# Speech Recognition: A Review on Its Methodological Aspects

[1]Kris Negi, [2]Ishika Sindhi, [3]Komal Garg, [4]Khushi Gambhir
[1234]Chitkara University Institute of Engineering and Technology,
Chitkara University, India

***Abstract :*** Human emotion recognition through machine intelligence has been a topic of interest from many years. Mechanisms such as speech processing, pattern recognition and machine learning has made it possible for a machine to understand the emotional state of a person with accuracy and simplicity. Features of speech such as pitch, wavelet, Mel Frequency cepstrum coefficients (MFCC) are used by these machines to classify the human emotions into several classes such as sad, anger, happy, neutral, etc. This paper features a survey of the various researches carried out on the speech emotion recognition (SER) domain. A brief explanation to the various stages of SER, classifiers such as K-Nearest Neighbour (K-NN) and Gaussian Mixture Model (GMM) and the performance analysis, using parameters such as F-measure, precision, recognition rate, etc. has been provided in this survey.

**Keywords - Speech recognition system, Machine learning, Gaussian mixture model, KNN, MFCC, Pitch.**

## I. Introduction

Human emotion can be considered as a mental state of a person which is often related with the person's mood or feelings. It plays a crucial role in defining personality and has a huge impact on the person's ability to handle certain situations. Making a machine to understand the human emotion makes the human-machine interaction less synthetic and more natural. In the recent years, there has been a huge progress in the human-machine interfaces, such as, progress in processing audio and visual data, in recording and storing information and much more. The technology sector has gone through a huge change and has seen massive advancements in the field of data science, artificial intelligence and machine learning. The sensors used nowadays have become more efficient in sensing and processing its surroundings. Even the machines, such as wearable computers, have become less of a burden on its users as a consequence of their high portability and intelligence. All these major advancements play a huge rule in helping humans in making a machine capable of understanding it's creator's feelings and thus making the human-computer interactions more friendly.

Human emotions can be synthesized through skin temperature, gesture recognition, facial expressions by using a high quality camera and similar devices. But using these methods for emotion recognition can be very complex and costly. Speech, on the other hand, can provide better results and can easily classify different human emotions without much complexity involved in the process. Speech is the primary medium used by humans for social interactions [1]. Using speech as a medium can provide a good research analysis in low cost. The tools used nowadays for speech emotion recognition (SER) include the signal processing toolbox of the MATLAB software, pattern recognition algorithms such as, GMM, K-NN, KMC among the many others.

Emotion recognition through speech plays a huge role in our day-to-day lives. It finds many applications in various fields like interactive systems, business, education, virtual reality, entertainment and medicine. The most common application is its use in smart/automated homes and call centres. In the field of medicine, SER is being used for automatic detection of fatigue, stress or depression, to identify an individual's personal wellness and to provide psychiatric aids. It serves as a great help for people with visual disorders. With more advancements, emotion recognition is taking a step further into tutoring systems and interactive games. To improve the performance of SER systems, emotion conversion can be employed, which converts the emotional features in a speech and regenerates the final output using new parameters. It uses data driven voice conversion techniques for spectral transformation. This paper is oriented towards discussing such techniques. The contents of this paper are divided as follows: The literature survey is discussed in section 2. Section 3 explains the speech emotion recognition systems. Section 4 provides a brief introduction to emotional speech database. Section 5 describes the feature extraction stage. Section 6 discusses the feature classification. The evaluation parameters are provided in section 7. Section 8 concludes the paper with a mention to future works on the research.

## II. Literature Survey

Busso et al. carried out two experiments in which one compared the emotional pitch with the neural speech while the other experiment concludes the importance of pitch in classifying different emotions. The latter measured the discriminative power of pitch for emotion classification [3]. As provided in [2], the work by Farooq et al. outlines the multi-resolution capabilities of wavelet packet transform and how it can derive the spectral features better than MFCCs, leading to an improved recognition. The biomedical research by Yao et al. generated a Bionic Wavelet Transform (BWT) which established an active control over auditory system. This led to an increase in selectivity and sensitivity of the recognition system. The clinical research by Wu et al. gave a variety of classifiers to plot the test data. It uses multiple classifiers like Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) in a single system. But the GMMs provide better results than the HMMs for Berlin Emotion Database (BES). This experiment resulted in an improvement in recognition. A review research by Ververidis et al. found that a classification of above 50% cannot be achieved for the 4 basic emotions in an automated system. The research describes how natural emotions are harder to classify as compared to simulated emotions. The ease of classification of emotions in increasing order is as follows: fear, happiness, sadness and anger. The basic objective of the emotion recognition systems is to enhance the human-machine interaction and to provide a better user experience. The next section discusses about the speech emotion recognition systems.

