IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

STUDY OF SPEECH EMOTION RECOGNITION ANALYSIS

[1]SILPA MARY ALAPATT, [2]PRAMEEJA PRASIDHAN

[1]Msc Scholar, [2]Assistant Professor

[1,2]Department of Computer Science

[1,2]St.Joseph's College (Autonomous), Irinjalakuda, Thrissur, India

ABSTRACT: Speech Emotion Recognition is a momentum research subject in light of its emotion handling. In this paper, we have done a short study on Speech Emotion Analysis alongside Emotion Acknowledgment. The initial segment of the paper is enhanced with a basic depiction. This paper incorporates the investigation of various kinds of emotions, elements to recognize those emotions and different classifiers to characterize them appropriately. Emotional speech databases, which are utilized in emotional speech recognition, are also analyzed critically. The Hidden Markov Model and Support Vector Machine were used in this study to discern human emotion through voice.

KEYWORDS: Speech Emotion Recognition, Feature Extraction, Recurrent Neural Networks, SVM**,** MFCC (Mel Frequency Cepstral Coefficient), LPCC (Linear Prediction Cepstral Coefficients), LPC (Linear Prediction Coefficients), Classifier, GMM (Gaussian Mixture Model), HMM (Hidden Markov Model), KNN (K-Nearest Neighbors), MLP (Multi Layer Perceptron), RNN (Recurrent Neural Network), MSF (Modulation Spectral Features), adaboost algorithm

INTRODUCTION:

The goal of an emotion recognition system is to emulate the mechanisms of human perception. Speech emotion recognition has a variety of applications, including decision making and physiological signal detection. A system can operate appropriately if emotion can be identified correctly from speech. Medical science, robotics engineering, call centre applications, and other fields could benefit from an effective emotion identification system. Speech Emotion Recognition manages this part of exploration in which machine can perceive emotions from speech like human. There are three ways to create an emotional speech database:

- Natural emotions: They are records of spontaneously occurring emotional states. It has a high ecological validity but suffers from a lack of available speakers, making annotation problematic.
- Simulated emotions: They are emotional states performed by professional or non-professional actors in accordance with emotion classifications or common settings. Although this method makes it simple to create an emotion corpus, it has been suggested that these emotions are more exaggerated than natural or induced emotions.
- Induced emotions: They necessitate the speaker's personal reflection on a previous incident and the instillation of the same emotion in him by recalling the complete situation of the previous incident.

The general architecture for Speech Emotion Recognition (SER) framework has three stages as shown in fig 1.

The emotional speech input to the system may be the acted data or collection of the speech data from real world situations. The relevant features were retrieved from the speech signal after the database, which was used as the training samples, was collected.

a. A speech processing system extracts relevant quantities from a signal, such as pitch or energy, for example.
b. A feature extractor, helps to reduce these numbers to a smaller number of features.
c. A classifier learns how to correlate features with emotions in a supervised manner.

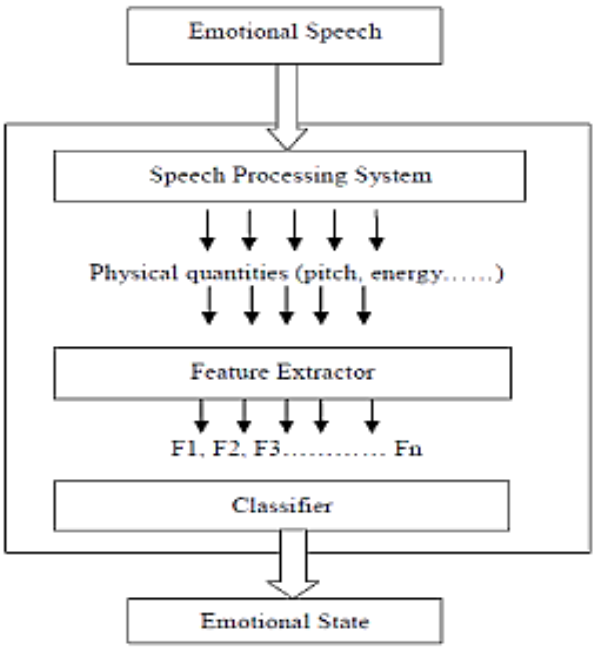

Fig1. Architecture for Speech Emotion Recognition (SER)

FEATURE EXTRACTION: Features that are separated from the vocal tract system are called system features or spectral features. The most famous spectral features are Mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs) and Perceptual linear prediction coefficients (PLPCs). The features extricated from the excitation source signal are called source features. Some source features include linear prediction (LP) and glottal volume velocity (GVV). Prosodic features are those features which are separated from long sections of speech, like, sentences, words and syllables. They are otherwise called supra-segmental features.

