Emotion Recognition On Speech Signals Using Machine Learning

Mohan Ghai, Shamit Lal, Shivam Duggal and Shrey Manik
Delhi Technological University
mohanghai13@gmail.com, shamitlal@yahoo.com, shivamduggal.9507@gmail.com,
shrey.manik@gmail.com

Abstract - **With the increase in man to machine interaction, speech analysis has become an integral part in reducing the gap between physical and digital world. An important subfield within this domain is the recognition of emotion in speech signals, which was traditionally studied in linguistics and psychology. Speech emotion recognition is a field having diverse applications. The prime objective of this paper is to recognize emotions in speech and classify them in 7 emotion output classes namely, anger, boredom, disgust, anxiety, happiness, sadness and neutral. The proposed approach is based upon the Mel Frequency Cepstral coefficients (MFCC) and energy of the speech signals as feature inputs and uses Berlin database of emotional speech. The features extracted from speech are converted into a feature vector, which in turn are used to train different classification algorithms namely, Support Vector Machine (SVM) , Random Decision Forest and Gradient Boosting. Random forest was found to have the highest accuracy and predicted correct emotion 81.05% of the time.**

Keywords - **Berlin database, Emotion recognition, Gradient Boosting, MFCC, SVM, Random Forest.**

Introduction

Emotions play a vital role in human communication. In order to extend its role towards the human-machine interaction, it is desirable for the computers to have some built-in abilities for recognizing the different emotional states of the user [2,5]. With the advent of technology in the recent years, more intelligent interaction between humans and machines is desired. We are still far from naturally interacting with the machines. Research studies have provided evidence that human emotions influence the decision making process to a certain extent [1-4]. Hence, it is desirable for machines to have the ability to detect emotions in speech signals. This is a challenging task which is drawing attention recently. Deciding which features to use to accomplish this task successfully is still an open question.

There can be many perceived emotions for a single utterance, each representing some part of the utterance. It is difficult to predict a single emotion as the boundary between these parts is hard to determine. Another issue is that the type of emotion generally depends upon the speaker, environment and culture. If a person is in a particular emotional state for a long time like depression, all other emotions will be temporary and the outcome of the emotion recognition system will be either the long term emotion or the short term (temporary) emotion. Presence of noise may also present some challenges in feature extraction.

A vital part of emotion detection system pipeline is the selection of appropriate features to classify speech segments in different emotional classes. For this task, hidden parameters like energy of signal, pitch, timbre, MFCC are preferred over the words and content of the speech itself. As these parameters vary continuously with time, audio is divided into frames of appropriate size. It is assumed that above-mentioned parameters don't change significantly over the duration of a frame. Thus the entire audio can be segmented in frames, each represented by a feature vector. These frames then can be used as training set for classification algorithms.

In the next section, the previous related work on speech emotion recognition systems is explained. Subsequent, Section 3 provides an insight on the database that is used for implementing the system. Section 4 provides the framework along with the approach used for feature extraction. In section 5, the algorithms proposed, Random Decision Forest,SVM and Gradient Boosting, are described in detail. Results obtained are mentioned in Section 6 and the next Section 7 concludes the paper

Berlin Database Of Emotional Speech

This database is easily and publically available and is among the most popular emotional databases used for emotion recognition. It comprises of 535 emotional

978-1-5090-6399-4/17/$31.00 ©2017 IEEE

utterances recorded from 5 female and 5 male actors (total of 10 speakers) spoken in German language [11]. The Berlin database consist of 7 emotions (anger, boredom, disgust, anxiety/fear, happiness, sadness and neutral). Each file has 16-bit PCM, mono channels, sampled at 16Khz. Total length of the 535 emotional utterances is 1487 seconds and average utterance length is around 2.77 seconds [9].

Related Work

Recognition of emotions in audio signals has been a field of extensive study in the past. Previous work in this area included use of various classifiers like SVM, Neural Networks, Bayes Classifier etc. The number of emotions classified varied from study to study, they play an important aspect in evaluating the accuracy of the different classifiers. Reduction in the number of emotions used for recognition has generated more accurate results as depicted below.
The following table summarises the previous study done on the topic.

