PHYSICAL FEATURES BASED SPEECH EMOTION RECOGNITION USING PREDICTIVE CLASSIFICATION

Mohammad Ahsan[1] and Madhu Kumari[2]

Department of Computer Science and Engineering, NIT Hamirpur, H.P., India

ABSTRACT

In the era of data explosion, speech emotion plays crucial commercial significance. Emotion recognition in speech encompasses a gamut of techniques starting from mechanical recording of audio signal to complex modeling of extracted patterns. Most challenging part of this research purview is to classify the emotion of the speech purely based on the physical characteristics of the audio signal independent of language of speech. This paper focuses on the predictive modeling of audio speech data based on most viable feature set extraction and deployment of these features to predict the emotion of unknown speech data. We have used two most widely used classifiers, a variant of CART and Naïve Bayes, to model the dynamics of interplay of crucial features like Root Mean Square (RMS), Zero Cross Rate (ZCR), Pitch and Brightness of audio signal to determine the emotion of speech. In order to carry out comparative analysis of the proposed classifiers, a set of experiments on real speech data is conducted. Results clearly indicate that decision tree based classifier works well on accuracy whereas Naïve Bayes works fairly well on generality.

KEYWORDS

Acoustic features, audio emotion recognition, speech emotions and predictive classifier.

1. INTRODUCTION

The human speech consists mainly of two types of information, first one is ‘what is said’ and the second one is ‘how it is said’. A lot of the work has been done till now on extraction of first kind of information but there is a lot of scope to extract the second kind of information that is how speech is delivered (emotion).

Emotion of a person does not change the content of the audio signals but it can totally change the meaning of that content. Therefore we are focusing on how to make a system able to accurately and precisely infer the meaning of the content of an audio signal. However, detecting the emotion from audio signals has many challenges. There are no clarified categories of emotions pertaining to the emotions and speech. Another critical research challenge in the emotion detection problem is to determine the features that influence the emotion in an audio signal. The precise feature extraction from audio signals for knowing the type of emotion is a very challenging task and depends strongly upon the application and database at hand.

There is considerable uncertainty in assuming a set as best feature set for classifying the audio data into a set of available emotional classes; it is even more difficult to choose the appropriate classifier (HMM or SVM or rule-based classifier). The data sets used in speech-emotion recognition are a combination of speech generated by speakers of different gender, in different speaking styles, in different languages. These different sources of speech pose complications because they directly affect the features such as pitch and energy[7].

DOI:10.5121/ijcsit.2016.8205

ech samples of human beings." In order to do this, first we have to find out which set of features would be more influential for classification and which classifier has the best performance with the selected features' set.

2. RELATED WORK

Automatic classification of speech in various emotions becomes possible due to the substantial enhancement in the field of computer science. There exist number of feature extraction methods which are currently being used by the researchers for assigning emotions to the speech samples. Emotional states experienced by the speakers get encoded in the acoustic signal and later can be decoded or seen in the form of specific patterns. In a telephonic conversation, these acoustic signals are decoded by the listener's ears to know the emotional state of the speaker. There are number of studies have been carried out so far for decoding the emotional state from the speech samples. Few of them are listed below.

Author	Features	Classification Methods	Results
Amir [1]	Pitch related, intensity related, speech rate, jitter, shimmer and duration related.	Distance measure and Neural Networks (**NN**).	DM method outperform the NN method.
Yang et al. [3]	Pitch related, Mel Frequency Cepstral Coefficients (MFCCs), Zero Cross Rate (ZCR) and energy related (total =306 features).	Bayesian Learning.	Harmony features are useful for emotion recognition, but not equal for all emotion dimensions.
Lee et al. [4]	Duration related- duration of longest voiced speech, ratio of duration of voiced and unvoiced region. Energy related- energy minimum, energy maximum, energy median, and energy range.	K-nearest neighbourhood (**k-NN**) and Linear Discriminant Classifiers (**LDC**).	Accuracy of emotion recognition get enhanced by combining the language and acoustic features.
Ververidis et al. [5]	Pitch, intensity, speech rate, and MFCCs.	Hidden Markov Models (**HMM**), Artificial Neural Networks (**ANN**), **LDA**, **k-NN** and **SVM**.	Pitch, short-term energy, MFCCs, formants, cross-section areas and Teager energy operator-based features are the most important features for emotion recognition.

