is the best line among them to separate.

A line is bad if it passes too close to the points because it will be noise sensitive and it will not generalize correctly. Therefore, we should find the line passing as far as possible from all points.

Therefore, the operation of the SVM algorithm is based on finding the hyperplane that gives the largest minimum distance to the training examples. Twice, this distance receives the important name of margin within SVMs theory. Hence, the optimal separating hyperplane is the line that maximizes the margin of the training data.

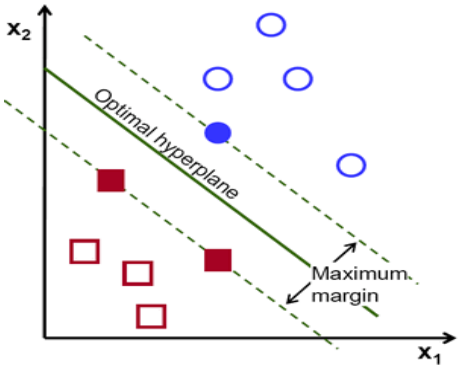

Fig 3: SVM Hyperplane

2) Decision Tree: Decision trees are classification trees that divide the feature space into several regions, and in each region, if a category of samples is dominant, they are marked with the category labels. A decision tree is a tree whose internal nodes are tests and whose leaf nodes are categories. Each internal node is responsible for testing one attribute and each branch from the node selects one value for the attribute. The leaf node predicts a specific class. The decision trees are not limited to boolean functions, but multiple categories.

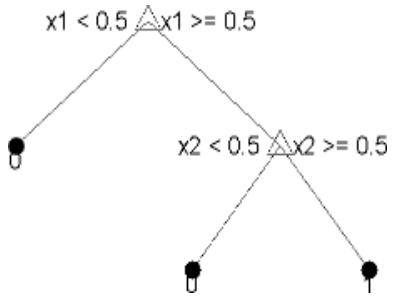

Fig 4: A simple Decision tree

Decision tree learning uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the

target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

3) Linear Discriminant Analysis: Linear discriminant analysis (LDA) is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

LDA takes multi-dimensional data, makes use of prior class information (Supervised Learning) and represents the data in a form which maximizes the distance between different classes.

It basically takes covariance of a class (of data) with itself, mean of the entire data, mean of each class, prior probabilities of the class. LDA also uses scatter within a class, scatter in between classes and tries to best separate 2 classes of data.

The resulting combination is used as a linear classifier, or, more commonly, for dimensionality reduction.

4) Multi Class Classification: Multi classification is the problem of classifying into three or more classes unlike two classes in binary classification.

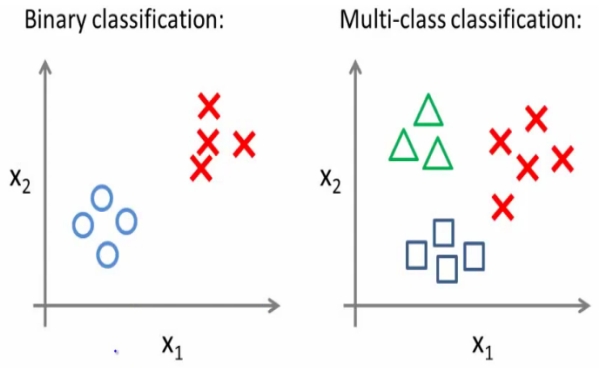

Fig 5: Binary vs Multiclass Classificaton

The problem of Multi class classification can be transformed into binary classification by using two methods

- One vs Rest
- One vs One

In One vs Rest strategy a single classifier per class is trained as shown in Fig 6. The samples of that class are treated as positives and the others are treated as negative. In this strategy the base classifiers are required to produce a real valued confidence score for its decision.

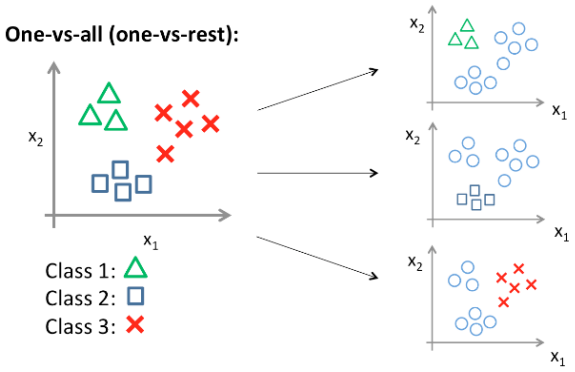

Fig 6: One vs All

In One vs One strategy the samples from a pair of classes from the original training set are taken and are distinguished. At prediction time voting is applied and the one that gets the highest number of votes will be predicted by the combined classifier.

We have used One vs One strategy in all our classifiers as all of them deal with multi class classification.