TABLE 1
Related Work

Study	Algorithm Used	# of EmotionS	Accuracy (%)
[Kamran Soltani, Raja Noor Ainon,2007] [1]	Two layer Neural Network	6	77.1
[Li Wern Chew, Kah Phooi Seng, Li-Minn Ang, Vish Ramakonar, 2011] [2]	PCA, LDA and RBF	6 (divided into three independent classes)	81.67
[Taner Danisman, Adil Alpkocak, 2008] [3]	SVM	4/5	77.5/66.8
[Lugger and Yang, 2007] [4]	Bayes Classifier	6	74.4
[Yixiong Pan, Peipei Shen, Liping Shen, 2012] [5]	SVM	3	95.1

Speech Emotion Recognition Framework

The main aim of the system is to recognise the emotional state of the human from his or her voice. The system extracts the best features from the audio signals such as energy and MFCC features, and summarizes them into a limited number of features, over which supervised learning algorithm classifier is used [2,11]. This classifier correlates features to emotion attached with the audio.

The system consists of five basic blocks : Emotional Speech Database, Extraction of Features, Feature Vector, Classifiers along with Emotion Output as depicted in figure 1.

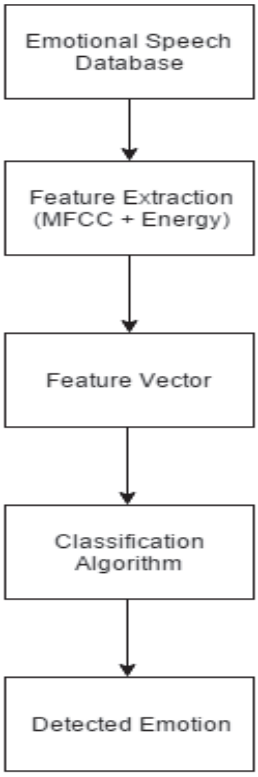

Figure 1. Speech Emotion Recognition Framework

A. Feature Extraction

Feature extraction process started by dividing an audio file into frames [19]. This was done by sampling the audio signal at 16000 Hz and selecting 0.025 sec as the duration for each frame. On the basis of above two parameters frame size was computed.

Large part of the audio signals have frequency elements only up to 8 kHz. According to the Sampling theorem 16 kHz is chosen as the ideal frequency for sampling.

It was assumed that on short time durations, the audio signal remains constant. This is why signals were framed into 20-40 ms frames. If the frame chosen is much smaller in duration then it doesn't have enough samples to get an appropriate spectral estimate, if it is larger in duration then the signal changes vividly through the entire frame.

For each audio sample an output vector was maintained which is a 7 element vector. Each emotion was given a label between 0-6. According to the audio label the

corresponding vector index was marked 1. All the other elements were marked as 0. Feature matrix was then generated by further breaking down frames into segments with each segment having 25 frames. Segment hop size was selected as 13 which gave the number of overlapping frames per segment. After segmentation, each segment was appended to the feature matrix to prepare the final feature matrix.

Most of the past researches in this field focussed on using single frames as data points for training their algorithms. Using this approach, some of the frames can be wrongly classified as belonging to a particular class when the entire audio belongs to some other class. This motivated us to combine 25 frames in segments and use these segments as data points. The benefits of this are two-fold. First, it leads to an increase of features for each data point. Second, it leads to an averaging effect where the classification of entire segment depends on the features extracted from all its constituents frames and a single frame cannot dictate the classification.

1) Mel Frequency Cepstral Coefficients(MFCC): MFCC represents parts of the human speech production and perception. MFCC depicts the logarithmic perception of loudness and pitch of human auditory system [1,2,5] . It also eliminates characteristics that are speaker dependent by removing the elementary frequency along with their harmonics.

The procedure for evaluating MFCC features [1,2] :

1. For each frame an estimate of periodogram of the power spectrum is made.
2. Add the energy in each filter after applying mel frequency filter.
3. Reduce each energy to logarithm of the corresponding energy.
4. Evaluate DCT of the filterbank energies obtained in step 3.
5. DCT coefficients from 1 to 13 are taken into consideration.