Rong et al. [6]	Pitch, intensity, ZCR, and MFCCs.	Decision tree classifiers, random forest and ERFTrees algorithm (for features selection).	Best recognition accuracy 72.25% can be achieved by Random Forest classifier by using ERFTrees algorithm.
Ayadi et al. [7]	Pitch, energy, speech rate, and MFCCs.	**HMM, GMM, ANN,** and **SVM**.	1. It is hard to decide which classifier performs best for emotion recognition task. 2. Speaker-dependent classification is easier than speaker-independent classification.

Speech emotion recognition is a supervised learning problem. It means, each sample carries a correct emotion class associated with it. As emotion is a result of human experience, there exist a difference between acted and natural emotions. It becomes a daunting task to model the complexity of human behaviours. Higher the difference among the features of two emotion classes higher is the chances of correct classification. Researchers listed above used different-different features and applied varying classifiers in order to extract the emotion from a speech sample. They used features in order of hundreds and used complex classifiers i.e. HMM and SVM. In our work, we used only most contributing or influential features like RMS, pitch, ZCR [3, 5, 6, 7, 10, 11] and brightness and by using simple classifiers i.e. decision-tree and naïve-bayes achieved 81% and 74% accurate results respectively.

3. PROPOSED SCHEME

In order to classify the audio speech signals into emotion the proposed scheme uses three steps namely data pre-processing, applying classifiers and evaluating applied classifiers based on comparative performance analysis.

3.1. Data pre-processing

3.1.1. Dataset description

In earlier work [10], data was recorded from actors who were asked to read the sentences in some pre-defined emotions. That technique might be useful because we can set microphones, cameras and noise-free environment easily but even then usually that sort of data remain different from a natural data. So that would not be much more effectual in human-machine interactions.

In order to form a good dataset, data should be recorded from natural events and must be divided into pieces of small length because emotions are temporary events and in order to effectively know the encoded emotion of an audio it is recommended to divide them into pieces of small length.

Dataset which we have used in this paper contain 100 samples of 25 sound clips, each of which is 30 seconds. In order to form a dataset, we have to extract or collect a set of sound clips and assign them some emotion class. So it is also a very important phase of automatic emotion detection problem and efficiency of the classifiers undoubtedly going to be effected by the dataset in hand.

3.1.2. Features Extraction

Here we extract features from audio samples and transform them into some appropriate format for further processing. For any classification problem, the classification result would be as much better as the features are distinctive with respect to the classes to be distinguished.

In our implementation, we extracted N_f (number of features) features for each of the N_s (number of samples) samples using MIRToolbox[8]. Now these samples are assigned to one of the Ne (number of emotional classes) emotions. The result of the features extraction process is a feature-vector set with a size of N_s*N_f and an emotion-vector, containing the target emotion for N_s samples with a size of N_s*1. This result of feature extraction process is used in making a classifier and finally this classifier is used to predict the emotion of a new sample.

3.1.3. Features Selection

All of the extracted features may not be equally important for the classification problem so here in this stage only those features are selected which are more useful for the problem in regard to some prior performance criterion. Classifiers face the curse of dimensionality with large number of features so in order to protect our classifier from this problem we need to remove the unwanted or unnecessary features from the features set.

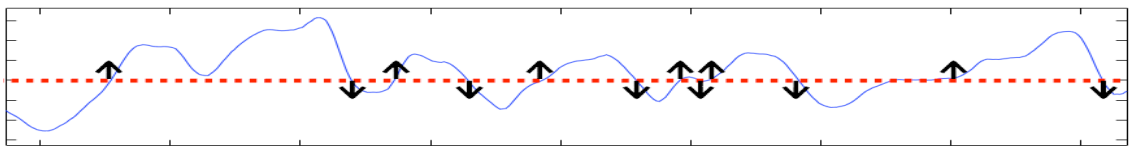
Commonly used acoustic features for the automatic emotion detection are: RMS, pitch, ZCR and brightness [1,6,9].

RMS (Root Mean Square): It represents the energy of the signal and can be calculated by taking the root average of the square of the amplitude. The global energy of the signal x:

$$x_{rms} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2} = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n}}$$

Pitch: It is the periodic time of a wave pulse generated by air compressed through the glottis from the lungs. With a change in the emotions of a person his/her biological characteristics like blood pressure and flow of air from the lungs also get changed. So by extracting this feature we can get help in knowing the emotion of a person.

ZCR (Zero Cross Rate): It is the calculation of how many times a signal is crossing its zero axis. Due to a change in the biological and psychological behavior of peoples with a change in their emotion, it also changed that how many times a signal crossed its zero axis.



