International Journal of Scientific Research in Science and Technology
Print ISSN: 2395-6011 | Online ISSN: 2395-602X (www.ijsrst.com)
doi : https://doi.org/10.32628/IJSRST229168

Speech Based Emotion Recognition Using Machine Learning

Dr. Loganathan R.[1], Arjumand Yusufi[2], Umar Rasool Yetoo[2], Waseem Ahmed R. [2], Zaki Hussain[2]
[1]Head of Department, Department of CSE, HKBKCE, Bangalore, Karnataka, India
[2]Student, Department of CSE, HKBKCE, Bangalore, Karnataka, India

Article Info
Volume 9, Issue 1
Page Number : 324-329

Publication Issue
January-February-2022

Article History
Accepted : 20 Jan 2022
Published : 28 Jan 2022

ABSTRACT

Emotion is a natural feeling which is distinguished from reasoning or knowledge, it is a strong feeling derived from one's circumstance or surroundings. 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 sub field 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. When implemented the Speech Emotion Recognition (SER) will be able to understand different human emotion such as anger, fear, happiness, sadness etc. Speech is a medium of expression of one's perspective or feelings to other. Emotion recognition from audio signal requires feature extraction and classifier training. The feature vector consists of elements of the audio signal which characterize speaker specific features such as tone, pitch, energy, which is crucial to train the classifier model to recognize a particular emotion accurately. Thus, with the help of SER we can make conversations between human and computer more realistic and natural. Automatic Speech Emotion Recognition (SER) is a current research topic in the field of Human Computer Interaction (HCI) with wide range of applications. The speech features such as, Mel Frequency cepstrum coefficients (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) are extracted from speech utterance. The Support Vector Machine (SVM) is used as classifier to classify different emotional states such as anger, happiness, sadness, neutral, fear, from Berlin emotional database.

Keywords : Speech Emotion Recognition, Human Computer Interaction, Mel Frequency cepstrum coefficients, Support Vector Machine

I. INTRODUCTION

Automatic Speech Emotion Recognition is a very recent research topic in the Human Computer Interaction (HCI) field. As computers have become an integral part of our lives, the need has risen for a more natural communication interface between humans and computers. To achieve this goal, a

computer would have to be able to perceive its present situation and respond differently depending on that perception[1]. Part of this process involves understanding a user's emotional state. To make the human-computer interaction more natural, it would be beneficial to give computers the ability to recognize emotional situations the same way as human does. Automatic Emotion Recognition (AER) can be done in two ways, either by speech or by facial expressions. In the field of HCI, speech is primary to the objectives of an emotion recognition system, as are facial expressions and gestures[2]. Speech is considered as a powerful mode to communicate with intentions and emotions In the recent years, a great deal of research has been done to recognize human emotion using speech information Many researchers explored several classification methods including the Neural Network (NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Maximum Likelihood Bayes classifier (MLC), Kernel Regression and K-nearest Neighbors (KNN), Support Vector Machine (SVM) The Support Vector Machine is used as a classifier for emotion recognition. The SVM is used for classification and regression purpose. It performs classification by constructing an N-dimensional hyperplanes that optimally separates the data into categories. The classification is achieved by a linear or nonlinear separating surface in the input feature space of the dataset. Its main idea is to transform the original input set to a high dimensional feature space by using a kernel function, and then achieve optimum classification in this new feature space.