The splitting of speech into short intervals known as frames is the basis for feature extraction. In the SER system, selecting relevant characteristics that contain information about emotions from speech signals is a crucial step. Prosodic features, such as energy and pitch, and spectral features, such as MFCC, MEDC, and LPCC, are two types of features.

PROSODIC FEATURES: In a spoken signal, energy is the most basic and important property. We apply the short-term function to extract the value of energy in each speech frame to produce statistics of energy features such as mean value, maximum value, variance, variation range, and contour of energy. As the pitch signal is dependent on the tension of the vocal folds and the subglottal air pressure, the mean value of pitch, variance, variation range, and contour differ in each of the seven main emotional states. From the pitch, the following statistics are derived and used in the pitch feature vector:

1. Maximum, Minimum, Median, Variance, Mean, Median, Median, Median, Median, Median, Median (for the pitch feature vector and its derivative).
2. Voiced and unvoiced speech average energies.
3. Rate of speech (the inverse of the average length of the vocal section of the utterance).

SPECTRAL FEATURES: (MFCC)Mel-Frequency Cepstrum coefficients, is the most essential aspect of speech having a straightforward calculation, strong distinguishing ability, and anti-noise properties. The frequency resolution of MFCC in the low frequency area is good, and the resistance to noise is also good. MEDC extraction is comparable to MFCC extraction. The main variation in the extraction procedure is that the MEDC uses logarithmic mean of energies after Mel Filter bank and Frequency wrapping, whereas the MFCC uses logarithmic mean of energies after Mel Filter bank and of Frequency wrapping.

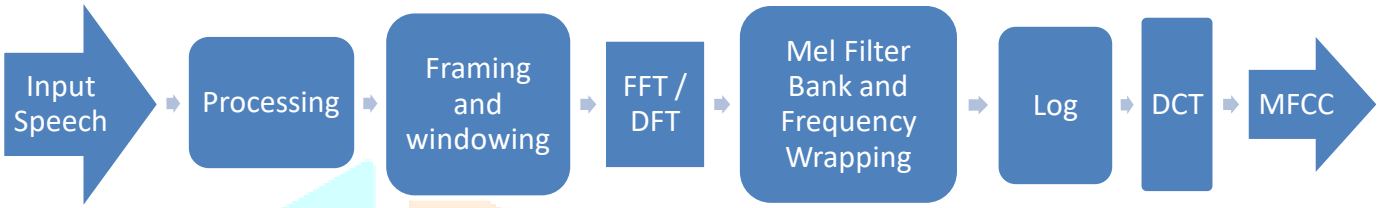

Fig 2. Schema of MFCC extraction

Mel frequency cepstrum coefficients (MFCCs) are Mel frequency cepstrum (MFC) coefficients, which are descended from power cepstrum. The Fourier Transform of the logarithm of a signal's spectrum is used to create a cepstrum. Cepstrum comes in a variety of forms, including complicated cepstrum, real cepstrum, phase cepstrum, and power cepstrum. Speech synthesis applications use the power cepstrum. Cepstrum frequency bands are linearly separated, whereas MFC frequency bands are equally spaced. As a result, MFCs can produce a more accurate representation of speech.

An auditory-inspired long-term spectro-temporal representation is used to extract modulation spectral features (MSFs). These characteristics are derived by simulating Spectro-temporal (ST) processing in the human auditory system, which takes regular acoustic frequency and modulation frequency into account. The spoken stream is initially decomposed by an auditory filter bank to produce the ST representation. The modulation signals are formed by computing the Hilbert envelopes of the critical-band outputs. The Hilbert envelopes are then subjected to a modulation filter bank for frequency analysis. The modulation spectra are the spectral contents of the modulation signals.

