Speech Emotion Classification using Machine Learning

Pooja Yadav,
Department of CSE,
ITM University,
Gurgoan, Haryana, India.

Gaurav Aggarwal
Department of CSE,
ITM University,
Gurgoan, Haryana, India.

ABSTRACT

In recent years, the interaction between humans and machines has become an issue of concern. This paper results from study of various researches related to the investigation of the six basic human emotions which include anger, dislike, fear, happiness, sadness and surprise. [1, 3] Feature extraction is done from various voice utterances recorded from different persons. The various features like pitch, energy, fundamental frequency are extracted from the utterances using respective feature extraction algorithms. After feature extraction procedure, the extracted features are classified under the basic six emotions using various machine learning algorithms. [1, 3, 4]And using the different algorithms the classification accuracy is measured for each algorithm respectively. Various acoustic and prosodic features are extracted from the speech recorded and then classified under different emotional category using machine learning tools. [7] This paper discusses how feature extraction through speech samples and then classification of the extracted features under different emotions is performed.

Keywords

Speech Emotion Recognition Models: Distributed speech recognition model; Feature Extraction; Classification Algorithms.

1. INTRODUCTION

Emotion is defined as the positive or negative state of a person's mind which is related with a pattern of physiological activities. Emotions describe the mental state of a person. Paul Ekman's research work led him to categorize emotion into six basic classes: HAPPINESS, ANGER, FEAR, DISGUST, SURPEISE and SADNESS. These six basic emotions blend to form complex emotions. For ex: - disgust and anger unify to take a new form of emotion that is contempt.[1] Human Machine interaction has become an issue of common concern over the past few decades. Speech is the most natural and efficient way of communication. [4] Earlier, traditional methods used manual processing of parameters from speech signals which were time consuming and costly. There are many applications which are earning profit by using emotion classification technique. In medical fields for online assessment of patients' disorders, Smartphone interface personalized to automatically choose song based on the current emotional state of a person. [5]

A number of speech processing technologies are there which helps in analyzing emotions from speech. Speech signals contain various information like age, sex, physical state of a person, etc. which can easily identify emotional state of a person. The classification of emotion from speech is done in a series of stages from extracting information from speech to classification of emotional content from speech.[3] The stages involved in speech emotion classification are: the speech signal is pre-processed and segmented by converting it into a wave file, after pre-processing of speech signal features extraction algorithm like Linear Prediction Coefficient (LPC), Cepstrum Coefficients algorithms are applied to extract emotional content from speech signal and at last the classification algorithms are applied to classify the emotional content.[4] The paper discuss about the Speech Emotion Recognition System, the various features contained in a speech signal and the different classification algorithms applied to classify the emotional content of speech. The features like pitch, energy, fundamental frequency, formants, etc are discussed how these are helpful in extracting the emotional information from speech signals. The various classification algorithms like Support Vector Machine (SVM) algorithm, Sequential Minimal Optimization (SMO) algorithm, Decision Tree algorithm, etc. are used for classification and the algorithm which provides the best emotion recognition rate is identified.