## III. SPEECH EMOTION RECOGNITION SYSTEMS

Speech emotion recognition can be achieved through a variety of approaches. Acousticphonetic approach is one such approach and is based on mechanism of feature extraction. Pattern Recognition is another approach that focuses on template matching algorithms. In the Artificial Intelligence approach, algorithms are dependent upon neural networks, stochastic of speech signal and the knowledge source. The Machine Learning approach is used to test the classifier's performance by training the classifier to learn the different human emotions with the help of extracted speech features. Another important approach is the Stochastic Modelling which uses hidden markov models. Algorthms also plays a major role in speech emotion recognition concept. Dynamic time warping algorithm is one of the most popular and accurate template-based algorithms used for emotion recognition [6].

The speech recognition process consists of six main stages which are also known as the pattern recognition cycle. It includes data set preparation in which the speech signals are obtained using a microphone and the identification of emotion classes takes place. Front end processing is used to pre-process the input speech signal and any detected noise in the input is deleted using the wavsurfer tool. Feature extraction is considered as one of the most crucial stage in the speech recognition process. It extracts all the important features from the pre-processed speech signal. The processes under this stage depend on various signal processing techniques such as, filter banks, linear predictive coding, cepstral techniques, etc. Feature selection stage plays an important role in limiting the feature sets and only the relevant features are selected for further processing. The selected features are passed on for classification which uses classifiers like Gaussian Mixture Model (GMM) and K-Nearest Neighbour (K-NN). The last stage in the cycle is emotion detection where the emotion is detected from speech input with the help of classifiers [2]. The accuracy and efficiency of speech recognition is affected by the surroundings of the system. Noise being a major obstacle, can alter the system's output and can have a negative impact on the efficiency of recognition systems.

Feature sets are a crucial part of the emotion recognition cycle but there are certain properties which a good feature set should have. They should help the system to easily distinguish between different speech signals and should work effectively without the need of excessive training data. Besides they should not get affected by the surroundings, that is, they should be immune towards noise. The emotion classification rate largely depends on the features used, categorization of emotion by classifiers, and the speech emotion database used.

It is imperative to have a collection of data stored as a repository for these algorithms as a pre-requisite for testing and experimentation purpose. Therefore, the next section discusses the databases in the context of Emotional Speech.

## IV. EMOTIONAL SPEECH DATABASE

A collection or record of speech emotion data sets is very crucial and it is useful for further studies and researches in the speech emotion recognition domain. It is difficult to maintain such a large data collection but a lot of work is being done in the perpetuation of such databases. The three groups of speech that can be found in such database are, natural, simulated and elicited. Natural speech refers to a speech in which all emotions are real. In a simulated or acted speech, speech is expressed in a professional manner, whereas, an elicited speech contains induced emotions [10]. The standard speech database used to extract various speech features are as follows, Berlin Emotion Speech Database (BES), Surrey Audio Visual Expressed Emotion (SAVEE), Danish Emotion Database (DES), Speech Under Simulated and Actual Stress (SUSAS) [11]. Out of the above mentioned databases, Berlin Emotion Speech Database (BES) is the most popular database used for speech processing as it contains voice recordings of 5 male and 5 female actors. The recordings represent 7 emotional states, namely, neutral, fear, happiness, disgust, anger, surprise, sadness. The database contains around 800 sentence records of twenty contestants and has a naturalness score of above 60% [2]. The next section discusses the process of feature extraction.

