Speech Emotion Recognition using Support Vector Machine

ISSN (e) 2520-7393
ISSN (p) 2521-5027
Received on 18thApril, 2019
Revised on 15th June,2019
www.estirj.com

Asif Ali[1], M. Moazzam Jawaid[2], Anees Muhammad[3], Noor u Zaman[4], Qadir Bakhsh Jamali[5]

[1, 3] *Institute of Information and Communication Technology MUET Jamshoro*
[2, 4] *Department of Computer Systems Engineering MUET Jamshoro*
[5] *Department of Mechanical Engineering, QUEST Nawabshah*

Abstract: The Human computer interaction (HCI) whereas emotion recognition of speech through machine learning has many outcomes like to monitor the psychological behavior of human; it is also used in lie detection. The emotion of human can be recognized by different ways mostly by speech or facial expression. To make more natural communication edge between human and computer; speech has one of the primary objectives to recognize an emotion as facial expression. But to classify the accuracy of the speech emotions, it has many flaws and should be overcome. Here in this research; to recognize the emotion of speech, we are going to use Support Vector Machine (SVM) it perfectly classifies the data into groups; that categorization is accomplished by linear or no-linear untying plane in an input features space of the data sets. Whereas, in this research four features of speech corpus are selected and extracted out those fundamental speech recognition features form each speech corpus and classified that data by using SVM and achieved the testing accuracy of 87.5% correctly.

Keywords: *Human Computer Interaction, Support Vector Machine, Speech Emotion, Pitch, Energy, Mal Frequency Cepstrum Co-Efficient, Zero-crossing*

1. Introduction

In this era of modern technology automatic speech recognition has enormous position and ample applications. It is the study of structure and alteration in speech that human speech contains during different emotional states. Therefore, to make an improved human–machine interface for speech emotion recognition: is very challenging one. As in our daily life computer has become a vital element. The requisite has increased for an additional ordinary communication interface among human and computers [1,2].

To accomplish such objectives: computers have got to enclose the capabilities that identify their current conditions and react accordingly depending on acknowledged observations. These process factors entail accepting a user's emotional situation. To build the HCI more ordinary to make productive machine learning it would be valuable to deliver computers with aptitude to identify expressive condition the identical means as human does [3]. In machine learning like HCI where emotions are express in many ways through using facial or gestures expressions like that speech is one of the deep-seated elements from which emotions are recognized. Recently many scholars have done their research work on finding accurate emotions from speech [4]. Whereas many issues are there: What kinds of features set are required for speech emotion recognition? How to combine these acoustic features and which classification method is fitting for classification [5]?

Here the projected technique seeks to recognize emotional state of speaker such as sad, anger, happy or neutral. For classifying the various emotions from some very fundamental features of speech signal like energy, pitch, MFCC and zero-crossing are exercised. Whereas Support Vector Machine is a classifier used to notice the emotion from given features set. SVM offers healthy categorization outcome on data encloses extremely huge number of variables [6].

2. Related Work

Many types of techniques are projected for speech emotion recognition from different researchers. Some of these are reviewed as Namrata Dave *et al.,* has evaluated some features extraction techniques which are MFCC, Perceptual Linear Prediction (PLP) and Linear Predictive Codes (LPC), he conferred their pros and cons. According to researcher LPC is frequently used for automatic speech emotion recognition whereas LPC parameters are not practical for speech emotion due to linearity. Whereas human voice is non-linear in nature and he observed that MFCC and PLP are best for speech emotion recognition [7]. Rajni *et al.*, have proposed emotion recognition from Hindi speech. The features analyzed from speech signal for getting feature vector are: pitch, formants, fundamental frequency and duration (ratio of pause). Then features are classified by using K Nearest Neighbor classifier. A proportional investigation shows that KNN classifier have accomplished most excellent results than another neural network classifier [8].