These 13 features represent the 13 MFCC.

A frame vector was maintained with each frame to store the features of the frame. For every audio frame 13 MFCC features were generated.

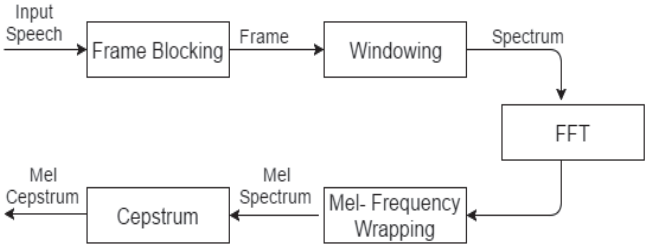

Figure 2. MFCC Feature Extraction

The Mel scale gives relationship between the actual measured frequency with the perceived frequency of a pure tone. Humans can differentiate slight transition in pitch at lower frequencies more appropriately in comparison to higher frequencies. The equation for changing frequency(f) to Mel scale is:

$$\text{Mel}(F) = 2595 \times \log10(1 + f/700) \quad (1)$$

2) Energy: The energy of the speech is extracted from the intensity of the speech. The energy associated with speech is continuously varying, therefore, energy associated with a short range of speech is of prime interest [19]. Energy can be voiced, unvoiced and silence classification of speech. Energy level for each frame is calculated by squaring the amplitudes of each frame and then summing up those values. Combining energy and MFCC features gives a clear insight to the emotion of the speech.

B. Classification Algorithms

1) SVM: SVM is a supervised learning algorithm used most widely for pattern recognition applications. The algorithm is simple to use and provides good results even when trained on limited size training dataset. More formally, it is an algorithm which constructs, in a high dimensional or infinite dimensional space, a hyperplane or set of hyperplanes which can be used for regression and classification tasks. It tries to learn a hyperplane that results in maximum separation between the data points belonging to different classes, leading to a better classifier. The data can be linearly separable or non-linearly separable. Non linearly separable data is classified by mapping it's feature space to a high dimensional space using a kernel function. The data points are then linearly separable in this high dimensional space. The optimization problem in SVM reduces to

$$a = \min(\frac{||w||^2}{2}) \text{ subject to } \forall k,\ y_k(< w.x_k > +b) \geq 1 \quad (2)$$

Where w is normal vector to the hyperplane and $< w.x_k >$ denote the inner product of w and x_k for each data point (x_k , y_k) in the training set.

Some of the widely used kernels are linear kernel and radial bias kernel.

Linear kernel function is given as:

$$kernel(x,y) = (x,y) \quad (3)$$

Radial bias kernel is given as:

$$kernel(x,y) = e^{\frac{-||x-y||^2}{(2\sigma^2)}} \quad (4)$$

In this paper, SVM is implemented using random bias kernel. The implementation is based on libsvm. The fit complexity of the algorithm increases quadratically with the number of samples it is trained on.

2) Gradient Boosting: Gradient boosting is used to produce a prediction model in the form of ensemble of weak prediction models. It works in a stage-wise manner and generalises the different boosting methods by optimising the technique for minimizing arbitrary differentiable loss functions by adding, at each stage, a new tree that best reduces (steps down the gradient of) the chosen loss function.

The basic idea of the algorithm is to develop new base-learners that are maximally connected with the loss function specifically it's negative gradient, related with the total group. The loss functions applied can be arbitrary chosen, in this paper deviance loss function is used for the Gradient Boost Classifier. After the initial fitting by the basic classifier, additional copies of the classifier are fitted on the same initial dataset but the weights of the samples that have been incorrectly classified are adjusted in order to make successive classifiers focus on more complex cases.

In this paper, we trained gradient boosting algorithm using deviance loss. Learning rate was set to 0.15. Number of boosting stages performed were 120. Maximum depth of the regression estimator was set to 4. These hyperparameters selection resulted in the highest accuracy on test set.

3) Random Forests: Random Forests uses an ensemble of learning methods and is used for regression, classification and other tasks. Random Forests works by constructing a large number of decision trees at the time of training and it outputs the mean prediction or mode of the class of the individual trees. Random decision forests prevents decision trees from overfitting the training data.