V. Results

This section describes about the results obtained for the emotion classification using different classifiers on the polish dataset. We finally make comparisons based on the accuracies of each emotion against each classifier used with the help of their confusion matrices.

A. Support Vector Machine:

The results obtained on Support Vector Machine are good enough. The Training accuracy obtained using SVM is 72.85 percent. 102 emotions were classified accurately out of 140 during re-substitution. We used a Multiclass SVM to make classification among the 4 emotions and 13 features.

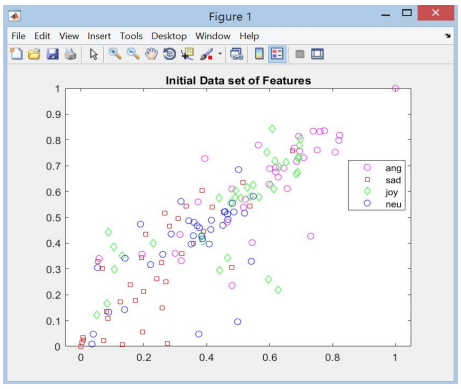

Fig 7: Initial Dataset of Features

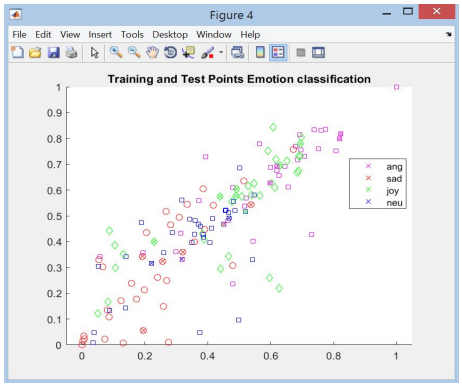

Fig 8: Classified Testpoints on Trained Model

Fig 7 shows the initial 2 features of the dataset. Testing accuracy of 70 percent is obtained on the test dataset of 20 records. 14 out of 20 emotions were recognized correctly. Fig 8 shows the Test Data points (represented by X) plotted on a graph of trained data points showing their classification. Fig 9 shows the regions of various emotions classified by the SVM classifier.

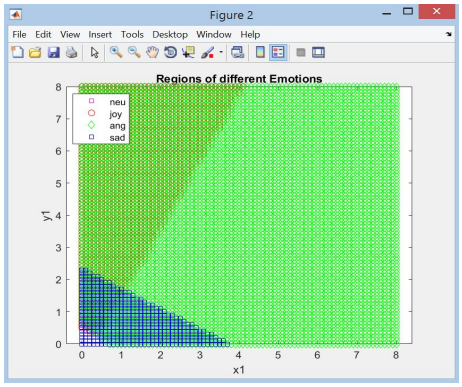

Fig 9: Regions of Emotions

B. Decision Tree:

The results obtained using Decision Tree classifier are very good and best among the three classifiers used. The Training accuracy obtained using SVM is 92.14 percent. 129 emotions were classified accurately out of 140 during re-substitution. We used a Multiclass Decision Tree to make classification among the 4 emotions and 13 features.

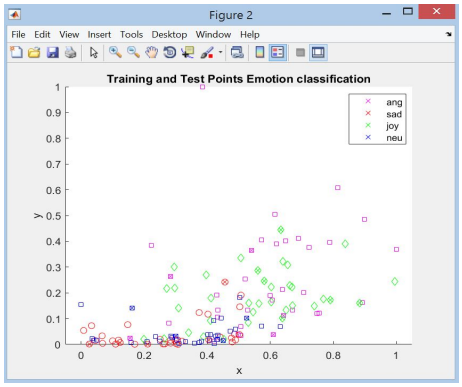

Fig 10: Initial Dataset of Features

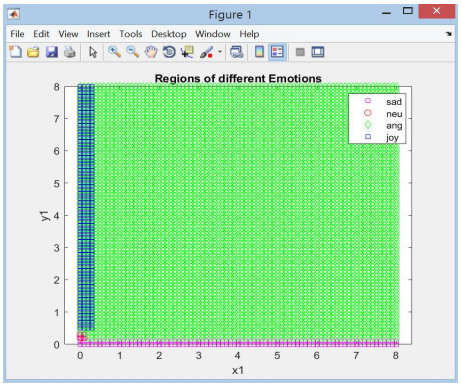

Fig 11: Regions of Emotions

A testing accuracy of 85 percent is obtained on the test dataset of 20 records. 17 out of 20 emotions were recognized correctly. Fig 10 shows the Test Data points (represented by X) plotted on a graph of trained data points showing their classification. Fig 11 shows the regions of various emotions classified by the Decision Tree classifier.