Brightness: It indicates the amount of high-frequency content in a sound. Its value is high if the emotion encoded in a signal is fear and low in case of angry emotion as per our implementation results.

3.2. Applying Classifiers

Automatic emotion detection problem is a supervised learning problem because supervised learning takes a known set of input data and known responses to the data, and seeks to build a predictor model that generates reasonable predictions for the response to new data.
Number of classifiers exist for the supervised learning, few of them are- HMM (Hidden Markov Model), SVM (Support Vector Machine), Classification-Tree classification, Naive-Bayes classification *etc.*

In this paper we have chosen Classification-Tree classification and Naive-Bayes classification methods because of the inherent simplicity and generality of these models.

3.2.1. Decision-Tree Based Emotion Recognition

Decision-tree based classifier uses a binary decision tree for classification of audio emotions. We are using 'ClassificationTree' method to predict the responses for new or unseen data, this method uses labelled data for training so that it can compute the most viable emotion class.

Model Construction:

Decision tree based predictive model for above problem is represented as:

ct = ClassificationTree.fit(FeatureVectors,Emotions), where ct is a decision tree, it represents a binary classification tree based on the input variables (also known as predictors, features or attributes) FeatureVectors and output (response) Emotions, where each branching node is split based on the values of a column of FeatureVectors.

Feature Vectors:

Feature vectors are encoded in a matrix of numeric feature values, where each column of a 'FeatureVectors' represents a feature, and each row represents an observation.

Emotions:

In the proposed scheme, it represents a numeric vector, vector of categorical variables (nominal or ordinal), logical vector, character array, or cell array of strings. Each row of 'Emotions' represents the classification of the corresponding row of 'FeatureVectors'.

Algorithms used by this classifier:

There are many specific decision-tree algorithms. Notable ones include: ID3 (Iterative Dichotomiser 3), C4.5 (successor of ID3) and CART (Classification and Regression Tree). These algorithms use some metrics to decide the classification is best or not.

Metrics:

Different metrics are used by the different classifiers for measuring the "best". These metrics are applied to each candidate subset, and the resulting values are combined (e.g., averaged) to provide a measure of the quality of the split. Few of them are described below:

Gini impurity:

Used by the CART (Classification and Regression Tree) algorithm, it is a measure of how often a randomly chosen element from the set would be incorrectly labeled if it were randomly labeled according to the distribution of labels in the subset. Gini impurity can be computed by summing

the probability of each item being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

To compute Gini impurity for a set of items, suppose i takes on values in {1, 2, … , m} and let f_i be the fraction of items labeled with value i in the set.

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i=1}^{m} \left(f_i - f_i^2\right) = \sum_{i=1}^{m} f_i - \sum_{i=1}^{m} f_i^2 = 1 - \sum_{i=1}^{m} f_i^2$$

Information gain:

Used by the ID3, C4.5 and C5.0 tree-generation algorithms. It is based on the concept of entropy from information theory.

$$I_E(f) = \sum_{i=1}^{m} f_i \frac{1}{\log_2 f_i} = -\sum_{i=1}^{m} f_i \log_2 f_i$$

In the proposed work, we are using a variant of CART which uses a blend of 'Gini impurity' and 'Information gain'.

This approach has advantages like- simple to understand and interpret, requires little data preparation, able to handle both numerical and categorical data and perform well with large datasets but it also has some limitations like- it can create over-complex trees that do not generalize well from the training data (known as overfitting), locally optimal decisions are made at each node. So in order to overcome these limitations we can use other approach known as Naive-Bayes approach.

