iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** IV **Month of publication:** April 2021

DOI: https://doi.org/10.22214/ijraset.2021.33656

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue IV Apr 2021- Available at www.ijraset.com

Speech Emotion Recognition: A Review Paper

Kanishka[1], Dhara Dhakad[2], Natasha Doshi[3], Mayank Sohani[4]
[1, 2, 3]*Undergraduates,* [4]*Assistant Professor, Dep. Of Computer Engineering, MPSTME, NMIMS, Shirpur Campus, Maharashtra, India*

Abstract: ***In the field of HCI, one of the most common researched topics is speech emotion recognition. Many researchers are working on systems to classify different emotions from human expression. This is done in order to make HCI and human interfaces more efficient and productive, as well as to build systems that behave intelligently like humans. We conducted an experiment to see if we could recognise emotions from human expression. Neutral, rage, excitement, and sorrow were among the emotions shown for the experiments.***
Necessary work has been done on this topic and carefully examination is also being done. We have researched on multiple algorithms which include CNN, SVM, FFT, and MFCC. We examined the fundamentals of a speech emotion recognition system and explored various pre-processing, feature extraction, and classification techniques for the system in this paper. Elicited, Prosodic, and Spectral features are the three types of features. The classifications were carried out for a variety of classifiers. Hidden Markov Model (HMM), Gaussian Mixtures Model (GMM), Support Vector Machine (SVM). Artificial Neural Network (ANN), and K-nearest neighbour (KNN) are some of the techniques used to distinguish various emotions from human expression. Dataset collection is one of the most important tasks. A lot of work has been done on collecting datasets on emotion recognition. In this research, we researched on various datasets like interactive emotional motion capture (IEMOCAP), MELD etc.

Keywords: K-nearest neighbour, utterance level, Speech emotion recognition, artificial intelligence, SVM, Deep Neural Network, CNN, RNN, ANN, MFCC, MMSE, KNN, GMM, FFT

I. INTRODUCTION

Human machine interaction is greatly used in many applications [1]. One of the many media of interaction is speech. For the past decades, psychologists from all around the world have their attention on understanding human emotions. Humans use voice and vocabulary to communicate verbally [2]. This allows for faster message exchange, concept transmission, and innovation dissemination.

Human communication encompasses not just what people say, but also how they say it. Emphasizing on a word changes the whole meaning of the sentence/command. Furthermore, nonverbal communication includes facial gestures, voice intonation, and actual words, which account for around 55%, 38%, and 7% of the message perception respectively. According to scientists, researching and evaluating feelings helps in the treatment of mental illnesses, which is an important aspect of universal human health care [3]. Sensing emotion from expression is a big obstacle in human-machine interaction. Since emotions affect processes like perception, focus, learning, memory, and decision-making, they have a significant influence on human behavior. Speech messages affected feelings as well as words and meanings.

Expression of emotion. The retrieval of the acceptable emotional state of a person's speech from that person's speech is identified as recognition. When two people communicate with each other, they can easily detect the underlying emotion in the other person's voice.

Speech signal processing is commonly used in a range of applications, including remote patient monitoring, non-contact medical diagnosis, and human- computer interaction. Speech emotion detection involves analyzing a speech signal for characteristics such as pitch, formant, and phoneme training to assess the correct emotion [4]. For feature extraction and testing of a speech signal, several algorithms have been tested.

Few of them are Artificial neural networks [5], linear prediction cepstrum coefficients, Mel Frequency cepstrum coefficients, combination of Linear Prediction coefficients and Mel Cepstrum coefficients, the Support Vector Machine: combination of HMM and SVM etc. [6]. Speech

signals can be interpreted to determine heart rate, allowing for remote diagnosis of a patient using only audio data. The detection of nasopharyngeal and vocal tract abnormalities is one of the medical applications.

©IJRASET: All Rights are Reserved

II. LITERATURE REVIEW

S.no	Methodology	Brief Description
1.	Acoustic signal processing, Human-computer interaction, Linguistic speech features, Affective computing, NSL	The aim of this paper is to highlight the various techniques for detecting emotional states in vocal expressions by providing a brief overview of the current state of research in this area [1]. Methods for extracting speech features from speech datasets, as well as machine learning methods based on classifiers, are also investigated.
2.	MFCC, FFT	Emotions are recognized in this paper using data from voice signals. The Mel Frequency Cepstral Coefficient (MFCC) technique was used to recognize the speaker's emotions from their expression.
3.	MFCC, ZCR, TEO, HNR, SVM	This paper proposes an emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine. Firstly, two sets of features are investigated which are: the first one, we extract 42-dimensional vector of audio features including coefficients of MFCC, ZCR, HNR and TEO. And the second one: we selected from the parameters previously extracted. Secondly, we use the SVM as a classifier method.
4.	Mel Frequency Cepstral Coeffiecient, Pattern Recognition Neural Network Gray Level Co-occurrence Matrix	In this analysis, a combination of PRNN and KNN algorithms is used to create an improved human speech emotion recognition system [4]. The six basic emotions of neutral, anger, pleasure, sorrow, surprise, and fear are listed and studied for their consistency over the speech emotions using other previously developed systems.
5.	The liquid state machine, The source-filter speech production model, HMM	This paper introduces a method that works directly on the speech signal, obviating the need for the time-consuming feature extraction stage. This approach also combines the advantages of the traditional source-filter model of human speech development with those of the recently introduced liquid state machine (LSM), a biologically influenced spiking neural network (SNN).
6.	Mel spectrogram, harmonic percussive, chromagram, mel frequency cepstral, GLCM	The paper's main idea is to use deep neural networks (DNN) and k- nearest neighbor (k-nn) to acknowledge emotion from voice, especially in a scary state of mind. The study of this concept has shown that this approach continues to play an important role in the medical and technological fields.
7.	DNN, CNN, HSF-DNN, LLD-RNN, MS-CNN	In this analysis, we used a deep neural network (DNN), a convolutional neural network (CNN), and a recurrent neural network (RNN) to create system that combined three different classifiers (RNN). The method was used to categorize and understand four different emotions. Frame- level low-level descriptors (LLDs), segment-level Mel-spectrograms (MS), and utterance-level outputs of high-level statistical functions (HSFs) on LLDs were provided to RNN, CNN, and DNN. Three separate models were created: LLD-RNN. MS-CNN, and HSF-DNN [7].
8.	MFCC, LPC, LPCC, LSF, PLP, DWT [8]	MFCC, LPC, LPCC, LSF, PLP and DWT are some of the feature extraction techniques used for extracting relevant information form speech signals for the purpose speech recognition and identification. These techniques have stood the test of time and have been widely used in speech recognition systems for several purposes.