Cross-validation is a technique used for evaluating models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Crossvalidation of data is used to detect overfitting(failing to generalise a problem). The error is computed as kfoldLoss of the partitions, which for our model is of the value 0.4214. The resubstitution loss is the loss calculated between the response training data and the model's predicted response values based on the input training data. It is returned as a scalar value and the value obtained is 0.0786.

For decision trees, the cross validation error is larger than the resubstitution error which shows overfitting. The best solution out is to find a simpler and less complex tree and this can be done by pruning the original tree. Pruning reduces the size of decision trees by removing sections of the tree that provide little power to classify instances. Pruning reduces the complexity of the final classifier, and hence improves predictive accuracy by the reduction of overfitting. This helps reduce the cross validation upto a certain point. The simplest tree with the least cross validation error is chosen as the best decision tree. The final value obtained after pruning the decision tree is 0.3786. This can be seen in the following fig 12.

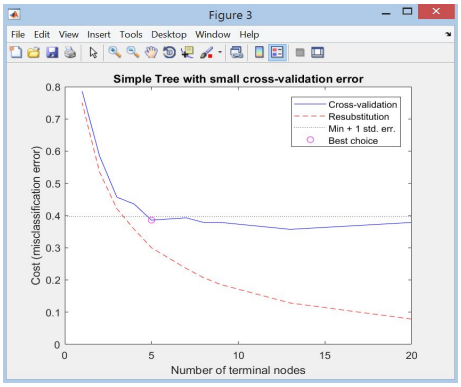

Fig 12: Cross Validation Error

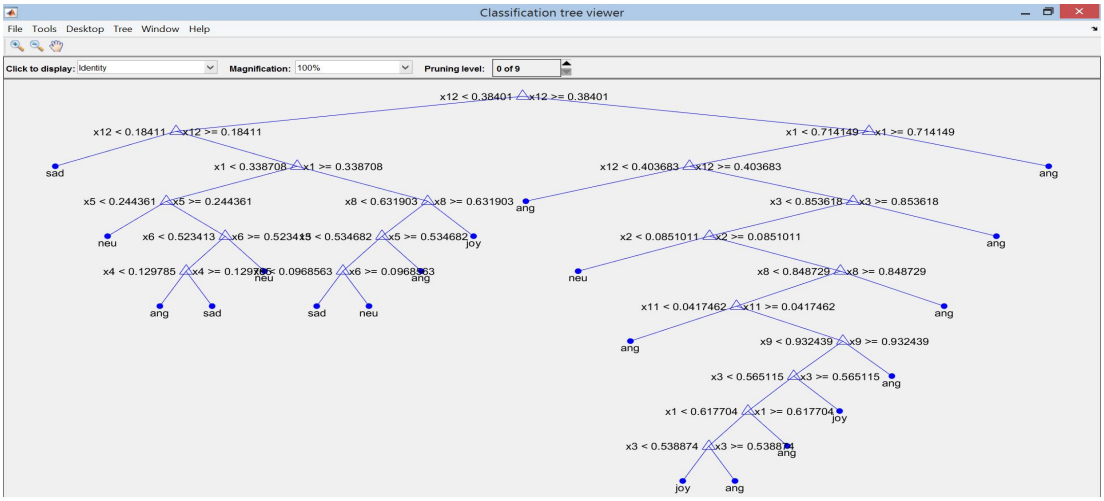

Fig 13: Complete Classification Tree

The above Fig 13 shows the complete classification tree and Fig 14 shows a simple Decision Tree classifier for classiftion of emotions.

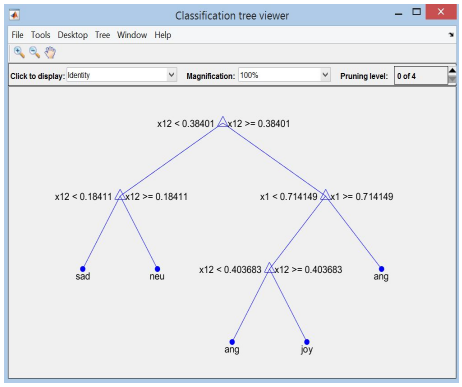

Fig 14: Simple Classification Tree

C. *Linear Discriminant analysis:*

The results obtained using Linear Discriminant analysis classifier are close enough to SVM. The Training accuracy obtained using SVM is 72.14 percent. 101 emotions were classified accurately out of 140 during re-substitution. We used a Multiclass LDA classifier to make classification among the 4 emotions and 13 features.