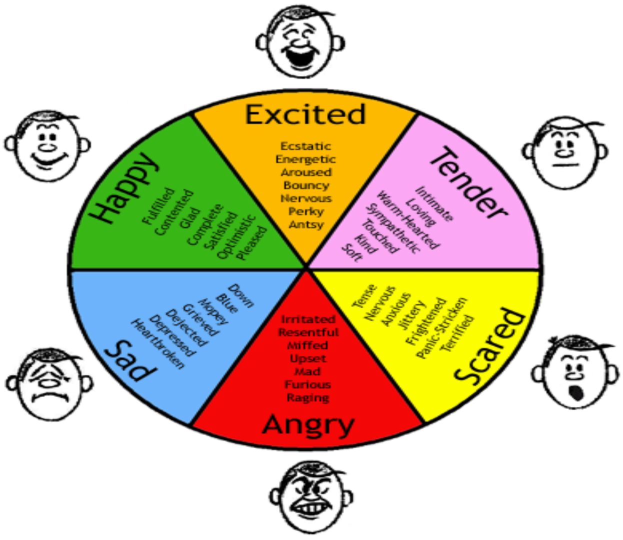

Fig.1

2. SPEECH EMOTION RECOGNITION SYSTEM

Speech is the most natural form through which humans communicate. Speech Recognition involves transformation of speech signal into a sequence of words with the help of an algorithm. Speech recognition is the capability of a machine to acknowledge the speech samples. [2, 5] The speech emotion recognition system is dependent on the naturalness of the database which contains speech as the input signal. The accuracy of the emotion recognition rate in SER system is dependent on the database used. The database which will be used as an input source for the system should hold real time world emotions. The basic architecture for SER system has the following steps shown in Fig.2 below:-[2]

I. A speech processing system is used for extracting suitable features from the speech signal like pitch or energy,

ii. The features extracted are then summarized into reduced set of features,

iii. A classifier is trained using supervised learning by providing example data how to correlate the features to the emotions.

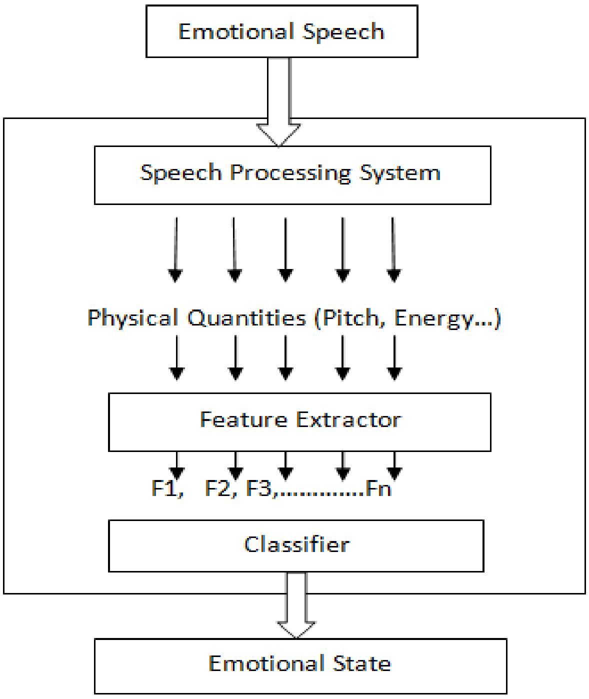

Fig.2

Speech Recognition model which have been used commonly is:-

2.1.DSR(Distributed Speech Recognition) Model,

ETSI published the first DSR in February, 2000. In this model, FRONT-END is used to transform the digitized speech into a stream of feature vectors and then the extracted features are sent to the BACK-END for further processing. The diagram below shows the working of front-end and back-end of the DSR model. [2]

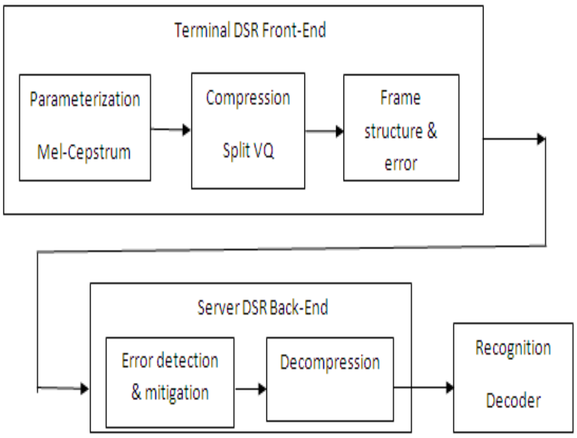

Fig.3

3. FEATURE EXTRACTION

The speech is partitioned into small intervals known as frames. The process of partitioning speech into frames based on the information they are carrying about emotion is known as feature extraction. [2, 6] Feature Extraction is a vital step in SER (Speech Emotion Recognition) System. Some of the features which helps to figure out emotions from the speech are-

3.1.PITCH: -

It is the main component of any speech which is defined as the lowness or highness of a voice as identified by the human ears. Pitch is dependent on the number or vibrations per second. The value of pitch parameter is extracted by using cepstrum in the frequency domain. [6] Pitch helps in identifying the neutral and angry emotions from speech samples.