Corresponding author Email address*: asif.jamali@usindh.edu.pk*

Shashidhar *et al.,* have offered emotion recognition using Hindi speech. In [9], he utilized the pitch, duration, energy and MFCC features of audio speech signal to classify the emotion. Han *et al.,* have proposed emotion recognition system using music clips. The MIR tool box is applied for feature extraction from music clip which are pitch, tempo, tonality, dynamic, timber. Here it is examined that the values of pitch of happy song are higher than sad song and sad songs have very low tempo while happy songs have higher tempo [10]. Iliou *et al.,* have anticipated a speech emotion detection from Berlin emotion corpus of German Language. Kinds of features are extracted for emotion detection are pitch, MFCC, energy and formants. For classification of emotion from features set researchers have comparatively investigate the three techniques which are Random Forest, SVM, Probabilistic Neural Networks, Multilayer Perception. This research shows that PNN has adequate results whereas for future work should embrace a hybrid method to examine emotion more perfectly [11].

Although it is hard to acquire an accurate outcome but we can show the deviation that occur when emotion alters by using MFCC algorithm, features are mined from which we can observe how alters occurs in pitch, frequency and other features. When emotion alters, we have made frame blocking and windowing steps of MFCC algorithms for a matching voice and a same sentence in tow diffident emotion and showed difference in pitch with change in emotion by using classifier algorithms SVM to classify different emotion [7].

In this research Berlin emotion database of German language is exercised for feature extraction. Mal Frequency Cepstrum Co-Efficient (MFCC) and Mel Energy Spectrum Dynamic Coefficients (MEDC) features are observed from a speech records in .wav format. By testing and outcomes, it is verified that systems is speaker and text autonomous. It is also observed that outcomes from LIBSVM using RBF and polynomial kernel are 93.75% and 96.25% correspondingly. Regarding Library for Support Vector Machines (libsvm) via Radial Basis Function (RBF) kernel and polynomial kernels it is observed that by varying the parameters of a kernel functions better result can be obtain [8].

Due to very less awareness about the field there are very limited researches going on in the field of speech processing. But a huge quantity of effort can be done by processing the spectral features efficiently to recognize the emotion. Here using 24 MFCC features an accuracy of 68% has been observed in the Berlin emotional database. Besides that, elevated accuracy can be achieved using the grouping of more features. To surmise, prospect work is to mine the delta features from each expression and then use the SVM hierarchical formation for classification. Also though increasing the sigma value from the default value one, considerable results may be achieved [12].

With the improvement of communication and security technologies, it has happened to critical to have sound of embedded biometric systems. Here research represents the comprehension of such technologies which demands trustful and rational biometric identity authentication systems. Multi-dimensional prototypes are not permitted due to Eigen-decomposition in multi-dimensional feature space and degeneration of dispersion matrices in tiny size sample. Simplification, dimensionality diminution and maximize the margins are restricted by reducing weight vectors. Therefore, the aim of examine a biometric identity system using SVM and Lindear Discriminant Analysis (LDA) with MFCCs and apply such system in real-time by Signal WAVE because results are not sufficient for classification and are very poor by using SVM only [10].

2.1 Algorithms applied in SER

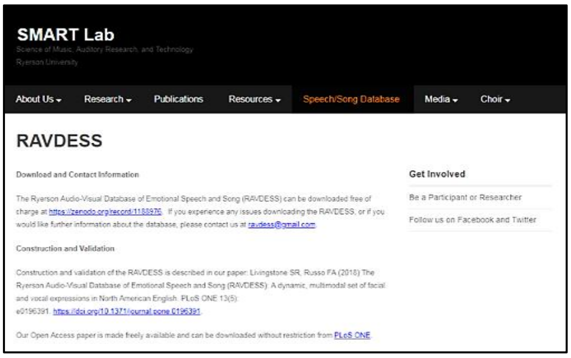
Many classification techniques are used for recognizing emotions from acoustic speech signal among them some techniques are very familiar classifier for speech processing which are HMM, GMT, ANN and SVM. Hidden Markov Model contains propose issues like establish the optical number of states and its output performance for SER is satisfactory. It is argued that four state HMM grants the satisfactory output. GM Technique characterizes key point with combined probability solidity methods. It is depending on supposition that all vectors are autonomous and so couldn't be molded temporal patterns. Baykal conducted a study with 74.83% usual classification precision achieved using a 16 elements GMMs [13]. The output is very comparable to HMMS. ANNs use a 2-layer Neural Network, it knows actually in non-linear mappings. As compare to other classifier ANN gives no satisfactory outcomes [14].