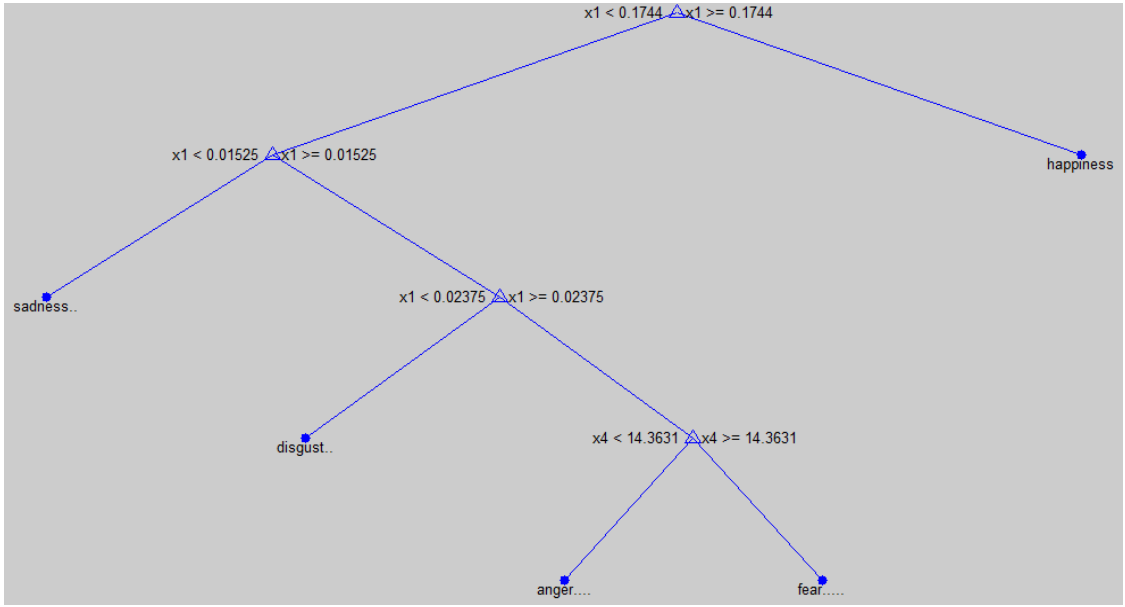


Figure 1. Classification-Tree

Once decision tree is computed from the training set, predictive method c1= ct.predict (FeatureVectors), predicts the emotion for the new data or observation by following the branches of tree until it reaches a leaf node and on reaching to a leaf node, it returns the classification of that node.

3.1.2. Naïve-Bayes Classification

The Naive Bayes classifier is used when features are independent of one another within a class, but it appears to work well in practice even when that independence assumption is not valid. For example, a fruit may be considered to be a guava if it is green, round, has pink flesh, hard seeds and about 400gm in weight. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is a guava, regardless of the presence or absence of the other features. It classifies data in two steps:

a) Training step: Using the training samples, the method estimates the parameters of a probability distribution, assuming features are conditionally independent given the class.
b) Prediction step: For any unseen test sample, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test sample according the largest posterior probability.

While the class-conditional independence between features is not true in general, research shows that this optimistic assumption works well in practice. This assumption of class independence allows the Naive Bayes classifier to better estimate the parameters required for accurate classification while using less training data than many other classifiers. This makes it particularly effective for datasets containing many predictors or features. It assigns a new observation to the most probable emotion's class.

Model construction:

Naïve-Bayes predictive model for above problem is represented as:

nb= NaiveBayes.fit(FeatureVectors, Emotions) , where 'nb' is the object of the Naïve-Bayes classifier. Parameter estimation for naïve Bayes uses the method of maximum likelihood.

Advantages:

It requires only a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because variables are assumed as independent, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

Bayes' theorem is used in modelling this classifier:

$$p(C \mid F_1,\ldots,F_n) = \frac{p(C)p(F_1,\ldots,F_n \mid C)}{p(F_1,\ldots,F_n)}.$$

Plainly the above equation is:

$posterior = \frac{prior \times likelihood}{evidence}$, denominator is constant because it contains the features' values, which are constant. So to know the posterior probability all we have to focus is the numerator of the fraction, which is equivalent to the joint probability model $p(C, F_1,\ldots,F_n)$.

Using Chain rule:

$$p(C \mid F_1,\ldots,F_n) = p(C)p(F_1,\ldots,F_n \mid C)$$
$$= p(C)p(F_1 \mid C)p(F_2 \mid C,F_1)p(F_3 \mid C,F_1,F_2)p(F_4,\ldots,F_n \mid C,F_1,F_2,F_3)$$

$$= p(C)p(F_1 \mid C)p(F_2 \mid C, F_1)...p(F_n \mid C, F_1, F_2, F_3, F_{n-1})$$

Applying assumptions of naïve Bayes, that each feature F_i is conditionally independent of every other feature F_j for j! =i.

$$p(F_2 \mid C, F_1) = p(F_2 \mid C)$$

$$p(F_3 \mid C, F_1, F_2) = p(F_3 \mid C), \text{ and so on.}$$

Thus, the above model can be expressed as:

$$\begin{aligned} p(C \mid F_1,...,F_n) \ \alpha \ & p(C, F_1,...,F_n) \\ \alpha \ & p(C)p(F_1 \mid C)p(F_2 \mid C)... \\ \alpha \ & p(C)\prod_{i=1}^{n} p(F_i \mid C) \\ = \ & \frac{p(C)\prod_{i=1}^{n} p(F_i \mid C)}{p(F_1,...,F_n)} \end{aligned}$$

The naïve Bayes classifier combines this model with a decision rule, known as maximum a posteriori. Naïve bayes classifier uses this function for classification:

$$classify(F_1,...,F_n) = \arg\max_c p(C = c)\prod_{i=1}^{n} p(F_i = f_i \mid C = c).$$

Parameter estimation:

A class prior may be calculated by two ways:

1) Assuming equiprobable classes i.e. priors=1/ (number of classes)
2) Calculating from the training set i.e. prior for a given class= (number of samples in the class)/ (total number of samples).