©IJRASET: All Rights are Reserved

9.	HMM classifier, MFCC	An HMM-based approach to emotion and speaking style identification is discussed in this paper. The SUSAS database [9] is used to evaluate the proposed feature extraction method's accuracy. For ten different style groups, we assess classification accuracy.
10.	Multilayer perceptron neural network, NSL	This paper explores the influence of age and gender on emotion recognition applications. The results of four different models were compared [10], and an association between age, gender, and emotion recognition accuracy was shown.
11.	Mel-Frequency Cepstral Coefficients	This paper presents the results of an experiment on understanding emotions in human speech. One of the main feature attributes considered the prepared dataset was the peak-to-peak distance obtained from the graphical representation of the speech signals.
12.	MFCC, GMM, KNN, GLCM	The spectral components of Mel Frequency Cepstrum Coefficients (MFCC), wavelet features of voice, and the pitch of vocal traces are used in this paper to implement a speech emotion recognition method.
13.	Connectionist context modeling, Polyphone context classes, PLP	This paper introduces a context-dependent hybrid connectionist speech recognition method that employs a collection of generalized hierarchical mixtures of experts (HME). The connectionist part of the framework is built in a modular fashion, allowing for distributed training on standard workstations.
14.	ACNN, LPCC	We present our findings in this paper on how representation learning on large unlabeled speech corpora can be used to improve speech emotion recognition (SER).
15.	CNN, VGGish, Log Mel-spectogram, GLCM	Current emotional datasets in the domain of speech emotion recognition (SER) have an unbalanced data distribution of emotional samples. Furthermore, various fragment areas in an utterance lead to SER in different ways. This paper proposes a new SER approach that combines a single first-order attention network with data balance to solve these two issues.
16.	Preprocessing, MFCC, CNN, LSTM, fast fourier transform	The most difficult aspects of emotion recognition are selecting emotion recognition corpora, identifying various speech characteristics, and selecting a suitable classification model. We use a CNN and 13 MFCC features with 13 velocity and 13 acceleration components.
17.	Mel-Frequency Cepstral Coefficients, ERB, HMM	Using two databases, SAVEE and IEMOCAP, we propose and analyze an emotion classification system that focuses on the distinctions between acted and spontaneous emotional expression. We investigated wavelet packet energy and entropy features applied to Mel, Bark, and ERB scales using Hidden Markov Model (HMM) as a classification method in this study.
18.	Speech enhancement metros, spectral Subtraction, wiener filter, MMSE Recognition methodology, emotional feature extraction, HMM	Speech signals are often distorted by various forms of noise in real-world situations. To mitigate this problem, a noise reduction process is performed before using enhancement algorithms to analyze emotional expression. For better emotion classification, three speech enhancement algorithms are introduced: spectral subtraction, wiener filter, and MMSE.
19.	Multi-level multi-head fusion attention, RNN	The Multi-Level Multi-Head Fusion Attention mechanism and a recurrent neural network are used in this paper to present a multimodal method for speech emotion recognition (RNN). Two modalities of feedback are included in the proposed structure: audio and text. Using the OpenSMILE toolbox, we measure the Mel-frequency cepstrum (MFCC) from raw signals for audio features.