- RAVDESS : The Ryson Audio-Visual Database of Emotional Speech and Song that contains 24 actors (12 male, 12 female), vocalizing two lexically-matched statements in a neutral North American accent.
- TESS : Toronto Emotional Speech Set that was modeled on the Northwestern University Auditory Test No. 6 (NU-6; Tillman & Carhart, 1966). A set of 200 target words were spoken in the carrier phrase "Say the word' by two actresses (aged 26 and 64 years).
- EMO-DB : As a part of the DFG funded research project SE462/3-1 in 1997 and 1999 we recorded a database of emotional utterances spoken by actors. The recordings took place in the anechoic chamber of the Technical University Berlin, department of Technical Acoustics. Director of the project was Prof. Dr. W. Sendlmeier, Technical University of Berlin, Institute of Speech and Communication, department of communication science. Members of the project were mainly Felix Burkhardt, Miriam Kienast, Astrid Paeschke and Benjamin Weiss.
- Custom : Some unbalanced noisy dataset that is located in data/train-custom for training and data/test-custom for testing in which you can add/remove recording samples easily by converting the raw audio to 16000 sample rate, mono channel (this is provided in create_wavs.py script in convert_audio(audio_path) method which requires ffmpeg to be installed and in PATH) and adding the emotion to the end of audio file name separated with '_' (e.g "20190616_125714_happy.wav" will be parsed automatically as happy).

Applications of Speech Emotion Recognition include psychiatric diagnosis, intelligent toys, lie detection, learning environment, educational software, and detection of the emotional state in telephone call center conversations to provide feedback to an operator or a supervisor for monitoring purposes.

II. SYSTEM IMPLEMENTATION

The importance of emotions in human-human interaction provides the basis for researchers in the engineering and

Figure 1. Speech Emotion Recognition System.

computer science communities to develop automatic ways for computers to recognize emotions. As shown

in fig. 1 the input to the system is a .wav file from Berlin Emotion Database that contains emotional speech utterance from different emotional classes. After that features extraction process is carried out. In feature extraction process two features are extracted MFCC [6], [7] and MEDC [8]. After that the extracted features and their corresponding class labels are given as input to the LIBSVM classifier. The output of a classifier is a label of a particular emotion class. There are total five classes angry, sad, happy, neutral and fear. Each label represents corresponding emotion class.

III. PROPOSED SYSTEM

3.1 Feature Extraction

In previous works several features are extracted for classifying speech affect such as energy, pitch, formants frequencies, etc. all these are prosodic features. In general prosodic features are primary indicator of speaker's emotional state. Here in feature extraction process two features are extracted Mel Frequency Cepstral Coefficient (MFCC) and Mel Energy spectrum Dynamic coefficients (MEDC). Fig. 2 shows the MFCC feature extraction process. As shown in Fig. 2 feature extraction process contains following steps:

- Preprocessing: The continuous time signal (speech) is sampled at sampling frequency. At the first stage in MFCC feature extraction is to boost the amount of energy in the high frequencies. This reemphasis is done by using a filter.
- Framing: it is a process of segmenting the speech samples obtained from the analog to digital conversion (ADC), into the small frames with the time length within the range of 20-40 ms. Framing enables the non-stationary speech signal to be segmented into quasistationary frames, and enables Fourier Transformation of the speech signal. It is because, speech signal is known to exhibit quasistationary behavior within the short time period of 20-40 ms.