## V. FEATURE EXTRACTION

Feature extraction is the most important stage in the speech emotion recognition (SER) systems. It consists of three main stages which are used to extract feature sets from an input speech signal. The first stage or the acoustic front end helps in speech analysis. The second stage compiles the extended feature vectors containing static and dynamic features. The third stage transforms the extended feature vectors to compact vectors and gives them to the recognizer. Features can be classified into acoustic and linguistic, out of which acoustic features are the most widely used [10]. The following features are extracted during this stage: pitch, frequency, energy, MFCCs, wavelet, etc and are explicated below:

### 5.1 Mel Frequency Cepstrum Coefficients (MFCC)

MFCC is one of the most popular spectral feature in use nowadays. Mel measures the recognition of frequency or pitch of a tone. Mel scale maps the actual frequency scale to perceived frequency scale. Mel scale is also used to model the speech recognition process similar to that of the human ear, that is, linearly below 1000 Hz but logarithmically above [9]. MFCCs are more insensitive to noise and give better recognition results as compared to other parameters. The extraction of MFCC feature follow a series of steps. First, the speech signal is pre-processed using windowing techniques and the Discrete Fourier Transformation of the framed signals is taken. Later, frequency wrapping is performed on the acquired magnitude spectrum to obtain the Mel scale which is further used to obtain uniformly spaced triangular filter banks. The filters get multiplied with magnitude spectra to acquire the MFCC features [4].

### 5.2 Pitch Features

Pitch, also known as the fundamental frequency (F0), is a very popular prosodic feature. It can be defined as the vibrational rate of vocal fold. Pitch is extracted using Subharmonic-to-Harmonic Ratio (SHR) which can be computed using spectrum compression techniques. The calculated SHR is then compared with pitch perception data to obtain the pitch features [5].

### 5.3 Zero Crossing Rate (ZCR)

ZCR is the measure of the number of times a speech signal transitions from positive to negative or vice versa. It can also be defined as the rate of sign changes in a speech signal. It finds huge application in fields like music information retrieval, voice activity detection (VAD) and speech recognition because of its ability to classify percussive sounds.

### 5.4 Linguistic Features

These features consist of spoken or written text which can be used to distinguish emotions. It focuses on the usage of particular words and grammar to help differentiate between the emotional states. The various approaches used for these type of analysis are semantic trees, keyword spotting, Bayesian networks, etc. [10]. The next section discusses on the process of Feature Classification.

## VI. FEATURE CLASSIFICATION

The classifiers are trained on data sets which in turn helps in predicting the emotional state. The features extracted from the speech signal form feature vectors and these vectors are mapped onto a data science model through learning. K-Nearest Neighbour (K-NN), Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) are some examples of the classifiers used in data science. This paper focuses on the K-NN and GMM classifiers and are presented below:-

### 6.1 Gaussian Mixture Model (GMM)

GMM is a type of probabilistic model which is used to represent normally distributed subpopulations amidst the overall population. This model is based upon the concept of unsupervised learning as it is not required to know which class a data set pertains to. It helps the model in learning subpopulations automatically. Gaussian mixtures model the emotions as a Gaussian density mixture where the Gaussian components represent spectral shapes that are emotion dependent. They can be used to represent arbitrary densities. The data in GMM looks multimodal, which means that the distribution of data contains more than one peak. Since a multimodal distribution can be assumed as multiple unimodal distributions, Gaussian distribution can also be used to model real-world unimodal data. Due to its ability to model very large data sets, GMMs are used extensively in feature extraction from speech signals and object tracking [12].

### 6.2 K-Nearest Neighbour (K-NN)

This algorithm classifies an unknown sample based on its K nearest neighbours. The nearest neighbours are calculated using the Euclidian distance from the unknown data point. Being a non-parametric algorithm, the probability distribution of input is not assumed by k-NN which makes it a useful algorithm when the input properties are unknown. This makes k-NN more robust as compared to other parametric algorithms. Furthermore, k-NN is a lazy learning algorithm, that is, in this algorithm data is generalized in the testing phase, rather than in the training phase. This increases the computation at the testing phase but it gives k-NN an added benefit of quick adaptiveness to change. Due to its effectiveness and simplicity, k-NN finds huge application in the field of machine learning which includes, gene expression analysis, handwritten letter identification, computer vision, and so on [13].

## VII. EVALUATION PARAMETERS

This section focuses on the parameters that are used to test a classifier's performance or are used to compare one classifier with another based on their efficiency, ease of use and accuracy. The results obtained from the training phase are cross validated using such parameters. Recognition accuracy, precision rate, recall, F-measure are some parameters used for such purposes.

