ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(39), DOI: 10.17485/ijst/2016/v9i39/102106, October 2016*

Emotional Analysis using Multinomial Logistic Regression

E. Poovammal*, Satyam Verma, Siddhant Sharma and Virendra Agarwal

Computer Science and Engineering Department, SRM University,
Chennai - 603203, Tamil Nadu, India;
poovammal.e@ktr.srmuniv.ac.in, satyamverma95@gmail.com,
siddhants29@gmail.com, virendraagarwal03@yahoo.com

Abstract

Background: Emotions have been widely used in psychology and behavior sciences as they are important elements of human nature. We expect a machine to behave like a human, in this digital era. Since the machines do not understand the emotional state of the speaker easily, it is not very easy to get a natural interaction between machine and man. Yet, many researchers are working and progressing in speech recognition. In this paper we aim to identify emotions present in human beings through analysis of speech signals. **Methods**: We make use of machine learning algorithm to choose the best features which effectively influence the emotional states of speech. **Findings**: We aim to identify various emotions that a human being goes through during verbal communication. **Applications**: Computation on the chosen best features of speech signals help us identifying the speaker's emotional state.

Keywords: Emotions, Logistic Regression, Machine Learning, SAVEE, Speech, Hypothesis

1. Introduction

The natural way of communication among humans happens using speech. This fact leads many of the researchers to think of speech signal as medium to work with. They want to work on such signals to improve the quality of interaction, which can happen between humans and machines. However, machines are not sufficiently intelligent to recognize human voice. Hence recognition of emotion expressed by speech signal can be used to teach the machine to recognize certain emotions in humans. This characteristic of identifying emotions can then be used in many fields which require interaction between man and machine. Emotion recognition using speech signal is a challenging field of study for many reasons.

First reason is that the features of speech which contribute for identifying or differentiating between emotions are not known. Most of the common extracted speech features such as pitch, and energy contours are directly affected by the variations introduced by the availability of different sentences, speed and style of speakers. Also, a single utterance may mean more than one perceived emotion. Different portions of the spoken utterance correspond to different emotion. Yet, identifying boundaries between these portions is not an easy task. One more challenge which is quite common is, every speaker expresses

**Author for correspondence*

his/her emotions differently depending on his or her culture and environment.

Many researchers believed that emotion can be characterized by two dimensions. They are activation and valence[1]. The amount of energy required to express a certain motion is referred to the term activation. The features such as pitch, quality of voice, timing and articulation of the speech signal are acoustic features. They are highly correlated with the emotion[2]. However, the parameter activation alone cannot distinguish among different emotions. For example, emotions such as anger and happiness correspond to activation with high value. Even though they are different emotions, both can have same activation value. In order to differentiate such emotions another parameter is considered and named as valence. While high accuracy is achieved in the process of classification between high-activation emotions and low-activation emotions, it is very difficult to get relatively good accuracy in classification between different emotions. Particularly, if the emotions take same level of activation values, differentiating the emotions is a great challenge.

When it is decided to have an automatic emotion recognizer, the challenge is to identify a set of emotions which can be recognized with high accuracy by the automatic emotion recognizer. Just like the white light is made up of seven different colors we consider emotion to be made up of seven primary emotions. Primary emotions are Anger, Disgust, Fear, Joy, Sadness, Surprise and Neutral[3]. These emotions are the most obvious and distinct emotions. In[4], similar work has been done but it is based on speech recognition of Punjabi numerals. In[5], similar work is based on speech synthesis of Sindhi numeric.

In[6] emotions such as Anger, Compassion, Fear, Disgust, Happy, Neutral, Sarcastic and Surprise are studied using IITKGP-SESC corpus. Simple Euclidian distance measure is used to achieve a classification accuracy of approximately 69% for female artist and 75% for male artist.

The basic emotions are Anger, Compassion, Fear, Happiness, Sadness and Dislike. They are investigated using a recognizer[7]. The recognizer was designed based on Discrete Hidden Markov Model and feature vector. Feature vectors are calculated based on coefficients called short time speech power coefficients of Mel frequency in their research. The accuracy achieved is 72.22% and 60% for two speakers for ungrouped classification and higher accuracies of 94.44% and 74% for grouped classification.

In this work, multinomial logistic regression method is used for identifying emotions in speech using utterances. The logistic regression method is applied on the existing database considering mean pitch, duration of speech and mean energy as features of speech signal and results are analysed.