In the voice recognition process, LPC is one of the good signal analysis approaches for linear prediction. The most powerful method for finding the fundamental parameter and computational model of speech is LPC. LPC is based on the assumption that a speech sample can be approximated by a linear combination of previous speech samples. We may extract these feature coefficients to identify the emotions present in speech since LPCC incorporates the characteristics of a given channel of speech, and the same person with different emotional speech will have distinct channel characteristics. A recurrence of computing the linear prediction coefficients is generally used as the LPCC computational method (LPC).

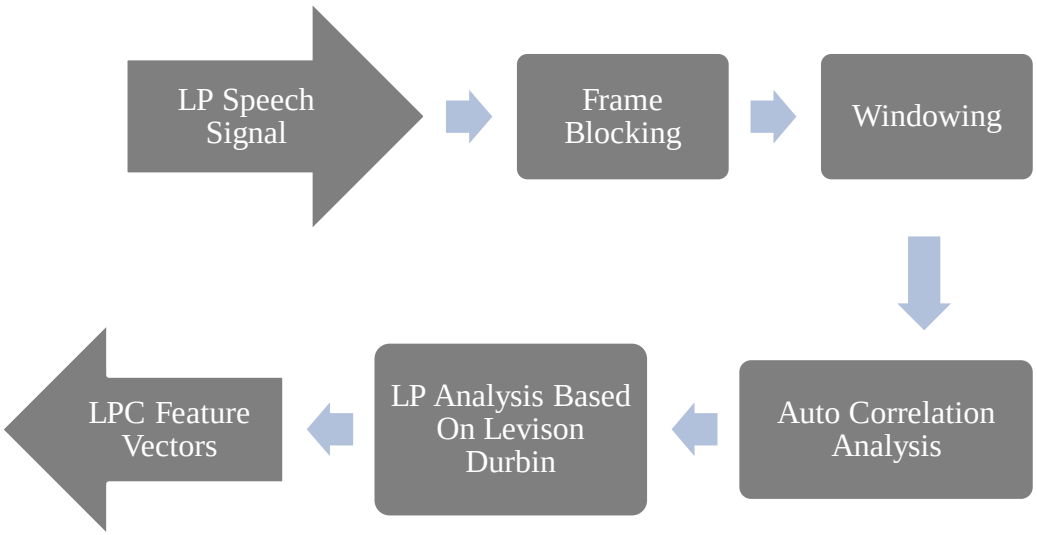

Fig. 3. Feature Extraction by LPC

CLASSIFIERS: - Humans can easily discern numerous types of emotions in the real world after years of practice and observation. A human-machine interface that can process emotional speech employs a similar idea through training and testing phases. The classifier in the interface that is supplied with emotions takes features from all the samples and creates a mixture for each emotion during the training phase. During the testing phase, the classifier captures features from the input emotional speech and compares them to all of the mixtures. The input file will be categorized into the emotion with the most features in common with the input file. Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), k-nearest neighbours (KNN), Support Vector Machines (SVMs), and the AdaBoost Algorithm are some of the classifiers now in use.

- Support Vector Machine (SVM): SVM, a binary classifier, is a simple and efficient computation of machine learning algorithms that is widely used for pattern recognition and classification problems, and it can have a very good classification performance under the conditions of limited training data when compared to other classifiers. The SVM's goal is to use kernel functions to turn the original input set into a high-dimensional feature space. As a result, this transformation can be used to address non-linear issues.

- Hidden Markov Model (HMM): Hidden Markov Models are statistical models that are used to characterize event sequences. The HMM is made up of a first-order markov chain whose states are hidden from the observer, hence the model's internal behaviour is kept hidden. These model states represent the data's temporal structure. For HMM-based emotion recognition, the database is first sorted by classification mode, and afterwards the features from the input waveform are extracted. The database is then updated with these features. The modes were used to create the transition matrix and emission matrix, which generates the random sequence of states and emissions from the model. Finally, the Viterbi algorithm is used to estimate the state sequence probability.