3.2.ENERGY: -

The intensity of the speech defines the energy level of speech. Energy level for each frame is calculated as: first the square of all the sample amplitudes is done and then summing up the values of all the squared sample amplitudes. [6]

3.3.PITCH DIFFERENCE and ENERGY DIFFERENCE:

The difference between the values of pitch or energy level of neighboring segments is used to categorize the speech parameters into emotions. The more the fluctuation the more it is easier to reveal the lively emotions like happiness and anger. [6]

3.4.FORMANTS:

Formants are governed by the shape of the vocal tract and are manipulated by different emotions. For example, the state of excitement results in obtaining the higher mean values of the first formant frequency. [6] The fundamental frequency (F0) helps in identifying the happy emotion from speech samples.

3.5.MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC): -

MFCC is the most vital parameter in speech which best describes the emotional state by using simple calculations. MFCC also provides good frequency resolution when the speech frequency is low. MFCC based parameters show the energy migration in frequency domain and also helps in identifying phonetic characteristics of speech.[3,9]

4. CLASSIFICATION ALGORITHMS

The Speech emotion classification systems are trained using the various classification algorithms. The system is trained using some data sets through machine learning algorithms. Various machine learning algorithms are used to recognize the basic human emotions from the given speech samples. The recognition rate is also achieved by applying various machine learning algorithms. It has been generalized that the recognition rate for audio alone is 75% and for that of video alone is 70%. [5]But if the speech is a joint audiovisual then the recognition rate achieved is 97%. [5] Some of the algorithms which have been used to train our systems are as: -

4.1.Sequential Minimal Optimization (SMO) Algorithm

SMO algorithm is used for solving the optimization problem which occurs when training the support vector machines. John Platt discovered the SMO algorithm in 1998 at Microsoft. [3] LIBSVM tool is used to implement the SMO algorithm. The SMO algorithm is as:-

- A Lagrange multiplier is found that contravene the KKT conditions for the optimization problem.
- Next multiplier is picked up for optimizing the pair.
- Steps 1 and 2 are repeated until convergence.

4.2.Support Vector Machine(SVM) Classifier

The main motive of the SVM classifier is to track down the hyper-planes with maximum obtainable margin that sets apart the data points into classes by identifying a weight vector and an offset. [6, 12] Support Vector Machine (SVM) classifier uses binary classification based on statistical learning theory.SVM transforms the original input set to a high dimensional feature space with the help of kernel function. [6] This renovation can also be used for transforming non-linear problems. SVM can have a very good classification performance even when there is a limited training data set. SVM has the capability to generalize new and accurate data by using the trained models designed in the learning phase. An adjustable weighted segmentation (AWS) is proposed to improve the accuracy rate of SVM classifier. [6,13] AWS is a very simple approach in which each segment is assigned with a weight vector based on the type of emotion and the weights assigned are adjustable according to the input data. [6, 13]

4.3.Decision Tree Algorithm

It is a hierarchical classifier in which each node signifies an choice between a number of alternatives and each leaf node signifies a decision to be taken. This classifier is similar to the if-then-else structure. In this algorithm, the mean and the Root Mean Square (RMS) of each and every feature in emotion class are calculated. [10]

$$\text{Mean}=\frac{\text{sum of all values under the dimension}}{\text{Total number of values}}$$

$$x_{\text{rms}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$$

4.4.K Nearest Neighbor (KNN)

KNN is the most simplest and traditional classification algorithm which does not use any parameters i.e. it is a non-parametric method.The value of K nearest neighbours is provided as input from the feature space.The classification under KNN is done on the basis of number of nearest neighbours of an object . The output obtained is a set of classes. The object is assigned to a particular class on the basis of majority of votes of its nearest neighbours. The classification of samples is done by calculating the distance of that particular sample to the training set. KNN algorithm is independent of prior assumptions. Optimal value of K is chosen so that better results are obtained. But it is observed that as the value of K grows larger the effect of noise is reduced while training the data set.