2. Database and Tools used

We have used the Surrey Audio-Visual Expressed Emotion (SAVEE) database for emotional analysis. The database is

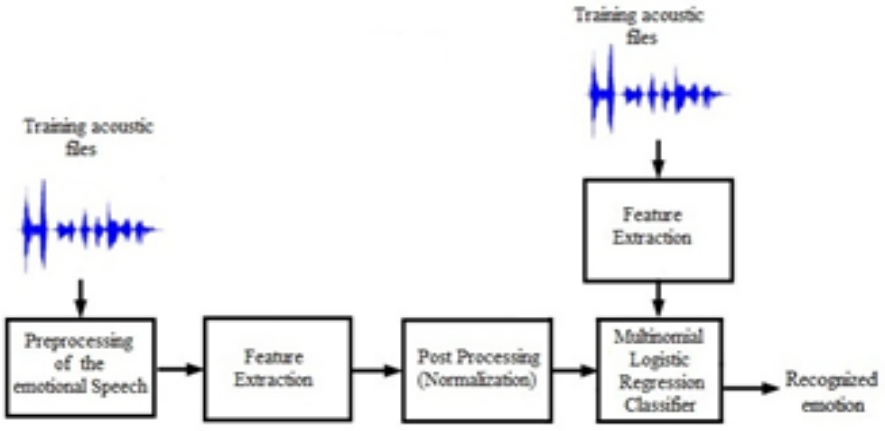

Figure 1. Emotion analysis model diagram.

recorded from four native English speakers, postgraduate students and researchers at university of Surrey aged 27 to 31 years[8]. The database considers seven types of emotions namely Anger, Disgust, Fear, Joy, Sadness, Surprise and Neutral. There are fifteen sentences for each emotion, consisting of three common, two emotion-specific and 10 generic sentences. Hence there are a total of 105 utterances per speaker. Hence the database consists of 420 utterances. Praat which is open source software has been used to construct spectrograms based on various features of the speech signals and for pitch and intensity analysis of speech signals[9]. Figure 1 shows the complete process of emotional analysis. First stage is about data processing and selecting important features. Feature Normalization ensures that data are in same range. After these preprocessing steps we use the data to train the Multinomial logistic regression classifier. Testing the classifier follows same data preprocessing steps and test data emotion is predicted.

3. Prosodic Features

Prosodic features (sometimes known as supra segmental phonology) are those aspects of speech which go beyond phonemes and deal with auditory qualities of sound. They play an important role in determining the emotion according to various researches in the area as well perceptual point of view. The prosodic features which have been considered in the study are 1. Mean pitch, 2. Duration of speech signals and 3. Mean energy. The duration of speech signal is determined in seconds. The mean pitch value for each utterance of a particular emotion is calculated using autocorrelation. The mean energy, duration of speech signal and mean pitch is calculated by using the Praat software, which is a open source software, downloaded from the link. Feature scaling[8] (Normalization) is applied to all the considered features to make sure that all instances of a feature are in the same range. Each speaker has uttered 105 sentences (7 emotions and 15 utterances per emotion) and hence each emotion in the database has 60 sentences (15 utterances by four different speakers). Figure 2 represents a 3-D scatter plot of the three features namely pitch, duration and energy for second speaker.

From Figure 2 it can be seen that Anger has the highest values for both energy and pitch while approximately Neutral emotion has the least value for pitch and energy. All the other emotions lie between these emotions. Each