- Gaussian Mixture Models (GMMs): Gaussian Mixture Models (GMMs) are commonly used to assess density and perform clustering. For this, the expectation-maximization technique is applied. Gausses are the component functions that constitute GMMs. The number of components in the mixture model refers to the number of these Gausses. Based on the quantity of training data points, the total number of components can be changed. However, as the number of components grows, the model becomes more complicated.

- K Nearest Neighbour (KNN): The Nearest Neighbour approach is the most classic of the supervised statistical pattern recognition methods; it does not make any a priori assumptions about the distributions from which the training samples are obtained. It entails a training set that includes all possible scenarios. The distance between a fresh sample and the nearest training instance is calculated, and the sign of that point determines the sample's categorization. Larger K values assist lessen the effects of noisy points in the training data set, and cross validation is frequently used to choose K.

- AdaBoost Algorithm: The AdaBoost algorithm is a type of adaptive classifier that generates a strong classifier from a weak one iteratively. The weak classifier is used to classify the data points in the training data set in each iteration. Initially, all data points are assigned equal weights, but as the classifier iterates, the weight of improperly categorized data points rises, causing the classifier to focus more on them in the following iteration. As a result, the Classifier's global error decreases, resulting in a stronger classifier.

- Recurrent neural Networks (RNN): Time series data can be learned using recurrent neural networks (RNN). While RNN models are good at learning temporal correlations, they have a vanishing gradient problem that gets worse as the training sequences get longer.

- A neural network model called a Multi Layer Perceptron (MLP) is used to derive a suitable output from a set of input data. MLP has three layers: input layer, hidden layer, and output layer, with the hidden layer that may have many instances. The MLP model is built as a connected graph, with each layer's nodes connected by a weighted edge to the nodes of the following layer. There are several nodes in each layer. Each node performs two functions that is input and output. Back propagation is a supervised learning approach used by MLP to train the network.

Table 1. Literature review on use of different classifiers in different research

Study group and year	**Database**	**Extracting feature**	**classifier**	**Accuracy rate**
X. Cheng et al (2012)	Mandarin emotional speech database	Pitch, MFCC	GMM	79.9% (female) 89.02% (male)
T. Seehapoch et al (2013)	Berlin, Japanese and Thai	F0, Energy, ZCR, MFCC	SVM	89.8%
Li Liu et al(2014)	Chinese tweets from Sina Weibo	MI, CHI, TF-IDF and ECE.	HMM	78% accuracy
Monorama Swain et al(2015)	Odiya Speech sambalpuri, cuttaki	MFCC, Delta MFCC, LFPC	SVM	They got 82.14% accuracy with SVM only using MFCC.
Akash shaw (2016)	Real users engaged with a machine agent.	Energy, Pitch, Formant frequency, MFCC	ANN	They got 85% accuracy
Raviraj Vishwambler(2017)	Marathi Speech Database	Cepstral features, NMF, Pitch	ANN	78%
Vishnu Vidyadhara Raju Vegesna et al(2018)	Telegu speech corpus	MFCC, Pitch, Intonation	GMM HMM	75% recognition rate for all observation

S.S Poorna et al(2018)	South Indian language	LPC, Energy	AN, SVM KNN	98.32% 94.84% 81.75%

RESEARCH PROBLEMS AND PROJECT IDEAS: -

- The majority of study findings on emotional speech recognition have relied on databases with a small number of speakers, which could lead to the usage of speaker-specific information such as speech utterances from the same speakers being used to train and test models. These generated models may yield poor results due to a lack of generality. As a result, a larger emotional speech database with a reasonable number of speakers and text cues is required. Because of the differences in speaker, text, and session, emotion identification experiments must be conducted on huge databases.

- Emotional expression is a universal occurrence that can occur regardless of speaker, gender, or language. A study of cross-lingual emotion identification could be another promising area for future research. The emotion recognition models created with a certain language's utterances should work admirably well with any test speech in a different language. One can organise languages based on their emotional similarity utilising trans lingual emotion analysis.

- In most studies, a single model (e.g., GMM, AANN, or SVM) is used to complete the emotion categorization job. Hybrid models can be investigated to see how well they perform in the area of emotion recognition. The primary principle behind adopting hybrid models is that they derive knowledge from many perspectives, and so the combination of evidence can improve performance if the evidence is complementary.