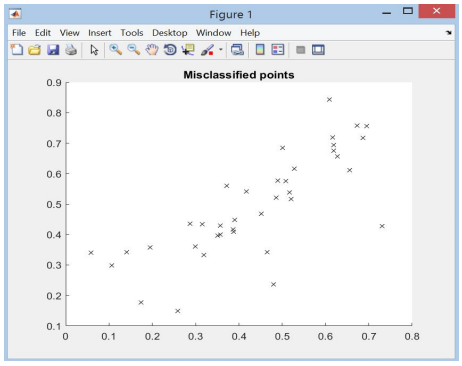

Fig 15: Misclassified Points

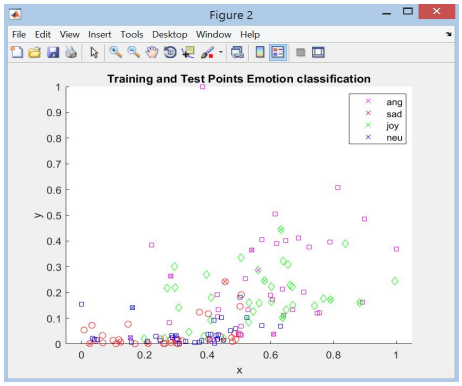

Fig 16: Classified Testpoints on Trained Model

A Testing accuracy of 60 percent is obtained on the test dataset of 20 records. 12 out of 20 emotions were recognized correctly. Fig 16 shows the Test Data points (represented by X) plotted on a graph of trained data points showing their classification. Fig 17 shows the regions of various emotions classified by the LDA classifier. The obtained LDA Cross-validation Error is 0.4571.

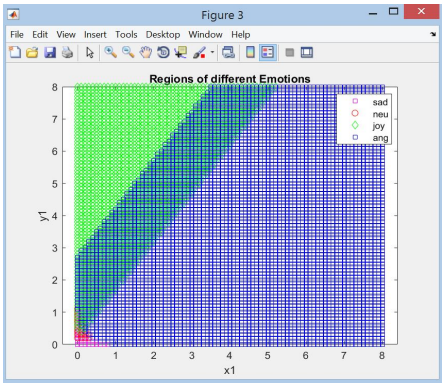

Fig 17: Regions of Emotions

D. Comparison:

The confusion matrices of the three classifiers are compared to understand the efficiency of the classifiers on each emotion.

Fig 18: Confussion Matrix

The confusion matrices in Fig 18 represent the classification of emotions: Anger, Sad, Joy, Neutral respectively. The Diagonal represents the number of features that were classified properly and the rest represent the misclassified emotions. Using this we can understand the individual efficiencies of the classifiers against a particular class in the dataset. Here we observe that SVM classifier could classify 21 emotions out of 35 as angry, 30 emotions out of 35 as Sad, 27 emotions out of 35 as Joy and 24 emotions out of 35 as Neutral. The LDA classifier could classify 20 emotions out of 35 as angry, 29 emotions out of 35 as Sad, 27 emotions out of 35 as Joy and 26 emotions out of 35 as Neutral. Where as Decision Tree classifier could classify 33 emotions out of 35 as angry, 31 emotions out of 35 as Sad, 33 emotions out of 35 as Joy and 32 emotions out of 35 as Neutral.

VI. Conclusion

Therefore we can conclude that the performance of decision tree classifier was best with the chosen polish dataset (35*4 = 160 records) for speech emotion recognition. The performance of SVM was better than the LDA classifier. But however the SVM classifier could give better results if the size of the dataset would increase.

VII. Further Research

In future this project could be improved for accuracy by applying more data preprocessing techniques and dimensional reduction of the features.

References

[1] Yashpalsing Chavhan, M. L. Dhore, and Pallavi Yesaware. Speech emotion recognition using support vector machine.

[2] X. Mao, L. Chen, and L. Fu. Multi-level speech emotion recognition based on hmm and ann. In *2009 WRI World Congress on Computer Science and Information Engineering*, volume 7, pages 225–229, March 2009.

[3] Tsang-Long Pao, Yu-Te Chen, Jun-Heng Yeh, and Pei-Jia Li. Mandarin emotional speech recognition based on svm and nn. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 1, pages 1096–1100, 2006.

[4] B. Schuller, G. Rigoll, and M. Lang. Hidden markov model-based speech emotion recognition. In *Multimedia and Expo, 2003. ICME '03. Proceedings. 2003 International Conference on*, volume 1, pages I–401–4 vol.1, July 2003.

[5] Piotr Staroniewicz and Wojciech Majewski. Cross-modal analysis of speech, gestures, gaze and facial expressions. chapter Polish Emotional Speech Database — Recording and Preliminary Validation, pages 42–49. Springer-Verlag, Berlin, Heidelberg, 2009.

[6] Vibha Tiwari. Mfcc and its applications in speaker recognition. 1, 01 2010.

[7] D. Ververidis, C. Kotropoulos, and I. Pitas. Automatic emotional speech classification. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–593–6 vol.1, May 2004.

[8] Zhongzhe Xiao, E. Dellandrea, Weibei Dou, and Liming Chen. Features extraction and selection for emotional speech classification. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2005.*, pages 411–416, Sept 2005.