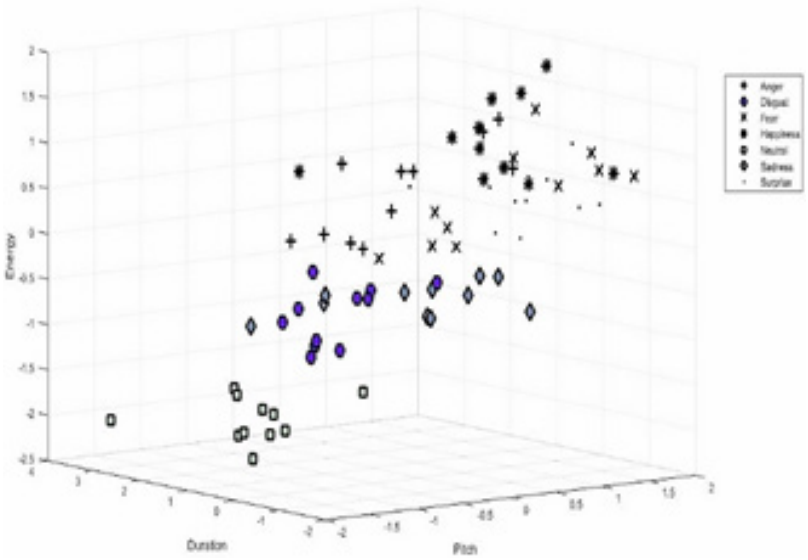

Figure 2. Scatter plot of pitch, energy and duration.

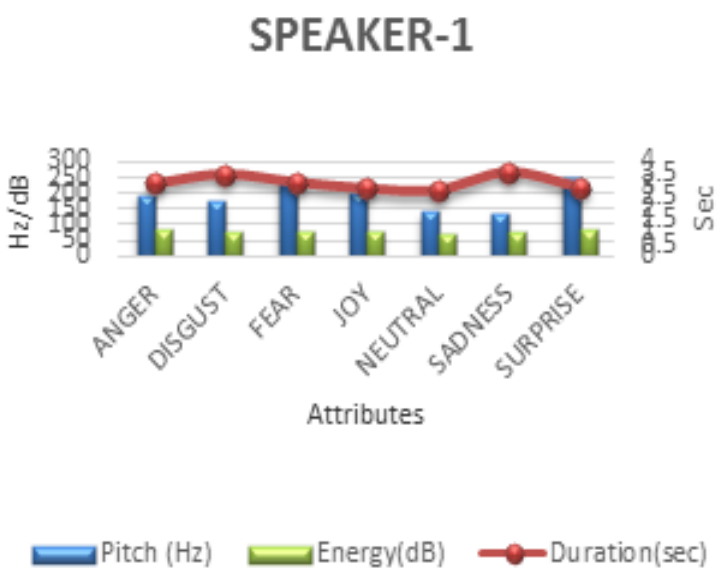

Figure 3. Emotion data of speaker-1.

Figure 4. Emotion data of speaker-2.

emotion occupies a particular area in the graph where it is more clustered and dominant than the other emotions.

Figure 3 to 6 show the mean values of all the features for each emotion for each speaker correct to two decimal places. Although this data is not used for computation but it is presented to provide a brief idea about the relative position of various emotions in Figure 2.

Figure 5. Emotion data of speaker-3.

Figure 6. Emotion data of speaker-4.

4. Emotion Evaluation

Multinomial logistic regression model has been implemented in MATLAB on the data-set[8] for predicting the classification. Emotions can be represented graphically using their features and thus a hypothesis can be derived which predicts the solution to the given classification problem and it is given by Equation (1).

$$h_{(\theta)(x)} = \theta_{(0)} + \theta_{(1)}x_{(1)} + \theta_{(2)}x_{(2)} + \cdots + \theta_{(n)}x_{(n)} \quad (1)$$

Where, $h_\theta(x)$ represents hypothesis result, $\Theta_{i's}$ represent parameters and $X_{i's}$ represents features.

The hypothesis mentioned in Equation (1) is given as input to the sigmoid function which classifies the result in different predefined categories. The sigmoid function is given by Equation (2).

$$g(z) = \frac{1}{(1 + e^{(-z)})} \quad (2)$$

The emotions predicted by the model should be close to the actual emotions. A cost function is used as intermediate result to achieve high accuracy of model which is given by Equation (3).

$$J(\Theta) = \frac{1}{m}\sum_{i=1}^{m}[-y^i \log(h_{(\Theta)}(x^{(i)})) - (1-y^{(i)})\log(1-h_{(\Theta)}(x^{(i)}))] \quad (3)$$

Where, $J(\Theta)$ represents cost of prediction, m represents number of samples, $h_\theta(x)$ represents predicted value and y represents actual value.