3. Methodology

3.1 Dataset

It is very important part of making an innovative speech emotion recognition system is the preference of the dataset. However, making such dataset which satisfies all needs and become a standard for all time research is very challenging. An English emotion speech dataset: RAVDESS [12] is evaluated for this research. The dataset is an open source database and effortless to access. This dataset has 24 professional actors (12 actors are male and 12 actors are female). Here in this research four speech emotions are going to be examined that are Happy, Anger, Sadness and Neutral.

For RAVDESS dataset the SMART LAB has pursued the procedure in which the actors were employed online as shown in Figure 1. A total of 58 actors are tested and all those tested videos were examined by author and investigators for accuracy. Finally, 24 actors with top cumulative score are accepted and actors were reserved for 4 hours and were paid for recording.


SMART Lab
Science of Music, Auditory Research, and Technology
Ryerson University
About Us
Research
Publications
Resources
Speech/Song Database
Media
Choir
Get Involved
Be a Participant or Researcher
Follow us on Facebook and Twitter

RAVDESS

Download and Contact Information

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) can be downloaded free of charge at https://zenodo.org/record/1188976. If you experience any issues downloading the RAVDESS, or if you would like further information about the database, please contact us at ravdess@gmail.com.

Construction and Validation

Construction and validation of the RAVDESS is described in our paper: Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

Our Open Access paper is made freely available and can be downloaded without restriction from PLoS ONE.

Figure. 1. SMART Lab for RAVDESS dataset

The detailed flowchart of that procedure by smart lab has been introduced by RAVDESS database using by SMART Lab as shown in Figure 2.