©IJRASET: All Rights are Reserved

20.	DNN, CNN, MFCC, SVM	This paper proposes and names BFN, which stands for robust automated speech emotional-speech recognition architectures based on hybrid convolutional neural networks (CNN) and feedforward deep neural networks. CNA, as well as HBN. BEN is a hybrid architecture that combines a bag-of-Audio-word (BoAW) and a feedforward deep neural network, CNA is based on CNN, and HBN is a hybrid architecture that combines a BFN and a CNA.
21.	Discrete wavelet transform, local binary pattern, local ternary pattern, NSL	In this research, a new lightweight effective SER method with low computational complexity was developed. The 1BTPDN approach is used on the RAVDESS, EMO-DB, SAVEE, and EMOVO databases.
22.	CNN, R-CNN, LPC, LPCC	A Residual Convolutional Neural Network (R-CNN) and a gender knowledge block are combined in the proposed algorithm. These two blocks receive the raw speech data at the same time. The R-CNN network extracts the requisite emotional data from the speech data and categorizes the emotion.
23.	Cuckoo search algorithm, weighted binary cuckoo algorithm, PLP, FFT	The algorithm in this paper is designed in two stages. To determine the value of each function, we first use the ensemble learning model random forest algorithm. For experimental comparison, we use emo-db and discover that combining the logistic regression and WBCS algorithms yields the best results.
24.	Fourier coefficients, SBC, MFCC, DFT, DFrFT	This research proposes a new feature extraction method based on adaptive time frequency coefficients to enhance speech emotion recognition accuracy. The GA-CS feature selection algorithm finds the most effective feature array by combining coefficients and the MFCC.
25.	End point detection, MFCC, GMM, SVM, PLP	In this paper, a smart music system is generated by detecting emotion using a voice speech signal as an input. The purpose of the speech emotion recognition (SER) system is to determine a person's emotional state through their expression. Rage, anxiety, boredom, satisfaction, and sadness are the five emotions identified in this report. Speech processing using the Berlin emotional database, then extracting suitable features and choosing appropriate pattern recognition or classifier methods to classify emotional states are all important aspects of implementing this SER framework. The machine platform automatically selects a piece of music as a cheer-up technique from the database of song playlists until the emotion of the speech is recognized.
26.	HMM dynamic programming, LPC	The types of features that can hold more detail about the emotional significance of each utterance are examined in this paper. The method entails determining which features contain the most details and combining these features to improve recognition rates. The phonetic units corresponding to sentences taken from the training base were modelled using the secret Markov model in this paper. Given the size of the training community and the number of people who attended the registration, the findings are very promising. This algorithm is based on the dynamic programming versatility of the hidden Markov model for sentences.
27.	GMM model	In this article, a speech database is used to characterize the emotions conveyed in speech. A database that tends to be semi-natural in appearance. The GEU semi natural emotion speech corpus is used to obtain emotion specific information using LP residual samples as features. Popular Hindi film dialogues were captured to establish the corpus.

©IJRASET: All Rights are Reserved

28.	MFCC, CNN, DBN, RNN	Rather than analyzing data and making decisions from a distance away from the data sources, we can move the decision-making phase closer to where the data resides. The suggested method can be used in several realistic applications, such as knowing people's feelings in everyday life and tension from pilots' or air traffic controllers' voices in air traffic management systems. When the proposed emotion recognition approach was compared to conventional approaches, it was found to be superior.
29.	SVM, parking instruction recognition, DTW	This paper suggests a voice recognition and speech emotion recognition device for emergency parking instructions. The process extracts the feature vector of the speech signal before using a support vector machine (SVM) to recognize the speech's emotion. When the emotion is irregular the dynamic time warping (DWT) algorithm is used to match the parking instruction prototype.
30.	Decision fusion, HMM, ANN	In this paper, we developed three types of classifiers for the four emotions of rage, depression, surprise, and disgust using Hidden Markor Models (HMM) and Artificial Neural Networks (ANN). The DS evidence theory was then proposed to conduct decision fusion among the three types of emotion classifiers for a good emotion recognition outcome.
31.	IRS, MSIN, WB, SWB, FB, LPCC	The aim of this study is to see if audio bandwidth influences human speech emotion recognition. Several standard telephony bandwidths are considered, ranging from full band to narrowband.

Table1: Literature review

III. METHODOLOGIES

A. Different Pre-Processing Methods

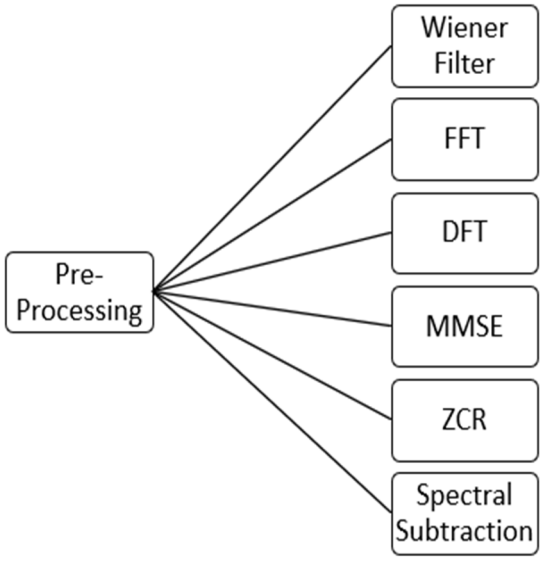


FIG.1: Pre-Processing Methods

As seen in FIG.1, we observe that the pre-processing stage used a lot of methods for speech emotion recognition. The signals from the input source are first pre-processed to make them more compatible, noise-free, and feature extraction- ready. In structures that use randomly spoken words, there is often the possibility that the spoken word will be preceded and followed by silence. The primary stage of pre-processing is the elimination of silence.

©IJRASET: All Rights are Reserved

1) *Wiener Filter:* The Wiener filter has the lowest mean squared error in terms of coefficient vector, making it the best filter to use in speech recognition pre-processing. It is used to low the amount of noise in speech signals [1].
2) *FFT:* The Fast Fourier Transform (FFT) is a DFT implementation that produces virtually equivalent results as the DFT, but it is much more efficient and quicker, reducing computation time significantly. It is simply a computational algorithm for computing the DFT quickly and efficiently. The fast Fourier transform is a term that refers to a set of fast DFT computation techniques.
3) *DFT:* The Discrete Fourier Transform (DFT) measures the spectrum of a finite-duration signal and is one of the most important tools in optical signal processing. The information in the sinusoids that make up a signal is often encoded. However, in some situations, the shape of a time domain waveform is meaningless for signals, and signal frequency content becomes extremely useful in ways other than as digital signals.
4) *MMSE:* In the last two decades, the method focused on minimum mean-square error (MMSE) estimation of the short-time spectral magnitude has been commonly used.
5) *ZCR:* The rate of sign changes of a speech signal frame is the number of times the signal changes value from positive to negative and vice versa divided by the duration of the frame. [11].
6) *Spectral Subtraction:* Boll proposed the spectral subtraction algorithm, which is a commonly used speech enhancement algorithm. It is one of the first effective noise- reduction algorithms in signal processing. The success of spectral subtraction stems from its simple and straightforward algorithm.