Random Forests algorithm works as follows:

1. Random Record Selection: Each decision tree is trained on approximately 2/3rd of the training data.
2. Random Variable Selection: Only some of the variables used for prediction are chosen. This selection is done randomly and node is split according to the most optimal split of the node.
3. Trees grow to maximum extent without any pruning.

Random Forests algorithm is based on bagging. In bagging, a random sample is selected with replacements repeatedly from the training set and trees are then fit to these samples: By averaging the prediction values or by taking the majority vote from all the decision trees we predict the values of the unseen sample. Gini importance determines whether parent node needs to be split into children nodes. If reducing the parent into children reduces the gini impurity or increases the intensity , the node is split.

$$I = G_{parent} - G_{split1} - G_{split2} \quad (5)$$

$$G = \sum p_i(1-p_i) \quad (6)$$

This summation is run over all the classes and p_i is the probability of class i. G stands for gini importance and I is the intensity.

In this paper, the random forests classifier was implemented using 15 forests. Gini impurity was used to measure the quality of the split of a node. Testing for different max depths, we decided to split nodes until each leaf had samples belonging to a single class only. Using these hyperparameters highest accuracy of 81.05% was achieved.

RESULTS

Three classification algorithms, namely Random Decision Forest, SVM and Gradient Boosting classified an audio signal into one of the 7 classes. Out of the three, Random Decision Forest achieved the highest accuracy of 81.05%. Considering that many past papers achieved similar or less accuracy when trained on less than 7 emotional output classes, we consider our accuracy as a significant improvement. For all the three algorithms, we achieved highest classification accuracy for samples belonging to anger class while least accuracy was achieved for those belonging to happiness class. The results for each algorithm is summarized in the following tables and graphs.

TABLE 2

SVM

Emotion	Emotion Recognised (%)						
	A	B	D	An	H	S	N
A	86.10	1.85	2.29	3.62	3.97	0.44	1.72
B	4.62	51.04	5.55	4.35	1.20	18.86	14.38
D	16.29	14.15	44.00	7.51	4.00	7.90	6.15
An	30.66	6.13	6.70	37.83	4.34	7.83	6.51
H	57.99	4.33	6.37	8.42	15.89	1.97	5.04
S	1.90	13.80	2.52	2.27	0.06	73.87	5.58
N	9.55	26.63	4.27	6.06	2.25	10.87	40.37
Accuracy						55.89%	

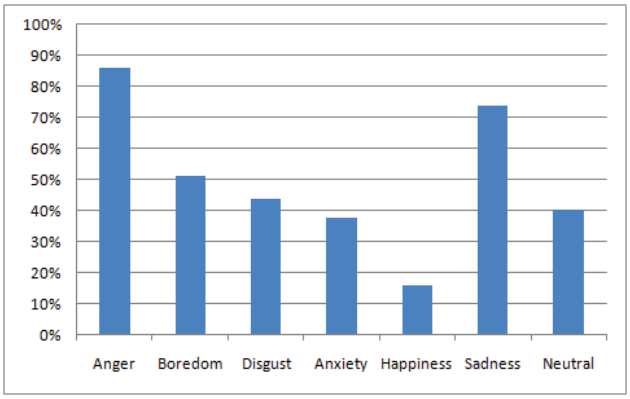

Figure 3. Graph showing the accuracy results for SVM for different emotions

TABLE 3

GRADIENT BOOSTING

Emotion	Emotion Recognised (%)						
	A	B	D	An	H	S	N
A	88.53	1.50	1.32	2.69	3.18	0.88	1.90
B	6.69	66.29	2.41	2.07	0.67	12.17	9.70
D	15.12	8.20	55.80	5.56	2.93	6.24	6.15
An	28.49	6.51	4.81	47.74	2.45	7.17	2.83
H	43.12	6.14	2.44	2.75	40.36	2.05	3.15
S	2.94	6.38	1.60	1.72	0.12	82.76	4.48
N	16.61	13.74	2.56	6.99	3.03	8.15	48.91
Accuracy						65.23%	