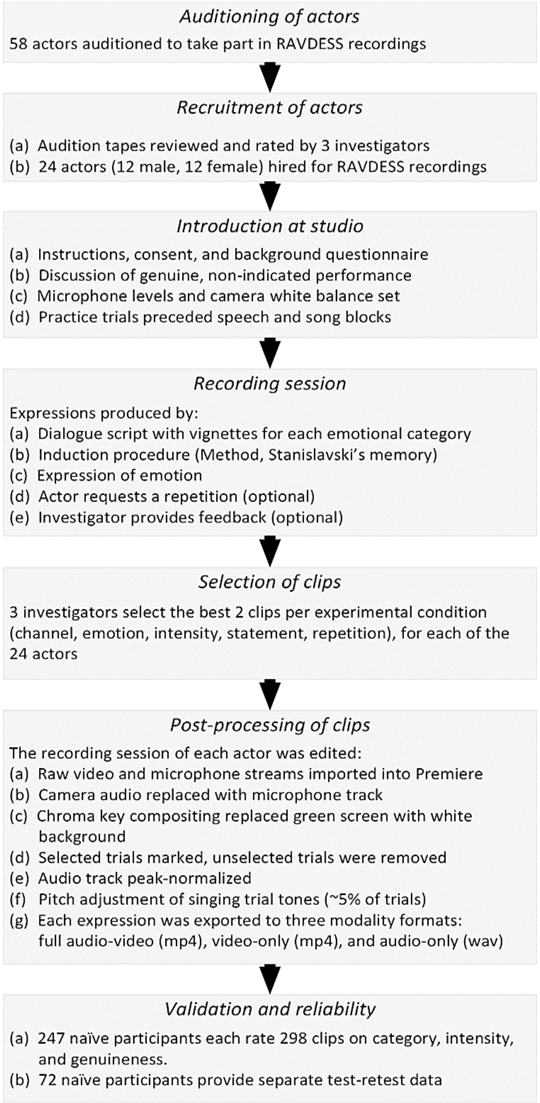

Auditioning of actors
58 actors auditioned to take part in RAVDESS recordings
Recruitment of actors
(a) Audition tapes reviewed and rated by 3 investigators
(b) 24 actors (12 male, 12 female) hired for RAVDESS recordings
Introduction at studio
(a) Instructions, consent, and background questionnaire
(b) Discussion of genuine, non-indicated performance
(c) Microphone levels and camera white balance set
(d) Practice trials preceded speech and song blocks
Recording session
Expressions produced by:
(a) Dialogue script with vignettes for each emotional category
(b) Induction procedure (Method, Stanislavski's memory)
(c) Expression of emotion
(d) Actor requests a repetition (optional)
(e) Investigator provides feedback (optional)
Selection of clips
3 investigators select the best 2 clips per experimental condition (channel, emotion, intensity, statement, repetition), for each of the 24 actors
Post-processing of clips
The recording session of each actor was edited:
(a) Raw video and microphone streams imported into Premiere
(b) Camera audio replaced with microphone track
(c) Chroma key compositing replaced green screen with white background
(d) Selected trials marked, unselected trials were removed
(e) Audio track peak-normalized
(f) Pitch adjustment of singing trial tones (~5% of trials)
(g) Each expression was exported to three modality formats: full audio-video (mp4), video-only (mp4), and audio-only (wav)
Validation and reliability
(a) 247 naïve participants each rate 298 clips on category, intensity, and genuineness.
(b) 72 naïve participants provide separate test-retest data

Figure. 2. Flowchart for RAVDESS database using by SMART Lab

3.2 Feature Extraction

To examine the emotion of speech; four basic features which are given below are extorted from each actor's speech. For that a system is projected as expressed in Figure 3.

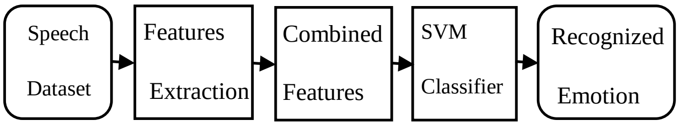


Figure. 3. Designed System for Speech Emotion Recognition

3.3 Simulation Tools

3.3.1 PRAAT

It is freely available software developed by Paul Boersma and David Weenink of the University of Amsterdam that is used mostly for analysis and reform of speech signals. For speech analysis PRAAT is very elastic tool as shown in Figure 4. It makes very high quality images for research articles and theses. It offers many standard procedures including speech analysis, articulatory synthesis, learning algorithms, neural networks and much more.

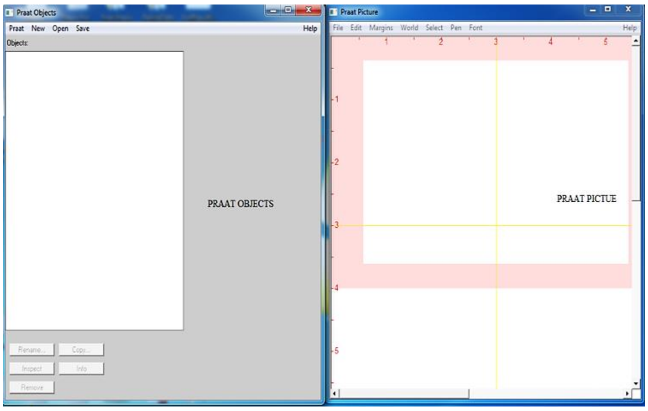


Figure. 1. IDE of PRAAT software

3.3.2 MATLAB

Millions of researchers apply MATLAB to examine and propose the systems to fulfill their projects requirements. So, to following that path for this research the MATLAB R2015a signal processing application have been used as shown in Figure 5. Audio toolbox is one of the Signal Processing applications which propose and examine speech, acoustic and audio processing systems. Using Audio toolbox, the MFCC features matrix of speech corpus has extracted.