Once posterior probabilities are computed for the Naïve-Bayes model from the training set, predictive method, c2= nb.predict(FeatureVectors), classifies each row of data in 'FeatureVectors' into one of the emotion's classes according to the Naive-Bayes classifier 'nb', and returns the predicted class as 'c2'.

3.3. Comparing Performance of Applied Classifiers:

The performance of any classifier refers to the ability of that classifier to correctly predict the class. In order to know the complete performance of the classifier we need to analyze the confusion matrix for that classifier and by calculating true positive (tp), false positive (fp), true negative (tn) and false negative (fn) from the confusion matrix, we eventually know the performance measures.

Accuracy

Accuracy is the measure of the overall correctness of a classifier and it is calculated as the sum of the correct classifications (a classification is said to be correct if the predicted class is same as the target class) divided by the total number of classifications.

Precision

Precision is the measure of the accuracy provided that a specific class has been predicted. It is calculated as: tp / (tp+fp).

Recall or Sensitivity

Recall is a measure of the ability of a classifier to predict a certain class from a data set. It corresponds to the true positive rate and calculated as: tp / (tp+fn), where tp+fn is the total number of test examples of the considered class.

Specificity

Specificity is related to Recall and corresponds to the true-negative rate. It is calculated as: tn / (tn+fp).

Table 1. Confusion matrix for Decision-Tree based classifier

		Predicted Classes				
		Anger	Disgust	Fear	Happiness	Sadness
Target Classes	Anger	371	0	129	0	0
	Disgust	0	363	137	0	0
	Fear	17	58	410	0	15
	Happiness	43	0	0	457	0
	Sadness	0	51	14	0	435

Performance measure:

Accuracy = Sum of correct classifications / Total number of classifications

= (371+363+410+457+435)/2500
= 2036/2500
= 0.8144

Precision = avg (tp / (tp + fp))

= (371/(371+17+43)+363/(363+58+51)+410/(410+129+137+14)+457/(457)+435/(435+15))/5
= (0.8608+0.7691+0.5942+1.0000+0.9667)/5
= 4.1908/5
= 0.8833

Recall = avg (tp / (tp + fn))

= (371/(371+129)+363/(363+137)+410/(410+17+58+15)+457/(457+43)+435/(435+51+14))/5
= (0.7420+0.726+0.8200+0.9140+0.8700)/5
= 4.072/5
= 0.8500

Specificity = avg (tn / (tn + fp))

=(1940/(1940+60)+1891/(1891+109)+1720/(1720+280)+2000/(2000+0)+1985/(1985+15))/5

= (0.9700+0.9455+0.8600+1.0000+0.9925)/5
= 4.768/5
= 0.9536

Table 2. Confusion matrix for Naive-Bayes classifier

Target Classes	Predicted Classes: Anger	Disgust	Fear	Happiness	Sadness
Anger	465	0	26	9	0
Disgust	0	472	21	0	7
Fear	0	59	395	0	46
Happiness	31	0	0	469	0
Sadness	120	322	0	0	58

Performance measure:

Accuracy = Sum of correct classifications / Total number of classifications

= (465+472+395+469+58)/2500
= 1859/2500
= 0.7436

Precision = avg (tp / (tp + fp))

= (465/(465+31+120)+472/(472+59+322)+395/(395+26+21)+469/(469+9)+58/(58+7+46))/5
= (0.7548+0.5533+0.8937+0.9811+.5225)/5
= 3.7054/5
= 0.7411

Recall = avg (tp / (tp + fn))

= (465/(465+26+9)+472/(472+21+7)+395/(395+59+46)+469/(469+31)+58/(58+120+322))/5
= (0.9300+0.9440+0.7900+0.9380+0.1160)/5
= 3.718/5
= 0.7436

Specificity = avg (tn / (tn + fp))