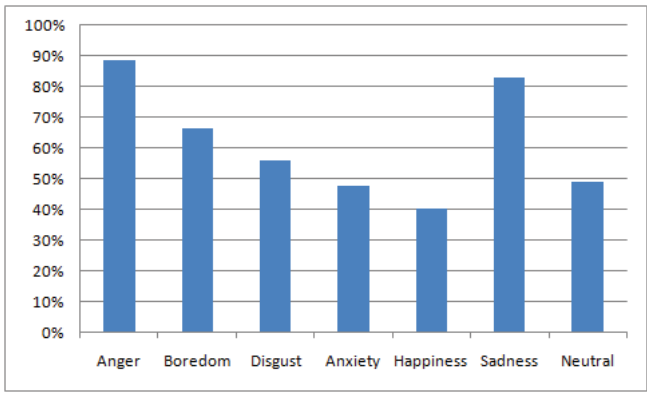

Figure 4. Graph showing the accuracy results for Gradient Boosting for different emotions

TABLE 4

RANDOM FORESTS

Emotion	Emotion Recognised (%)						
	A	B	D	An	H	S	N
A	93.38	0.54	0.67	1.34	2.91	0.18	0.98
B	4.20	87.65	1.05	1.31	0.99	1.58	3.22
D	7.18	3.74	83.91	1.25	1.15	1.15	1.63
An	19.66	2.98	0.79	70.90	2.18	2.18	1.29
H	27.21	4.22	0.65	3.57	61.66	0.57	2.11
S	1.96	3.80	0.52	0.63	0.40	91.83	0.86
N	11.51	8.25	2.14	2.54	4.13	2.22	69.21
Accuracy						81.05%	

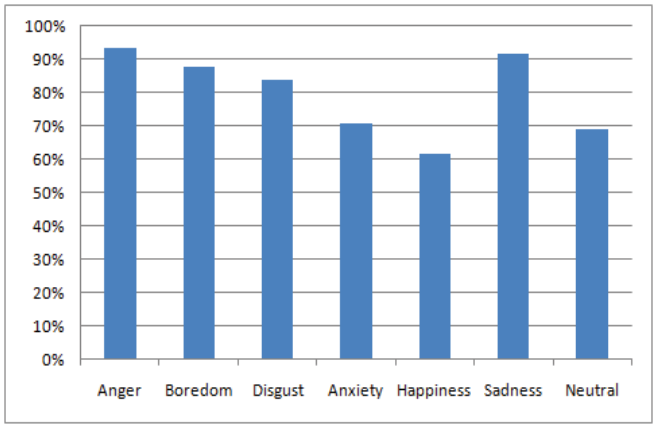

Figure 5. Graph showing the accuracy results for Random Forest for different emotions

N - Neutral, **S -** Sadness, **H -** Happiness, **An -** Anxiety, **D -** Disgust, **B -** Boredom, **A -** Anger.

Conclusion

Emotion recognition in speech signals has become potentially a major topic for research in the field of interaction between humans and computers due to its wide degree of applications in the recent times.

In this paper, an approach to emotion recognition in audio signals based upon Random Decision Forest, SVM and Gradient Boosting classifiers was presented. The performance of the classifiers can boost up significantly if suitable features are properly obtained. Hence, we concluded that inclusion of energy as a feature along with other 13 MFCC features led to better assessment of the emotion attached with the speech. It can be seen that integrating frames into overlapping segments led to a greater continuity in samples and also resulted in each data point having many more features. Also treating each segment as an independent data point increased the size training set many folds leading to an increase in accuracy when using different classification algorithms.

Considering the different classification strategies the maximum accuracy of 81.05 % is obtained for the database by using Random Decision Forest classifier.

Future Work

The paper presents only the analyses of seven human emotions using speech signals. It can be expanded to predict more human emotions. Also more tuples can be collected and better feature engineering can be applied in the future to enhance the result of the prediction algorithms. The classification algorithms wrongly predicted some of the samples belonging to happiness class as belonging to anger class. This can be rectified by extracting more features to better distinguish between these two class.