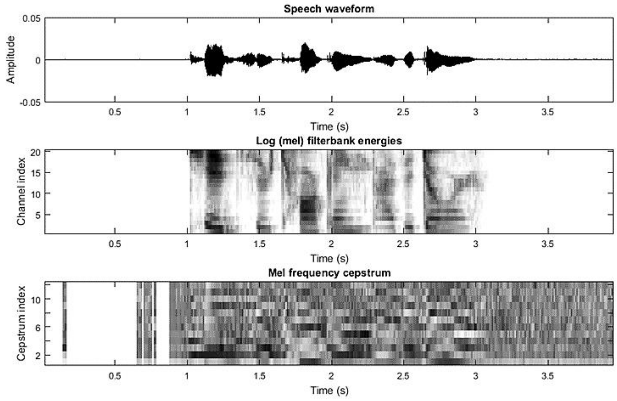


Figure. 5. MFCC spectral view of speech signal

3.4 Energy (Time-Domain)

There are many features of speech signal among them energy is also one of fundamental feature. Energy commonly declares the intensity or volume of speech signal. Energy gives information that can be used to distinguish sets of emotions. Anger and Happy have augmented level of energy, whereas Sadness has low level of energy. Mean of energy is taken into consideration in proposed emotion recognition system [6-7]. The energy of speech signal can be expressed as follows:

$$E = \sum_{n=0}^{N-1} |x(n)w(n-m)|^2 \qquad (1)$$

Where *E* is energy, signal sequence is *x(n)* and w(n-m) is the window series with limited time.

3.5 Pitch

The speech wave frequencies that ear recognize is known as pitch. Higher pitch represents higher speech frequency and lower pitch has lower speech frequency. Among many methods for calculating pitch from speech frequency I have applied auto-correlation method that can be express as follows:

$$R(K) = \sum_{i=1}^{N-1} x(n)x(n+k) \qquad (2)$$

Where *R(k)* is auto-correlation, signal frame is *x(m)* and sift of parameter is x(n+*k).*

3.6 MFCC

In speech processing MFCC is a perfect demonstration of short time power spectrum of speech. MFCC are coefficients of cepstral acquired from a spectrum filtered by Mel scale as shown in Figure 6. These scales enthused by the features of human inspection, that are totally diverse from the normal Cepstrum, as like that the bands of the frequency are logarithmically located permitting approximating the auditory system of human which response more closely than the linearly spaced frequency bands obtained directly from the FFT or DCT. A data vector of all frames of speech signal is achieved in matrix forms that are the features coefficients MFCC. In this output matrix the rows symbolize the resultant frame numbers and columns represent corresponding feature vector coefficients. The spectrogram view for Angry, Happy, Sad and Neutral speech are shown in Figure 6 to Figure 9 respectively.