5. CONCLUSION

In this paper, the study of how emotions are recognized from speech using various machine learning algorithms is discussed. The recognition rate is calculated by applying various classification algorithms and the algorithms which provides the best recognition rate is identified. The emotion recognition rate is dependent on the types of features extracted and the selection of the classification algorithm. From the study, it has been evaluated that the SVM and the SMO algorithms are better classification algorithms which gives higher accuracy in emotion recognition rates. Our future work will include the recording of speech samples of small children to adult speech samples and then extracting the features like pitch and fundamental frequency from those recorded data set. After extraction of the feature vectors we will be applying the classification algorithms to recognize emotions from the recorded speech dataset. The classification algorithms will help in evaluating that how emotional state of a person changes from a child to adult. Our proposed work is based on using Machine learning technique to develop a speech emotion recognition system with more correctness and efficient than the already existing systems. For identification of voice of children and adults, we need to create a database for making it more robust. Without database it will be difficult for the system to differentiate between emotions of child and adults. The proposed topic on which I have decided to work is “Characterization of emotion from specch using machine learning algorithm”. We will be recording audio messages of small children from 4 to 8 years and also of some adult males and females. After that the audio samples are recorded, the samples are converted into monowave format. The next step will be extraction of features from the speech samples. After the features are extracted from the speech samples, we will be applying one of the above classification algorithms. This will help us in classifying how the emotion of human beings is affected as tha age increases.

6. REFERENCES

[1] “Theories of Emotion”. Psychology.about.com. 13 September 2013. Reteieved 11 November 2013.

[2] Gaulin, Stevens J.C. and Donald H. McBurney. Evolutionary Psychology. Prentice Hall. 2003. ISBN 978-0-13-111529-3, Chapter 6, p 121-142.

[3] http://www.ijarcsse.com

[4] Casale, S., Russo, A., Scebba, G., & Serrano, S. (2008, August). Speech emotion classification using machine learning algorithms. In Semantic Computing, 2008 IEEE International Conference on (pp. 158-165). IEEE.

[5] Hassan, E. A., El Gayar, N., &Moustafa, M. G.(2010,November). Emotions analysis of speech for call classification. In Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on (pp.242-247). IEEE.

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

[7] Yang, N., Muraleedharan, R., Kohl, J., Demirkol, I., Heinzelman, W. R., &Sturge-Apple, M. (2012, December). Speech-based emotion classification using multiclass SVM with hybrid Kernel and thresholding fusion. In SLT(pp. 455-460)

[8] Chen, Y. Y.,Chen, B. W., Wang, J. F., & Chen, Y. C. (2010, Novenber). Emotion aware system based on acoustic and textual features from speech . InAware Computing (ISAC), 2010 2nd International Symposium on (pp . 92-96). IEEE.

[9] Planet, S., &Iriondo,I. (2012,June). Comparison between decision-level and features for spontaneous emotion rfcognition. In Information Systems and Technologies(CISTI),2012 7th Iberian Conference on (pp. 1-6). IEEE.

[10] Khan A &Baharudin B. (2011,September). Sentiment classification using sentence-level semantic orientation of opinion terms from blogs. In National Postgraduate Conference(NPC),2011(pp. 1-7).IEEE.

[11] Srieam, S., & Yuan, X (2012,March). An enhanced approach for classifying emotions using customized decision tree algorithm . In Southeastcon 2012 proceedings of IEEE (pp. 1-6), IEEE.

[12] Houjeij, A., Hamieh, L., Mehdi, N., & Hajj, H. (2012, April). A novel approach for emotion classofocation based on fusion of text and speech. In Telecommunications (ICT), 2012 19th International Conference on (pp. 1-6). IEEE.

[13] Kudiri, K. M., Said, A. M., & Nayan, M. Y. (2012, November). Emotion detection using average relative amplitude features through speech. In Control System, Computing and Engineering (ICCSCE), 2012 IEEE International Conference on (pp. 115-118). IEEE.

[14] Wu, Y. H., Lin, S. J., & Yang, D. L. (2013, September). A mobile emotion recognition system based on speech signals and facial images. In Computer Science and Engineering Conference (ICSEC), 2013 International (pp. 212-217). IEEE.