### 7.1 Recognition accuracy

It is the accuracy of each known speech emotion input to the total trained speech data in percentage [7].

### 7.2 Precision Rate

It is a fraction of correctly recognized emotion per class to the correctly recognized emotions for all classes or it denotes the probability of truly detected emotions [7].

### 7.3 Recall

Recall denotes the probability of actually detected emotions [8].

### 7.4 F-Measure

It can be considered as a combination of accuracy and precision rate. It is used to obtain the total performance of a system based on the correct results and not by the incorrect recognition observations [7].

In all the four evaluation parameters mentioned above, the GMM classifier stands superior than the K-NN algorithm for all the emotion classes [2].

## VIII. CONCLUSIONS AND FUTURE WORKS

As per the survey conducted, Gaussian Mixture Model gives the best result for 'anger' emotion, with a high recognition accuracy and the worst result for 'surprise' with the least accuracy. On the other hand, K-NN shows the highest recognition for 'happy' emotion with a large accuracy and the 'fear' and 'surprise' gets recognized at the lowest rate. Furthermore, the precision, recall and F-measure evaluation parameters show the supremacy and robustness of GMMs over K-NN algorithms.

The future works on this research will include a more practical approach towards the speech recognition systems by taking various aspects, such as language and text into consideration. A more detailed study on the recognition systems will be done to make the future research more relatable and realistic as possible. An attempt to combine two classifiers together will be made to get the best of both in one system.

## REFERENCES

**[1]** M. ElAyadi, M.S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", Pattern Recognition, vol. 44, no.3, pp. 572–587, March 2011.

**[2]** Rahul B. Lanjewar, Swarup Mathurkar and Nilesh Patel, “Implementation and Comparison of Speech Emotion Recognition System using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques”, Procedia Computer Science 49(2015), pp. 50-57.

**[3]** Chang-Hsein Wu and Wei-Bin Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic and Semantics Labels”, IEEE Trans. On Affective Computing, Vol 2, No.1, 2011, pp. 567-569.

**[4]** Rahul B. Lanjewar, D.S. Chaudhari, “Speech Emotion Recognition: A Review”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 2, March 2013, pp. 68-71.

**[5]** Xuejing Sun, ‘Pitch Determination and Voice Quality Analysis using Subharmonic-To-Harmonic Ratio’, Department of Communication Sciences and Disorders, Northwestern University, 1999, pp. 561-563.

**[6]** Anusuya M., Katti S. Speech Recognition by Machine: A Review, Department of Computer Science and Engineering Sri Jaya chamarajendra College of Engineering Maysore, India, 2009, pp. 181-205.

**[7]** Stavros Ntalampiras and Nikos Fakotakis, ‘Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition’, IEEE Trans. on Affective Computing, Vol. 3, No. 1, 2012, pp. 116-125.

**[8]** Er. Ubeeka Jain, Amandeep Sandu, “Emotion Detection from Punjabi Text using Hybrid Support Vector Machine and Maximum Entropy Algorithm”, International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), Vol. 4, Issue 11, November 2015, pp. 89-93.

**[9]** Ghazi A., Daoui C., Idrissi N., Fakir M., Bouikhalene B. Speech Recognition System Based On Hidden Markov Model Concerning the Moroccan Dialect DARIJA. Global Journal of Computer Science and Technology, 2011: 975-981.

**[10]** S. Ramakrishnan, Ibrahiem M.M. El Emary, “Speech Emotion Recognition Approaches in Human Computer Interaction”, Springer, September 2011, pp. 1467-1478.

**[11]** D. Kaminska, A. Pelikant, “Recognition of human emotion from a speech signal based on Plutchik’s model”, International Journal of Electronics and Telecommunications, vol. 58, no. 2, pp. 165-170, June 2012.

**[12]** John McGonagle, “Gaussian Mixture Model”, retrieved from https://brilliant.org/wiki/gaussian-mixture-model/

**[13]** Akshay Padmanabha and Christopher Williams, “K-nearest Neighbors”, retrieved from https://brilliant.org/wiki/k-nearest-neighbors/