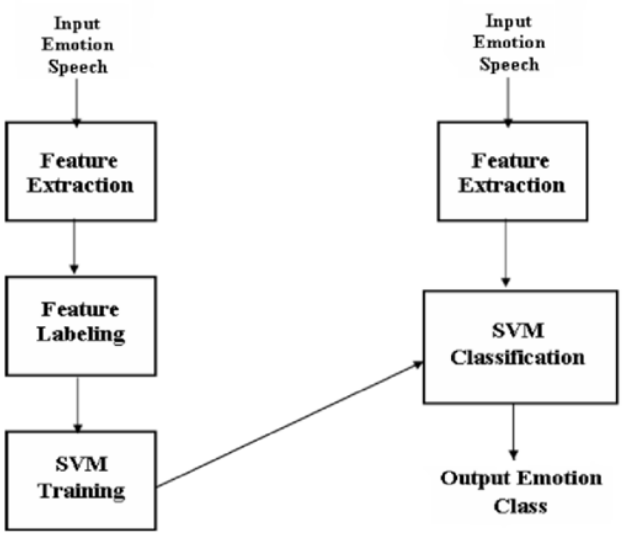

Figure 2. MFCC feature extraction

- Windowing: Windowing step is meant to window each individual frame, in order to minimize the signal discontinuities at the beginning and the end of each frame.
- FFT: Fast Fourier Transform (FFT) algorithm is ideally used for evaluating the frequency spectrum of speech. FFT converts each frame of N samples from the time domain into the frequency domain.
- Mel Filterbank and Frequency wrapping: The mel filter bank [8] consists of overlapping triangular filters with the cutoff frequencies determined by the center frequencies of the two adjacent filters. The filters have linearly spaced centre frequencies and fixed bandwidth on the mel scale[5]. • Take Logarithm: The logarithm has the effect of changing multiplication into addition. Therefore, this step simply converts the multiplication of the magnitude in the Fourier transform into addition
- Take Discrete Cosine Transform: It is used to orthogonalise the filter energy vectors. Because of this orthogonalization step, the information of the filter energy vector is compacted into the first number of components and shortens the vector to number of components.

- Another feature Mel Energy spectrum Dynamic coefficients (MEDC) is also extracted. It is extracted as follows: the magnitude spectrum of each speech utterance is estimated using FFT, then input to a bank of 12 filters equally spaced on the Mel frequency scale. The logarithm mean energies of the filter outputs are calculated En(i), i= 1.....N. Then, the first and second differences of En(i) are calculated. MEDC feature extraction process. The MEDC feature extraction process contains following steps shown in figure 3:

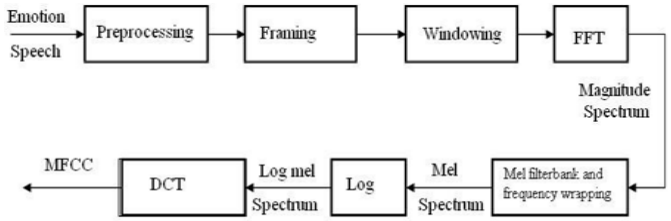

Figure 3. MEDC feature extraction

- Preprocessing, Framing, Windowing, FFT & Mel filter bank and Frequency wrapping processes of MEDC feature extraction are same as MFCC feature extraction.
- Take logarithmic mean of energies: In this process a mean log of every filter energy is calculated. This mean value represents energy of individual filter in a filter bank.
- Compute 1st and 2nd difference: The final Mel energy spectrum dynamics coefficients are then obtained by combining the first and second differences of filter energies.

3.2 Feature Labelling

In Feature labelling each extracted feature is stored in a database along with its class label. Though the SVM is binary classifier it can be also used for classifying multiple classes. Each feature is associated with its class label e.g. angry, happy, sad, neutral, fear

3.3 SVM Classification

In general, SVM is a binary classifier, but it can also be used as a multiclass classifier. LIBSVM [9], [10] is a most widely used tool for SVM classification and regression developed by C. J. Lin. Radial Basis Function (RBF) kernel is used in training phase. Advantage of using RBF kernel is that it restricts training data to lie in specified boundaries[9]. The RBF kernel nonlinearly maps samples into a higher dimensional space, so it, unlike the linear kernel, can handle the case when the relation between class labels and attributes is nonlinea[8]r. The RBF kernel has less numerical difficulties than polynomial kernel. The other classifiers that can be implemented into the project are:

- RandomForestClassifier.
- GradientBoostingClassifier.
- Recurrent Neural Networks

IV. EXPERIMENTATION AND RESULTS

Berlin Emotion database contains 406 speech files for five emotion classes. Emotion classes Anger, sad, happy, neutral, fear are having 127, 62, 71, 79 and 67 speech utterance respectively. The LIBSVM is trained on MFCC and MEDC feature vectors using RBF and Polynomial kernel functions. The LIBSVM is used to test these feature vectors. The experimentation is carried out by varying cost values for RBF kernel and degree values for Polynomial kernel. Both gender independent and gender dependent experiments are performed. Using RBF kernel at cost value c=4, it gives recognition rate of 93.75% for gender independent case, 94.73% for male and 100% for female speeches. The recognition rate using Polynomial kernel at degree d=4 is 96.25% gender independent, 97.36% for male and 100% for female speeches.