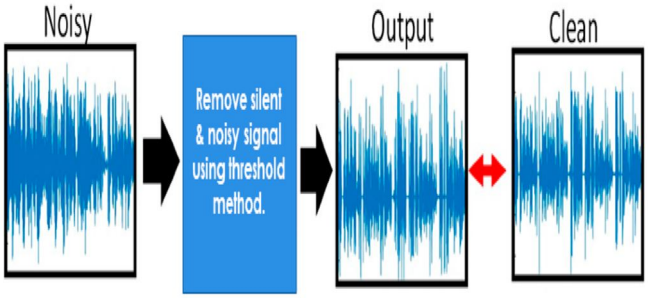


FIG.2: Pre-Processing

B. Feature Extraction Methods

Feature extraction is achieved by converting the speech waveform to a parametric representation for subsequent processing and analysis at a low data rate. As a result, excellent and quality characteristics are used to determine suitable classification. Speech feature extraction techniques used and discussed in the papers include Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), Line Spectral Frequencies (LSF), Discrete Wavelet Transform (DWT), and Perceptual Linear Prediction (PLP).

1) *Mel Frequency Cepstral Coefficients (MFCC):* Mel frequency cepstral coefficients (MFCC) was originally suggested for identifying monosyllabic words in continuously spoken sentences but not for speaker identification. MFCC computation is a replication of the human hearing system intending to artificially implement the ear's working principle with the assumption that the human ear is a reliable speaker recognizer. MFCC features are rooted in the recognized discrepancy of the human ear's critical bandwidths with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to retain the phonetically vital properties of the speech signal. The mel-frequency scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000 Hz. Pitch of 1 kHz tone and 40 dB above the perceptual audible threshold is defined as 1000 mels and used as reference point.

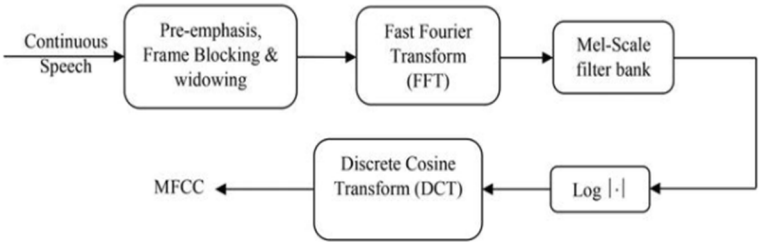


FIG. 3: Block diagram of MFCC processor [8]

©IJRASET: All Rights are Reserved

2) *Linear Prediction Coefficients (LPC):* Linear prediction coefficients (LPC) imitate the human vocal tract and gives robust speech feature. It evaluates the speech signal by approximating the formants, getting rid of its effects from the speech signal and estimate the concentration and frequency of the left behind residue. The result states each sample of the signal as a direct incorporation of previous samples. The coefficients of the difference equation characterize the formants, thus, LPC needs to approximate these coefficients. LPC is a powerful speech analysis method and it has gained fame as a formant estimation method [12].

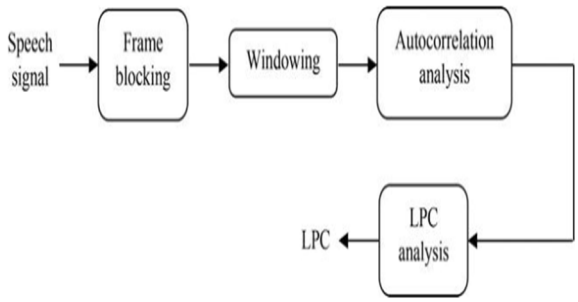


FIG. 4: Block diagram of LPC processor [8]

3) *Linear Prediction Cepstral Coefficients (LPCC):* Cepstral coefficients derived from the spectral envelope determined by LPC are known as linear prediction cepstral coefficients (LPCC). The letters LPCC stand for the coefficients of the Fourier transform illustration of the logarithmic magnitude continuum LPC. Cepstral analysis is commonly used in the field of speech processing because of its ability to perfectly symbolise speech waveforms and characteristics with a limited number of features.

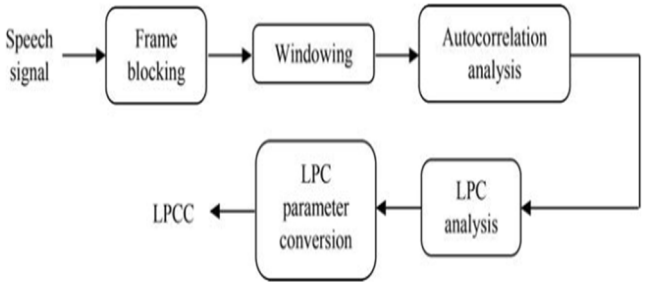


FIG. 5: Block diagram of LPCC processor [8]

4) *Line Spectral Frequencies (LSF):* Individual lines of the Line Spectral Pairs (LSP) are known as line spectral frequencies (LSF). LSF defines the two resonance situations taking place in the inter-connected tube model of the human vocal tract. The model takes into consideration the nasal cavity and the mouth shape, which gives the basis for the fundamental physiological importance of the linear prediction illustration. The two resonance situations define the vocal tract as either being completely open or completely closed at the glottis. The two situations beget two groups of resonant frequencies, with the number of resonances in each group being deduced from the quantity of linked tubes. The resonances of each situation are the odd and even line spectra correspondingly and are interwoven into a singularly rising group of LSF.