= 1849/(1849+151)+1619/(1619+381)+1983/(1983+47)+1991/(1991+9)+1947/(1947+53))/5
= (0.9245+0.8095+0.9915+0.9955+0.9735)/5
= 4.6945/5
= 0.9389

Table 3. Comparative Results

	Accuracy	Precision	Recall (or sensitivity)	Specificity
Classification-tree	0.8144	0.8833	0.8500	0.9536
Naïve-bayes	0.7436	0.7411	0.7436	0.9389

4. CONCLUSION

Speech analysis from various perspectives has led many research issues, among these emotion or emotion recognition from audio clips have gained widespread attention due to its commercial utility. This work is an effort to understand the contribution of various physical features of speech data e.g. Root Mean Square (RMS), Zero Cross Rate (ZCR), Pitch and Brightness towards the emotion of the signal without considering the language component. In order to achieve this objective, the proposed scheme uses two most widely used classifiers, decision tree and Naïve Bayes to model predictive classifiers. Results of the experiments (Table-3) suggest that decision tree based classifier is advantageous over Naïve-Bayes classifier in terms of accuracy whereas Naïve Bayes classifier is less sensitive towards noise and more generic as compared to decision tree based classifier for audio emotion recognition. However it is yet to be seen how these classifiers will perform if we include the linguistic part of the speech in the analysis.

REFERENCES

[1] Noam Amir, (2001) "Classifying emotions in speech: A comparison of methods", in proceedings of European conference on speech communication and technology (EUROSPEECH'01), Escandinavia.

[2] Antonio Damasio, (1994) "Descartes' Error: Emotion, Reason, and the Human Brain", in U.K., Putman, London.

[3] B. Yang & M.Lugger, (2010) "Emotion recognition from speech signals using new harmony features", in Journal Signal Processing, Volume 90 Issue 5, Pages 1415-1423.

[4] Chul Min Lee & Shrikanth S. Narayanan, (2005) "Toward detecting emotions in spoken dialogs", in IEEE Transactions on Speech and Audio Processing, Volume 13, Pages 293–303.

[5] Dimitrios Ververidis & Constantine Kotropoulos, (2006) "Emotional speech recognition: Resources, features, and methods", in Speech Communication, 48(9), Pages 1162–1181.

[6] Jia Rong, Gang Li & Yi-Ping Phoebe Chen, (2009) "Acoustic Feature Selection for Automatic Emotion Recognition from Speech", in Journal of Information Processing and Management, Volume 45, Pages 315-328.

[7] Moataz EI Ayadi, Mohamed S. Kamel & Fakhri Karray, (2011) "Survey on speech emotion recognition: features, classification schemes, and databases", in Pattern Recognition, Volume 44(3), Pages 572–587.

[8] Oliver Lartillot & Petri Toiviainen, (2007) "A Matlab Toolbox for Musical Feature Extraction from Audio", in Proc. of the 10th Int. Conference on Digital Audio Effects (DAFx-07), Bordeaux, France.

[9] B. Schuller, G. Rigoll, & M. Lang, (2003) "Hidden markov model-based speech emotion recognition", in Proceedings of the 28th IEEE international conference on acoustic, speech and signal processing (ICASSP'03), Volume 2, Pages 1-4.

[10] Carlos Busso, Murtaza Bulut & Shrikanth Narayanan, (2013) "Toward Effective Automatic Recognition Systems of Emotion in Speech", in Social Emotions in Nature and Artifact, New York, USA.

[11] Kun Han, Dong Yu & Ivan Tashev, (2014) "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine", in Proceedings of Interspeech.

AUTHORS

Mohammad Ahsan was born in Kairana, India, in 1990. He received the B.Tech. degree in computer science and engineering from ITS Engineering College, Greater Noida, India, in 2012, and the M.Tech. degree in computer science and engineering from the National Institute of Technology Hamirpur, Himachal Pradesh, India, in 2015. He is currently pursuing Ph.D. in the field of Social Network Analysis from the National Institute of Technology Hamirpur, Himachal Pradesh, India.

Madhu Kumari was born in Shamli, India, in 1980. She is currently working as an assistant professor in the department of computer science engineering at National Institute of Technology Hamirpur, Himachal Pradesh, India. She got her masters and doctoral degree from Jawaharlal Nehru University, New Delhi, India. Her previous work was mostly focused on the exploration of different variants of reinforcement learning in simulated robocup soccer domain. She has worked on different aspects of computational advertising based on sponsored search auctions. Her current research is more aligned towards the deployment of machine learning techniques on application milieu like data analytics.