Table 1. Confusion matrix of the RBF LIBSVM classifier (Gender Independent)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0

Neutral	0	6.25	0	93.75	0
Fear	0	0	30.76	0	69.24

The Confusion matrices using RBF kernel gender independent, male and female are shown in Table 1, 2 and 3. Table 4, 5 and shows Confusion matrices using Polynomial kernel gender independent, male and female.

Table 2. Confusion matrix of the RBF LIBSVM classifier (Male)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	16.66	0	83.34	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	14.85	85.15

Table 3. Confusion matrix of the RBF LIBSVM classifier(Female)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

Table 4. Confusion matrix of the Polynomial LIBSVM classifier (Gender Independent)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	0	0	100	0
Fear	7.69	0	15.18	0	76.92

Table 5. Confusion matrix of the Polynomial LIBSVM classifier (Male)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	0	0	100	0
Fear	0	0	14.28	0	85.72

Table 6. Confusion matrix of Polynomial LIBSVM classifier (Female)

Emotion	Emotion Recognition (%)				
	Angry	Sad	Happy	Neutral	Fear
Angry	100	0	0	0	0
Sad	0	100	0	0	0
Happy	0	0	100	0	0
Neutral	0	0	0	100	0
Fear	0	0	0	0	100

V. CONCLUSION

In this paper and throughout this project, we showed how we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of above-mentioned setups like human-computer interaction, Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.

VI. FUTURE WORK

A few possible steps that can be implemented to make the models more robust and accurate are the following

An accurate implementation of the pace of the speaking can be explored to check if it can resolve some of the deficiencies of the model.

Figuring out a way to clear random silence from the audio clip.

Exploring other acoustic features of sound data to check their applicability in the domain of speech emotion recognition. These features could simply be some proposed extensions of MFCC like RAS-MFCC or they could be other features entirely like LPCC, PLP or Harmonic cepstrum.

Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models. This will improve the accuracy of the system because in some cases the expression of emotion is contextual rather than vocal.

Adding more data volume either by other augmentation techniques like time-shifting or speeding up/slowing down the audio or simply finding more annotated audio clips.

VII. REFERENCES

[1]. M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC." 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, 2017, pp. 2257-2260.Christopher. J. C. Burges, A tutorial on support vector machines for pattern recognition, DataMining and Knowledge Discovery, 2(2):955-974, Kluwer Academic Publishers, Boston, 1998.

[2]. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., Emotion recognition in human-computer interaction, IEEE Signal Processing magazine, Vol. 18, No. 1, 32-80, Jan. 2001.

[3]. Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001.

[4]. M. D. Skowronski and J. G. Harris, Increased MFCC Filter Bandwidth for Noise-Robust Phoneme Recognition, Proc. ICASSP-02, Florida, May 2002.

[5]. Fuhai Li, Jinwen Ma, and Dezhi Huang, MFCC and SVM based recognition of Chinese vowels, Lecture Notes in Artificial Intelligence, vol.3802, 812-819, 2005

[6]. 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.

[7]. Burkhardt, Felix; Paeschke, Astrid; Rolfes, Miriam; Sendlmeier, Walter F.; Weiss, Benjamin A Database of German Emotional Speech. Proceedings of Interspeech, Lissabon, Portugal. 2005

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

Cite this article as :

Dr. Loganathan R., Arjumand Yusufi, Umar Rasool Yetoo, Waseem Ahmed R., Zaki Hussain, "Speech Based Emotion Recognition Using Machine Learning ", International Journal of Scientific Research in Science and Technology (IJSRST), Online ISSN : 2395-602X, Print ISSN : 2395-6011, Volume 9 Issue 1, pp. 324-329, January-February 2022. Available at doi : https://doi.org/10.32628/IJSRST229168

Journal URL : https://ijsrst.com/IJSRST229168