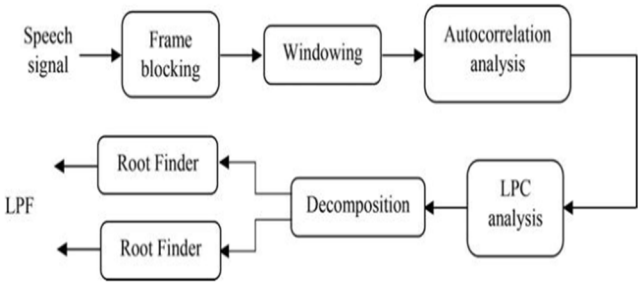


FIG. 6: Block diagram of LSF processor [8]

©IJRASET: All Rights are Reserved

5) *Discrete wavelet transform (DWT):* Wavelet Transform (WT) theory is centred around signal analysis using varying scales in the time and frequency domains. Discrete wavelet transform (DWT) is an extension of the WT that enhances the flexibility to the decomposition process. It was introduced as a highly flexible and efficient method for sub band breakdown of signals. In earlier applications, linear discretization was used for discretizing CWT.

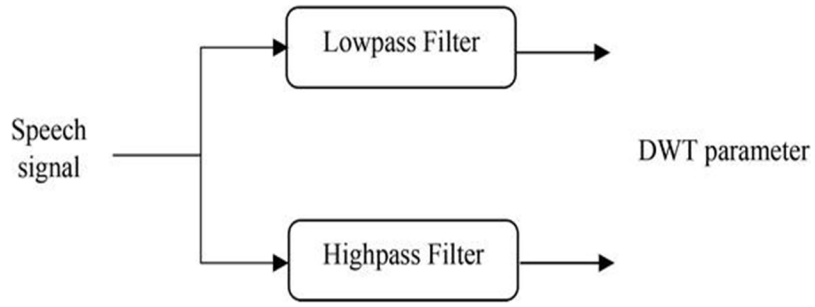


FIG. 7: Block diagram of DWT processor [8]

6) *Perceptual Linear Prediction (PLP):* Perceptual linear prediction (PLP) technique combines the critical bands, intensity-to-loudness compression, and equal loudness pre-emphasis it the extraction of relevant information from speech. It is rooted in the nonlinear bark scale and was initially intended for use in speech recognition tasks by eliminating the speaker dependent features [13]. PLP gives a representation conforming to a smoothed short-term spectrum that has beer equalized and compressed similar to the human hearing making it similar to the MFCC. In the PLP approach, several prominent features of hearing are replicated and the consequent auditory like spectrum of speech is approximated by an autoregressive all-pole model. PLP gives minimized resolution at high frequencies that signifies auditory filter bank-based approach yet gives the orthogonal outputs that are similar to the cepstral analysis. It uses linear predictions for spectral smoothing; hence, the name is perceptual linear prediction. PLP is a combination of both spectral analysis and linear prediction analysis.

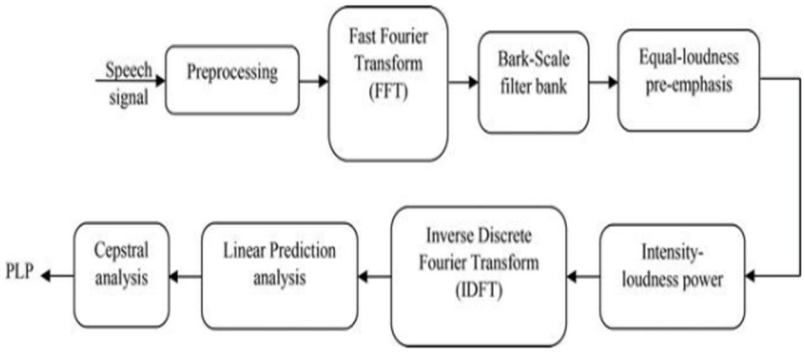


FIG. 8: Block diagram of PLP processor [8]

C. *Classification Methods*

Classification is a form of supervised machine learning in which the algorithm learns from the data it is given and then applies what it has learned to classify new observations. To put it another way, the training dataset is used to achieve better boundary conditions, which can then be used to evaluate each target class; once these boundary conditions have been established, the next step is to predict the target class [14]. In this paper, we have mentioned the various classification methods that were used in the speech emotion recognition.

1) *Dynamic Time Warping (DTW):* Dynamic Time Warping (DTW) technique compares words with reference words. It is an algorithm to measure the similarity between two sequences that can vary in time or speed. In this technique, the time dimensions of the unknown words are changed until they match with that of the reference word [15].
2) *K-Nearest Neighbour (KNN):* K nearest neighbours (KNN) is a supervised machine learning algorithm. A supervised machine learning algorithm's goal is to learn a function such that f(X) = Y where X is the input, and Y is the output. KNN can be used both for classification as well as regression [12] [16].

©IJRASET: All Rights are Reserved

3) *Pattern Recognition Neural Network (PRNN):* Pattern recognition is the method of using machine learning data to identify regularities and similarities in data. These parallels can now be discovered using statistical analysis, historical evidence, or the machine's own prior knowledge. The pattern of the signal formed by the PRNN is recognised by the emotional waves [4].
4) *Support Vector Machine (SVM):* The training data is represented as points in space divided into categories by a simple gap as large as possible in a support vector machine. New examples are then mapped into the same space and classified according to which side of the distance they fall on [17] [18].
5) *Gaussian Mixture Model (GMM):* Speech emotions are described as a mixture of Gaussian densities in GMM. The interpretation that the Gaussian components reflect certain general emotion spectral shapes, as well as the capacity of Gaussian mixtures to model arbitrary densities, inspire the use of this model [12] [19].
6) *Artificial Neural Network (ANN):* ANNs are computers whose architecture is modelled after the human brain. Hundreds of simple processing units are linked in a complicated communication network. Each simple processing unit represents a real neuron that fires when it receives a strong signal from another connected unit or sends out a new signal [20].
7) *Hidden Markov Modelling (HMM):* The most widely used pattern recognition technique for speech recognition is Hidden Markov Modelling (HMM). It is a mathematical model with a collection of output distributions based on the Markov Model. In comparison to the knowledge-based and template-based approaches, this method is more general and has a solid mathematical basis. Speech is broken down into smaller audible entities in this form, and each of these entities represents a state in the Markov Model. There is a transition from one state to another, according to the probabilities of transition [13] [21].