References

[1] K.V .Krishna Kishore, P.Krishna Satish, "Emotion Recognition in Speech Using MFCC and Wavelet Features", *3rd IEEE International Advance Computing Conference (IACC)*, 2013.

[2] Yixiong Pan, Peipei Shen and Liping Shen, "Speech Emotion Recognition Using Support Vector Machine*", International Journal of Smart Home, 2012*

[3] Ashish B. Ingale and Dr.D.S.Chaudhari,, "Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine", *International Journal of Advanced Engineering Research and Studies, Vol. 1,Issue 3*, 2012.

[4] Davood Gharavian, Mansour Sheikhan, Alireza Nazerieh, Sahar Garoucy, "Speech emotion recognition using FCBF feature selection method and GA-optimized fuzzy ARTMAP neural network", *Neural Computing and Applications*, Volume 21, Issue 8, pp 2115–2126, 2011

[5] Li Wern Chew, Kah Phooi Seng, Li-Minn Ang, Vish Ramakonar, Amalan Gnanasegaran, "Audio-Emotion Recognition System using Parallel Classifiers and Audio Feature Analyzer", *Third International Conference on Computational Intelligence, Modelling & Simulation,* 2011.

[6] Zeng, M. Pantic, G. Roisman, T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 31, no. 1, pp 39-58, Jan. 2009.

[7] J. S. Park, Ji-H. Kim and Yung-H. Oh, "Feature Vector Classification based Speech Emotion Recognition for Service Robots" *IEEE Transactions on Consumer Electronics*, vol. 55, no. 3, pp. 1590-1596,Aug. 2009.

[8] Xia Mao, Lijiang Chen, Liqin Fu "Multi-Level Speech Emotion Recognition based on HMM and ANN" in *World Congress on Computer Science and Information Engineering*. 2009.

[9] Taner Danisman, Adil Alpkocak, "Emotion Classification of Audio Signals Using Ensemble of Support Vector Machines", *Perception in Multimodal Dialogue Systems Volume 5078 of the series Lecture Notes in Computer Science pp 205-216, Springer-Verlag Berlin Heidelberg*, 2008.

[10] S. Casale, A. Russo, G. Scebba, S. Serrano, "Speech Emotion Classification using Machine Learning Algorithms", *The IEEE International Conference on Semantic Computing,*2008.

[11] Kamran Soltani, Raja Noor Ainon, "SPEECH EMOTION DETECTION BASED ON NEURAL NETWORKS", *IEEE International Symposium on Signal Processing and Its Applications, ISSPA* 2007.

[12] Lugger, M., Yang, B.: An Incremental Analysis of Different Feature Groups In Speaker Independent Emotion Recognition. *In: 16th Int. Congress of Phonetic Sciences ,* 2007.

[13] S.Kim, P.Georgiou, S.Lee, S.Narayanan. "Real-time emotion detection system using speech: Multi-modal fusion of different timescale features", Proceedings of *IEEE Multimedia Signal Processing Workshop, Chania, Greece,* 2007.

[14] Eun Ho Kim, Kyung Hak Hyun, Soo Hyun Kim and Yoon Keun Kwak "Speech Emotion Recognition Using Eigen-FFT in Clean and Noisy Environments" in16th *IEEE International Conference on Robot & Human Interactive Communication,,* August 26, 2007, Korea.

[15] Chul Min Lee, and Shrikanth S. Narayanan, "Toward detecting emotions in spoken dialogs", *IEEE Transaction on Speech and Audio Processing,,* vol. 13, no. 2, pp. 293- 303,Mar. 2005.

[16] T. Vogt and E. Andre. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. *In IEEE, editor, Int'l Conf. Multimedia and Expo*, pages 474–477, Jul 2005.

[17] M. Song, J. Bu, C. Chen, and N. Li, "Audio- Visual Based emotion recognition: A new Approach," in Proc. *IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 1020– 1025. 2004.

[18] http://www.expressive-speech.net/, Berlin emotional speech database.

[19] Nwe, T. L., Wei, F. S., & De Silva, L.C., " Speech based emotion classification", *TENCON 2001 Proceedings of IEEE Region 10 International Conference on Electrical and Electronic Technology*, Vol. 1. pp. 297-301), IEEE.