- Verification of emotions is important in real-time applications such as call analysis in emergency services such as ambulance and fire departments to verify the sincerity of demands. In this regard, appropriate features and models might be investigated within the framework of emotion verification.

EXPERIEMENTAL STUDY: - The system receives emotional speech input by recording voice samples from diverse speakers. The male and female speakers recorded voice samples in five different emotions: anger, happiness, sadness, surprise, and neutral, which will be detected by a speech emotion detection system. These speech samples were recorded using a common comment. The various voice samples were collected for the purpose of training the classifiers and system testing.

- Experimental Results using HMM: The database is initially sorted according to the mode of classification, in this example 5 for five modes, for emotion recognition using Hidden Markov Model (HMM). The features from the waveform input are then extracted. The database now includes these characteristics. The modes were used to create the transition matrix and emission matrix, which generates the random sequence of states and emissions from the model. Finally, the viterbi algorithm is used to estimate the state sequence probability. The matching of mode with the database is defined by the probability of this HMM, and the result tag can be labelled as the most match mode.

Table 2. Emotion Recognition Rate using HMM

Emotions	Recognized emotions(%)				
	Anger	Happy	Sad	Surprise	Neutral
Anger	83.33	16.67	0	0	0
Happy	0	57.14	14.29	28.57	0
Sad	0	0	62.50	12.50	25.00
Surprise	28.57	0	0	71.43	0
Neutral	0	0	25.00	0	75.00

- Experimental Results using SVM: First and foremost, the above-mentioned required qualities are calculated. All of the values calculated in the preceding phases will be fed into the Support vector machines, which will be used to train the classifier. The values of the features of the testing voice sample are calculated once again by the Support vector machines. The trained voice sample is then compared to the testing voice sample based on attributes collected from the testing voice sample. Support vector machines will detect the smallest difference between the trained and test voice samples during the comparison. The emotion will be recognised using this differential SVM classifier.

Table 3. Emotion Recognition Rate using SVM

Emotions	Recognized emotions(%)				
	Anger	Happy	Sad	Surprise	Neutral
Anger	71.42	14.29	0	14.29	0
Happy	0	57.14	14.29	28.57	0
Sad	0	0	71.43	0	28.57
Surprise	23.39	14.28	0	63.33	0
Neutral	0	0	25.00	0	75.00

CONCLUSION: -

Nowadays there is increasing interest in speech emotion recognition research through applications like call centre analytics, human-machine and human-robot interfaces, multimedia retrieval, surveillance activities and behavioural health informatics. In this paper, an overview of SER approaches for extracting audio features from speech samples is covered, as well as a brief explanation of several classifier algorithms. The accuracy of speech emotion recognition is dependent on the emotional speech database, the combination of features extracted from those databases for training the model, and the types of classification algorithms used to classify the emotions into appropriate emotion classes (e.g., happy, sad, anger, surprise, etc.). The study also discusses some major research issues and scope in the field of speech emotion recognition. As the accuracy of the system is greatly dependent on the emotional speech database utilized in it, it is critical to record accurate emotional speech databases.

REFERENCE: -

1. Saikat Basu, Md. Aftabuddin, Jaybrata Chakraborty, and Arnab Bag, "A review on emotion recognition using speech" (ICICCT 2017), *International Conference on Inventive Communication and Computational Technologies,*

2. Mohamed Ali Mahjoub, Kosai Raoof, Youssef Serrestou, Leila Kerkeni, and Mohamed Mbarki, 2018, "Speech Emotion Recognition: Methods and Cases Study", *10th International Conference on Agents and Artificial Intelligence.*

3. Akalpita Das, Prof. P.H. Talukdar, Laba Kr. Thakuria, and Purnendu Acharjee, "A brief study on speech emotion recognition," *International Journal of Scientific & Engineering Research, Volume 5, Issue 1, January-2014,* ISSN 2229-5518.