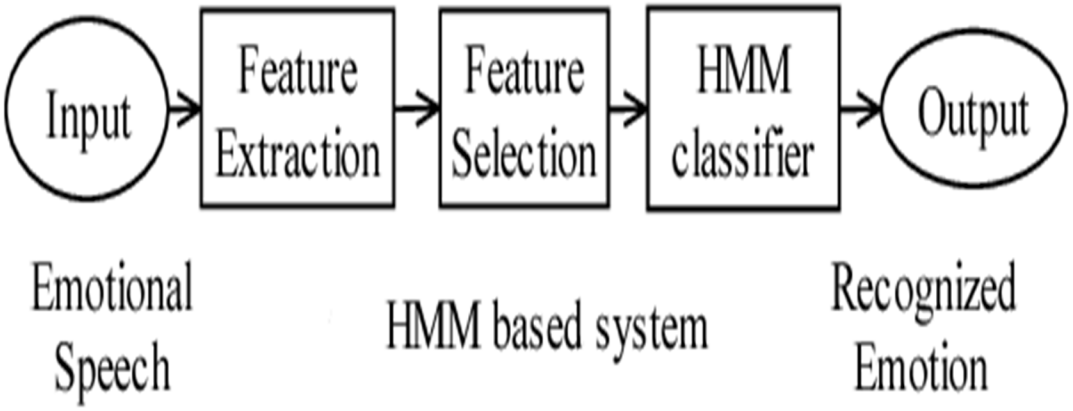


FIG 9: HMM Based System [13]

8) *Vector Quantization (VQ):* Vector Quantization (VQ) is a method for mapping vectors from a large vector space to a finite number of regions within that space. This method is based on the theory of block coding. Each area is referred to as a cluster, and it can be represented by a code- word at its middle. The compilation of all code-words is known as a code book [22] [23].
9) *Convolution Neural Network (CNN):* Convolutional neural networks are a form of deep neural network that is frequently used to analyse visual imagery. Centred on the shared-weight architecture of the convolution kernels that search the hidden layers and translation invariance characteristics, they are also known as shift invariant or space invariant artificial neural networks (SIANN). Multilayer perceptions are regularised variants of CNNs [16] [24]. Multilayer perceptions are usually completely connected networks, meaning that each neuron in one layer is linked to all neurons in the next layer. These networks' "completely connectedness" makes them vulnerable to data overfitting. Regularization approaches widely used include varying weights as the loss function is reduced and arbitrarily trimming connectivity. CNNs take a different approach to regularisation: they take advantage of the hierarchical pattern in data and use smaller and simpler patterns embossed in the filters to assemble patterns of increasing complexity [25].
10) *Recurrent Neural Network (RNN):* For sequential data, recurrent neural networks (RNNs) are a powerful model. RNNs can be prepared for sequence labelling problems where the input-output alignment is uncertain using end-to-end training methods like Connectionist Temporal Classification. The combination of these approaches with the Long Short-term Memory RNN architecture has shown cutting- edge results in cursive handwriting recognition [16] [26].

©IJRASET: All Rights are Reserved

IV. COMPARISON OF VARIOUS APPROACHES AND METHODS:

A. Comparative Study

S.no	MFCC	GLCM	GMM	SVM	KNN	NSL	LPC	LPCC	HMM	RNN	CNN	FFT	PLP
1.	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
2.	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
3.	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
4.	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
5.	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
6.	✓	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
7.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
8.	✓	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✓
9.	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
10.	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
11.	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
12.	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
13.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓
14.	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗
15.	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
16.	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
17.	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
18.	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗
19.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗
20.	✓	✗	✗	✓	✗	✗	✗	✗	✗	✗	✓	✗	✗
21.	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗
22.	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✓	✗	✗
23.	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
24.	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
25.	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓
26.	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗
27.	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
28.	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
29.	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
30.	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
31.	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗

Table2: Comparative Study

©IJRASET: All Rights are Reserved

B. Advantages and Disadvantages

Sno.	Method	Advantages	Disadvantages
1.	MFCC	• It estimates the human system response more strongly than other systems. • It simplifies the computation and present better result of recognition along with reducing utilized time.	• It does not give accurate results for background noises.
2.	LPC	• For speech parameters, it gives very accurate result and is relatively efficient for computation. • LPC is an approach to minimize the sum of the squared differences between the original and the estimated speech signal	• LPC has quantization error, bad performance in a noisy environment. • Unvoiced and nasalized sounds are not accurately represented by LPC.
3.	LPCC	• It is more dependable, quick, and performs well. • It has simple and high performance as compared to MFCC.	• In LPCC, proper ordering is required. • It is highly susceptible to the quantizer noise. • It has low vulnerability to noise.
4.	LSF	• The ability to localize spectral sensitivity. • It can identify animal voice.	• When compared to alternative techniques that operate on the raw input data itself, careful use of processing and analysis methods in the LSF domain could result in a reduction in complexity.
5.	DWT	• It can be used for partitioning the variance of input elements on a large scale.	• For successful speech analysis, it does not have an accurate number of frequency bands. • DWT does not fulfil any of the criteria for direct use of parameterization.
6.	PLP	• PLP is comparatively better than LPC as it suppresses the speaker dependent function. • It has ability to accurately reconstruct autoregressive noise. • It has improved speaker free acknowledgment execution and strong to noise.	• It has least sensitivity to spectral tilt. • It is dependent on overall spectral balance result.