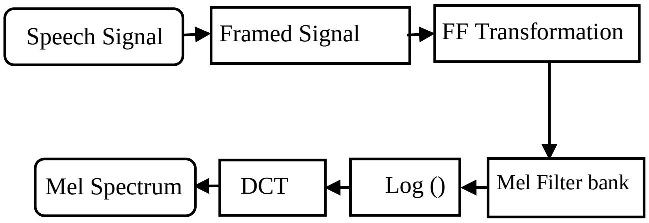


Figure. 5. Features Extraction Schema for MFCC

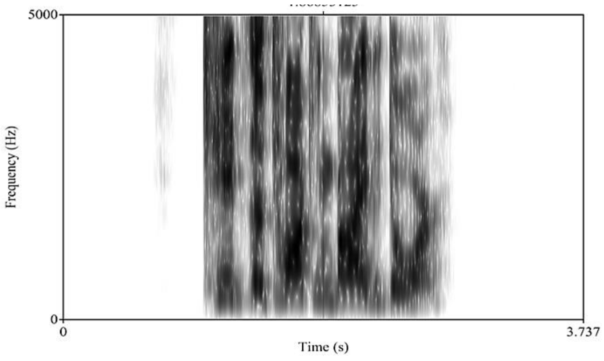


Figure. 6. Spectrogram view of Angry speech

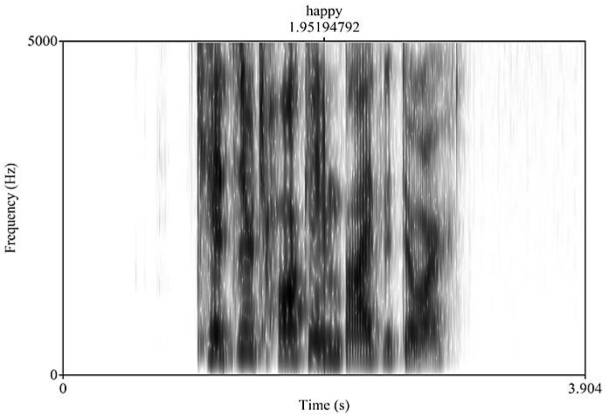


Figure. 7. Spectrogram view of Happy speech

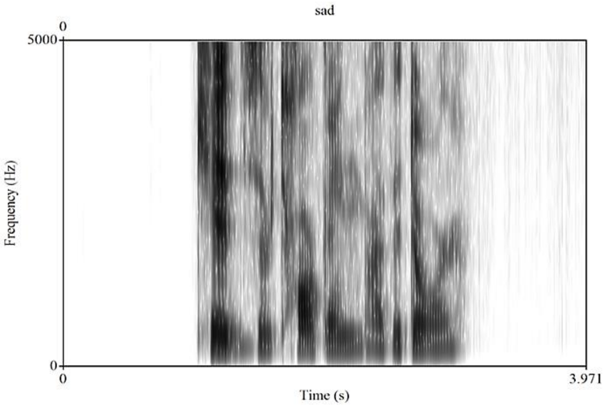


Figure. 8. Spectrogram view of Sad speech

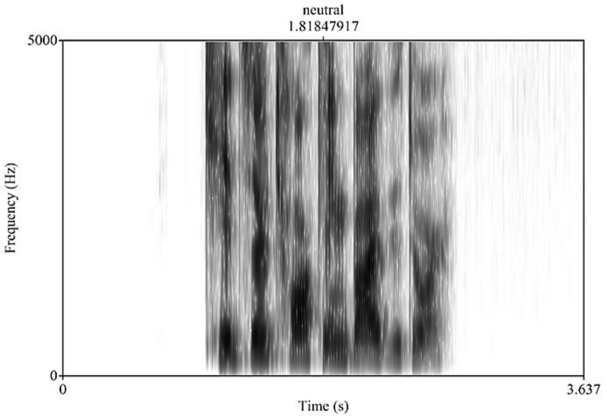


Figure. 9. Spectrogram view of Neutral speech

3.6 Zero- Crossing Rate

The number of times at given frame/interval the amplitude of speech signal cross through a value of zero is known as zero-crossing rate. In speech analysis ZCR is one of temporal feature. It proposes to the number of times speech samples change sign in a given frame. The ZCR for each unit time can be express as follows:

$$Z = n_c\left(\frac{f}{n}\right) \qquad (3)$$

Where *nc* is zero crossing per frame, sampling frequency is *f* and *n* is length of frame.

3.7 Support Vector Machine (SVM)

The SVM is a supervised classifier method commonly represented by extricating hyperplane as Figure 10. From a specified labeled (supervised) data, the SVM produce best possible hyperplane which classify latest examples; either it falls into that class or not.

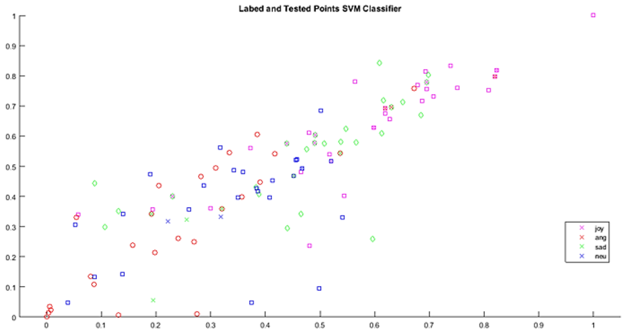


Figure. 10. Labeled and tested key points using SVM classifier

4. Results

SER study is one of the rising phases due to very complex task of designing and extracting emotions form speech corpus. Many researchers are involved in extracting out speech emotion speech corpus of different languages. The aim of SER is to give a very interactive HCI interface that interact with the humans and behave as like their emotion. In Table 1 it is accomplished a matrix of confusion, to understand the efficiency on finding emotions. 84 speech emotions were classified exactly out of 96 speech corpus.