The cost function gives the difference between predicted and actual values. For any classification model, the value of the cost function should be in minimum. The local minimum of the cost function is found out using gradient descent algorithm which calculates Θ_i the which will minimize the cost function. Thus the values of parameters of hypothesis are found out using gradient descent algorithm which is given by Equation (4).

$$\frac{\partial J(\Theta)}{\partial \Theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\Theta(x^{(i)}) - y^{(i)})x_j^{(i)} \quad (4)$$

The extracted data from the audio signal is first normalized so that all instances of the features are in the same range. Each actor has 15 utterances for a particular emotion and a total of 105 utterances. Each actor's data is divided into 4 sets using Jack-knife procedure where, in each round three sets containing 4,4,3 instances each are used for training and one set containing 4 instances is used for testing. The result of all the 4 speakers is averaged

Table 1. Learned logistic regression parameters for each class

Emotion	$\theta_{(0)}$	$\theta_{(1)}$	$\theta_{(2)}$	$\theta_{(3)}$
ANGER	-110.4346	-76.4746	3.2006	94.9863
DISGUST	-2.4170	-1.1505	-0.5106	-0.5068
FEAR	-3.3004	2.6396	0.6431	-0.6224
JOY	-2.0978	0.0163	0.3747	0.9417
NEUTRAL	-33.4814	-11.3092	-2.4981	-20.2376
SADNESS	-2.5564	0.5960	0.4192	-1.7329
SURPRISE	-3.7922	2.0705	0.5554	1.6099

to compute the final efficiency of the given model. Let us first consider the human classification accuracies for the 7 emotion classes.

The description below shows the calculation of the second speaker. We have extracted three features for emotional analysis which are *energy, pitch and duration.* In the hypothesis equation $x_{(1)}$ corresponds to values of energy, $x_{(2)}$ corresponds to values of pitch and $x_{(3)}$ corresponds to values of duration. One v/s all classification method which is used for classification of multiple classes was implemented while training multiple regularized logistic regression classifier. This algorithm treats one class as a positive and all other classes as negative and finds parameter values for the class. As a result we have seven different values of $\theta_{(0),}\theta_{(1),}\theta_{(2),}\theta_{(3)}$which is shown in the Table 1.

Once training is done we move further to test the hypothesis, which is done through one-versus-all prediction algorithm which calculates the probability that an instance belongs to each class using the logistic model. That class will be chosen as predicted class for which the probability is highest. The result of the prediction is shown through confusion matrix. The Table 2 shows the confusion matrix on train data which has 11 instances for each emotion and Table 3 shows confusion matrix on test data which has 4 instances for each emotion.

Training set accuracy achieved on second speaker is 76.62% and test set accuracy is 82.14%.

Emotional analysis was done over the four speaker's data with 10 participants. The Mean is averaged over four actors' data. Based on standard error (n = 40) and 95% Confidence Interval (CI) the Mean was arrived. For the audio data, disgust was highly confused with neutral. The emotion fear was confused with sadness and surprise. As expected the emotions happiness and surprise were not very much distinguishable from one another[8]. The results indicate that emotions expressed by third speaker and fourth speaker were better recognized by humans than that of speaker 1 and speaker 2. From the result we can also conclude that speaker-4 is able to reflect the emotions in his utterances better than his counterparts. The results also indicate that human participants were able to recognize approximately 279 emotional utterances out of a total of 420 utterances.

Multinomial logistic regression is thus applied on the emotion data samples. For speaker-1 each emotion consists of 15 utterances. Using the jack-knife procedure each emotion class is divided into 4 sets. First three sets are used for training the model and the last set is used for testing the model. Similarly other emotion classes are also divided into 4 sets each and finally all the training and test-

Table 2. Confusion matrix of unimodal classifier on training set

Actual/ Predicted	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	11	0	0	0	0	0	0
Disgust	0	6	0	2	2	1	0
Fear	0	0	8	1	0	0	2
Happiness	0	0	0	9	0	0	2
Neutral	0	1	0	0	10	0	0
Sadness	0	3	0	0	0	8	0
Surprise	0	0	3	1	0	0	7

Table 3. Confusion matrix of unimodal classifier on test set

Actual/ Predicted	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Surprise
Anger	4	0	0	0	0	0	0
Disgust	0	3	0	0	1	0	0
Fear	0	0	2	0	0	0	2
Happiness	0	0	0	3	0	0	1
Neutral	0	0	0	0	4	0	0
Sadness	0	0	0	0	0	4	0
Surprise	0	0	1	0	0	0	3

ing sets are merged to form one data-set for a particular speaker. Thus each speaker has 77 utterances for training the model and 28 utterances for testing the model. Thus logistic regression is applied on each speaker's data-set and training and test efficiencies are calculated. The results obtained are given in the Table 4.