Table 3: Advantages and Disadvantages of Feature Extraction Methods

©IJRASET: All Rights are Reserved

V. APPLICATIONS

METHOD	APPLICATION
MFCC	• It is used to find airline bookings. • It is also used as voice recognition system as security. • It is used in speaker recognition systems. • MFCC can automatically recognize numbers spoken into a telephone.
LPC	• For the development of mobile robots, the LPC method is widely used in musical and electrical companies. • It is used for total analysis of violins and other stringy instruments in telephone firms.
LPCC	• It is used in speaker recognition due to its efficiency.
LSF	• Its feature in speech recognition field is yet to be investigated [27].
DWT	• It's best for analysing signals with information in the low frequency range. • DWT exceptionally intended for dissecting a limited arrangement of perceptions over the arrangement of scales.
PLP	• It provides better performance to cross speaker ASR.
DTW	• DTW is used in mobile applications. • It is used in small embedded systems [28]. • It is used for small vocabulary application.
KNN	• It is used in rooms where there is a lot of multi-channel distortion, including echo and reverberation.
SVM	• It is used to identify different classes of data. • It is used for classification and regression [29].
GMM	• It is used in diagnostic tool for therapists.
Wiener Filter	• It removes additive noise. • It is used for improving the SNR by significant amounts in a car environment.
FFT	• quick multiplication of large-integer and polynomial • efficient multiplication of matrix–vector [30].
DFT	• Used for recognizing musical instruments. • for speech signals analysis implemented in an embedded system.
ZCR	• It is used for analyses. • It is used for speech-music discrimination [31].
Spectral Subtraction	• It is used in speech recognition signal processing to remove background noise. • This algorithm estimates the amplitude of noise.

Table 4: Applications

VI. CHALLENGES/ RESEARCH GAP

The difficulties one might face is within the ethics that we have like how will be the gathered information will be shared? Who will pay for the associated costs? How will be the success and value of the implemented service will be evaluated? Who will make decisions on the gathered data?

VII. CONCLUSION

Combining voice and video will increase the number to 95% in accuracy of understanding the emotion. To understand complicated emotions only LP cannot be used we would have to introduce different features like spectral and prosodic. The cuckoo algorithm is converted into binary weighted cuckoo algorithm and it keeps the most relevant features with a higher probability and uses the cuckoo algorithm's combined random walk strategy which gives us local optimum while converging faster. Even though MFCC gives better results for less noisy data, the hidden Markov model gave satisfactory results with limited number of speakers and small data size. With larger data set CNN-LSTM can produce much better results.

©IJRASET: All Rights are Reserved

Recognize emotions in real life using 4 different types of noise, MMSE and spectral subtraction can improve the recognition rate at the airport noise and babble noise but it is not very effective to car noise. Automatic recognition of speech emotions can be done with LSM, the vocals tracts and source components of speech signals are analysed on ERB model performs great on emotion recognition task however it is a general framework.

Future scope of emotion recognition through speech is everywhere. If this intelligence would be able to comprehend the meaning of emotions then it can create a new connection between social and technical systems that humans are part of. Huge implementation can be done in the military and air forces, with the noisy environments the created intelligence can be used to benefit the officers in communication and control. Commands given in the maps and traveling by the robots can be certainly improved which can create a different market for the voices in maps. And it can give emotional voice commands. It can be also used in the medical field added with video and audio which can depict the condition of the person it is interacting with. It can be used in telecommunications services to better understand the costumer needs and help them with better services by improving and implementing.