4. Ratnadeep R. Deshmukh, Dr. Vishal Waghmare, Shaikh Nilofer R., Rani. P. Gadhe, and Pukhraj P. Shrishrimal , "Emotion Recognition from Speech: A Survey", *International Journal of Scientific and Engineering Research, Volume 6, Issue 4, April-2015.ISSN 2229-5518..*

5. A. Bag, M. Mahadevappa, J. Mukherjee, S. Kumar, S. Basu, N. Jana, S. Kumar, and R. Guha "Emotion recognition based on physiological signals utilising valence-arousal model," *In Image Information Processing (ICIIP), 2015 Third International Conference on, pages 50–55. IEEE, 2015.*

6. S. Maity, V. A. Kumar, S. Chakrabarti, K. S. Rao and S. G. Koolagudi, "IITKGP-SESC: Speech Database for Emotion Analysis," *Communications in Computer and Information Science, JIIT University, Noida, India: Springer, ISSN: 1865-0929 edition August 17–19 2009.*

7. K. S. Rao and S. G. Koolagudi, "Exploring speech features for classifying emotions along valence dimension," *in Springer LNCS (S. C. et al., ed.), (IIT Delhi), 3rd international Conference on Pattern Recognition and Machine Intelligence (PReMI- 09), pp. 537–542, Springer-verlag, Heidelberg, Germany, December 2009.*

8. Lawrence R Rabiner and B H Juang, "Fundamentals of Speech Recognition", *Englewood Cliffs, NJ: PTR Prentice-Hall, 1993*

9. S. Basu, A. Bag, M. Mahadevappa, J. Mukherjee, and R. Guha, "Affect detection in normal groups with the help of biological markers", *In India Conference (INDICON), 2015 Annual IEEE, pages 1–6, . IEEE, 2015.*

10. Dr. Purnendu Bikash Acharjee and Somi Kolita, "Speech emotion Recognition using Non-linear Classifier- A Review", *International Journal of Engineering Research & Technology, Volume 08, Issue 05, May 2019.*

11. Gerhard Rigoll, Björn W. Schuller, and M. Lang, "Hidden Markov model based voice emotion recognition", *Proceedings of the IEEE ICASSP Conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 1-4, April 2003.*

12. Norhaslinda Kamaruddin and Abdul Wahab, "Features extraction for speech emotion", *Journal of Computational Methods in Science and Engineering, vol. 9, no. 9, 2009, pp. 1–12. ISSN: 1875-8983 (Print) and 1472-7978.*

13. Milton A., S. Sharmy Roy and S. Tamil Selvi (2013), "SVM Scheme for Speech Emotion Recognition using MFCC Feature", *International Journal of Computer Applications, Volume 69, Number 9.*

14. Alex Graves and Navdeep Jaitly, "Toward end-to-end speech recognition with recurrent neural networks", (2014) *International Conference on Machine Learning, 32*.

15. Ashish B. Ingale and Dr.D.S.Chaudhar, "Speech Emotion Recognition Using Hidden Markov Model and Support Vector Machine", *International Journal of Advanced Engineering Research and Studies, Volume 1, Issue 3, April 2012 E-ISSN 2249–8974.*

16. Shashidhar G. Koolagudi and K. Sreenivasa Rao, "Emotion recognition from speech: A review", *International Journal of Speech Technology, 15(2):99–117 (2012).*

17. R. R. Deshmukh and Kishori R. Ghule, "Feature Extraction Techniques for Speech Recognition: A Review", *International Journal of Scientific & Engineering Research, Volume 6, Issue 5, May-2015 143 ISSN 2229-5518.*

18. S. R., Gunn(1998), "Support Vector Machines for Classification and Regression", *PhD thesis*.

19. Dr. B. Venkateshulu, P. Chandrasekhar, and G. S. D. Sree (2016), "SVM Based Speech Emotion Recognition Compared with GMM-UBM and NN", *IJESC, 6.*

20. M. S. Kamel, F. Karray, and M. E. Ayadi, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", *Pattern Recognition Volume 44, Issue 3, pp.572-587, 2011*.

21. D. S. Chaudhari and Ashish B. Ingale, "Speech Emotion Recognition", *International Journal of Soft Computing and Engineering (IJSCE), ISSN 2231-2307 Vol.2, Is.1, March 2012.*