As mentioned in Table 4, the multinomial logistic regression model provides decent efficiency for emotion recognition. Using the jack-knife procedure we get an average efficiency of 70.31%. Speaker-2 has the highest efficiency of 82.14% while Speaker-3 has the lowest efficiency of 60.71%. Speaker-1 has an efficiency of 67.85%

Table 4. Average subject classification and model accuracy (%)

MODALITY	SUBJECT ACCURACY (%)	MODEL ACCURACY (%)
Speaker - 1	53.2	67.85
Speaker - 2	67.2	82.14
Speaker - 3	71.2	60.71
Speaker – 4	73.7	71.43
Mean	66.5 ± 2.5	70.53

Table 5. Comparison of models on SAVEE database

Model Used	Emotions Classified	Average Efficiency (%)
SVM	Anger, Disgust, Fear, Sadness, Surprise, Neutral and Happiness.	79.1
KNN	Anger, Disgust, Fear, Sadness, Surprise, Neutral and Happiness.	85
Neural Networks	Anger, Disgust, Fear, Sadness, Surprise, Neutral and Happiness.	85

while Speaker-4 gives an efficiency of 71.43%. The results obtained are comparable with those obtained by using other models like Support Vector Machine (SVM), Hidden Markov Model (HMM) and Neural Networks as shown in Table 5.

5. Conclusion

In this paper, a multinomial logistic regression approach is used for recognition of emotion in speech. Logistic regression model has been used for in this paper as no comprehensive study has been done using this model for emotional analysis. The basic emotions considered for emotion analysis are Anger, Disgust, Fear, Joy, Sadness, Surprise and Neutral. The speech signal is pre-processed using the Praat software and feature scaling is applied to bring all instances of a feature on same scale. It can be concluded that even a relatively simple method like logistic regression can be used to obtain decent efficiency in emotion recognition.

The efficiency of the system can be further increased if one takes into account other types of features like linguistic features, discourse information, or facial features. The work done at this stage is based on uni modal approach. This can be further extended to bimodal approach which can use fisher face algorithm. The bimodal approaches include facial expression at different emotion, gesture recognition and other features. The level of accuracy which can be achieved may depend on the speaker. If the speaker is not able to reflect his/her emotion in the utterance, the scope for achieving high accuracy is very less. Also further studies can be made on how the shape and size of the vocal tract and other internal organs affect a particular emotion.

We can follow the same procedure of emotion classification and extend it to help recognize the emotion expressed by animals through voice analysis. Proper database is required to be made with the help of domain expert for proper model training.

Various other classifiers like the SVM, HMM and GMM are known to provide good efficiency to identify emotions using speech. The study can be further expanded by comparing the efficiency obtained from using different types of classifiers and models preferably using real time data.

6. References

1. Fernandez R. A computational model for the automatic recognition of affect in speech [Doctoral dissertation]. Massachusetts Institute of Technology; 2004 Feb.
2. Cahn JE. The generation of a ECT in synthesized speech. Journal of the American Voice I/O Society. 1990 Jul; 8:1–9.

3. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011 Mar 31; 44(3):572–87.
4. Sharma K, Singh P. Speech recognition of Punjabi numerals using synergic HMM and DTW approach. Indian Journal of Science and Technology. 2015 Oct 16; 8(27).
5. Ratanpal BS, Sahni S. On speech synthesis of Sindhi numeric. Indian Journal of Science and Technology. 2015 Oct 18; 8(27).
6. Koolagudi SG, Maity S, Kumar VA, Chakrabarti S, Rao KS. IITKGP-SESC: Speech database for emotion analysis. International Conference on Contemporary Computing; Springer Berlin Heidelberg. 2009 Aug 17. p. 485–92.
7. Nwe TL, Foo SW, De Silva LC. Speech emotion recognition using hidden Markov models. Speech Communication. 2003 Nov 30; 41(4):603–23.
8. Surrey Audio-Visual Expressed Emotion (SAVEE) Database. Available from: http://personal.ee.surrey.ac.uk/Personal/P.Jackson/SAVEE/
9. Praat: Doing phonetics by computer. Available from: http://www.fon.hum.uva.nl/praat/