Table. 1. Matrix of Confusion

Emotion Recognition				
Emotions	**Sad**	**Anger**	**Happy**	**Neutral**
Sad	20	2	0	1
Anger	1	19	1	1
Happy	1	1	23	0
Neutral	2	2	0	22

The Table 1 shows the output of emotions those are Sad, Anger, Happy, and Neutral respectively. All diagonal shows the feature points while other points which are misclassified emotions. The SVM classified 20 emotions out of 24 as sad emotion, 19/ 24 emotions as anger, 23/24 emotions as happy and 22/24 emotions as neutral.

5. Conclusion

In the present days, it is one of the most momentous research areas; to judge the emotions of an actor from their speech. Basically, it has a central role in HCI. Whereas it is one of the composite tasks to judge the emotion of a speaker through their speech. So, for achieving that we have applied SVM for acquiring better classification accuracy. MFCC, Energy, Zero Crossing and Pitch features are extracted from 24 actors' speech utterances and emotions are classified using SVM.

The output of SVM classifier has testing accuracy of 87.5% as it has correctly classified the 84 speech emotions from 96 speech utterances.

In future this research could be improved for accuracy by applying more data preprocessing techniques and testing the system by assigning further feature values to achieve the better outcome of system.

References

[1] Luo Q. "Speech emotion recognition in E-learning system by using general regression neural network", *WIT Transactions on Engineering Science,* 2014.

[2] Koolagudi, Shashidhar G., and K. S. Rao. "Emotion recognition from speech: a review", *International Journal of Speech Technology,* Vol. 15,.Issue 2, 2012.

[3] Swati Johar, "Emotional Speech Recognition", *Emotion, Affect and Personality in Speech, pp. 35-41. Springer (2016)*

[4] Haotian Guan, Zhilei Liu, Longbiao Wang, Jianwu Dang, Ruiguo Yu1, "Speech Emotion Recognition Considering Local Dynamic Features", *11th International Seminar, ISSP 2017, Tianjin, China*, October, 2017

[5] Leila Kerkeni, Youssef Serrestou , Mohamed Mbarki , Kosai Raoof, Mohamed Ali Mahjoub, "Speech Emotion Recognition: Methods and Cases Study", *10th International Conference on Agents and Artificial Intelligence (ICAART),* 2018

[6] Siddique Latif, Adnan Qayyum, M Usman, Junaid Qadir,"Cross Lingual Speech Emotion Recognition: Urdu vs Western Languages", *International Conference on Frontiers of Information Technology (FIT),* Dec, 2018

[7] Namrata Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition", *International Journal for Advance Research in Engineering and Technology,* Vol.1, Issue VI, July 2013

[8] Rajni, Dr. Nripendra Narayan Das," Emotion Recognition from Audio Signal", *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET),* Vol. 5, Issue 6, June 2016

[9] Shashidhar G. Koolagudi, Ramu reddy,Jainath Yadav , K.Sreenivasa Rao. *"IITKGP-SEHSC:Hindi speech corpus for Emotion Analysis"*, IEEE (2011)

[10] Kee Moe Han, Theingi Zin, Hla Myo Tun, "Extraction of Audio Features for Emotion Recognition System Based on Music", *International Journal of Scientific & Technology Research,* Vol. 5, Issue 06, June 2016

[11] Theodoros Iliou, Christos-Nikolaos Anagnostopoulos," Classification on Speech Emotion Recognition- A Comparative Study", *International Journal on Advances in Life Sciences,* Vol. 2 no 1 & 2, year 2010

[12] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.

[13] Baykal, N. *Speech Emotion Recognition Using Auditory Models,* Doctoral dissertation, Tesis doct. Middle East Technical University, (2013).

[14] Atassi, H., Riviello, M. T., Smékal, Z., Hussain, A., & Esposito, A. Emotional vocal expressions recognition using the COST 2102 Italian database of emotional speech. In Development of multimodal interfaces: active listening and synchrony (pp. 255-267). Springer, Berlin, Heidelberg, (2010).