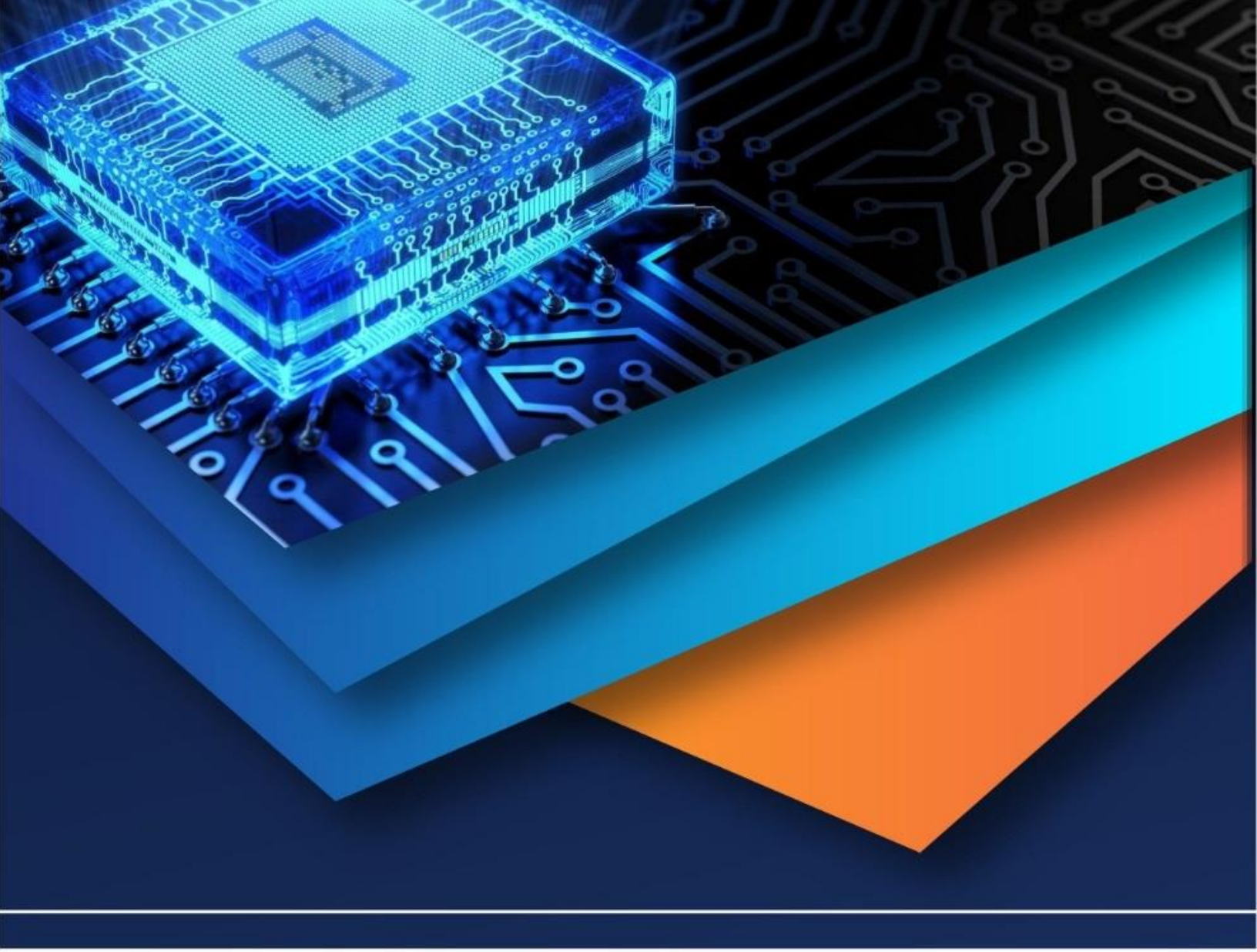
REFERENCES

[1] D. a. M. H. S. Lugović, "Techniques and Applications of Emotion Recognition in Speech," in CIS, Croatia, 2016.
[2] S. R. R. G. K. H. a. A. U. R. M. S. Likitha, "Speech based human emotion recognition using MFCC," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Chennai, India, 2017.
[3] Y. B. A. Hadhami Aouani, "Speech Emotion Recognition with deep learning," Procedia Computer Science, vol. 176, pp. 251-260, 2020.
[4] A. J.Umamaheswari, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," in International Conference on Machine Learning, Big Data, Cloud and Parallel Computing, Faridabad, India, 2019.
[5] R. L. a. P. Gournay, "Biologically inspired speech emotion recognition," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 2017.
[6] R. B. P. a. P. A. K. Tarunika, "Applying Machine Learning Techniques for Speech Emotion Recognition," in 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Bengaluru, India, 2018.
[7] Z. W. W. L. Y. L. J. P. Zengwei Yao, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," Speech Communication, vol. 120, pp. 11-19, 2020.
[8] S. A. A. a. N. K. A. Rashid, Some Commonly Used Speech Feature Extraction Algorithms, 2018.
[9] M. k. a. N. Ellouze, "Pitch and Energy Contribution in Emotion and Speaking styles Recognition Enhancement," in he Proceedings of the Multiconference on "Computational Engineering in Systems Applications, Beijing, China, 2006.
[10] R. D. M. A.-A. Ftoon Abu Shaqra, "Recognizing Emotion from Speech Based on Age and Gender Using Hierarchical Models," Procedia Computer Science, vol. 151, pp. 37-44, 2019.
[11] S. S. B. A. a. A. P. J. Assel Davletcharova, "Detection and Analysis of Emotion From Speech Signals," in Procedia Computer Science, 2015.
[12] S. M. N. P. Rahul B. Lanjewar, "Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) Techniques," Procedia Computer Science, vol. 49, pp. 50-57, 2015.
[13] M. F. a. A. W. J. Fritsch, "Context-dependent hybrid HME/HMM speech recognition using polyphone clustering decision trees," in 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing, Munich, Germany, 1997.
[14] M. N. a. N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019.
[15] S. Z. X. T. a. X. Z. G. Chen, "Speech Emotion Recognition by Combining a Unified First-Order Attention Network With Data Balance," IEEE Access, vol. 8, pp. 215851-215862, 2020.
[16] J. C. a. M. A. S. Basu, "Emotion recognition from speech using convolutional neural network with recurrent neural network architecture," in 2017 2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2017.
[17] M. P. a. S. K. K. R. Chakraborty, "Spontaneous speech emotion recognition using prior knowledge," in 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 2016.
[18] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," in 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, Tunisia, 2016.
[19] H. Y. S. K. a. G. L. N. Ho, "Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network," IEEE Access, vol. 8, pp. 61672-61686, 2020.
[20] A. A. M. K. H. F. A. H. a. A. I. H. M. Ezz-Eldin, "Efficient Feature-Aware Hybrid Model of Deep Learning Architectures for Speech Emotion Recognition," IEEE Access, vol. 9, pp. 19999-20011, 2021.
[21] Y. Ü. S. a. A. Varol, "A Speech Emotion Recognition Model Based on Multi-Level Local Binary and Local Ternary Patterns," IEEE Access, vol. 8, pp. 190784-190796, 2020.
[22] T. -W. Sun, "End-to-End Speech Emotion Recognition With Gender Information," IEEE Access, vol. 8, pp. 152423- 152438, 2020.
[23] Z. Zhang, "Speech feature selection and emotion recognition based on weighted binary cuckoo search," Alexandria Engineering Journal, vol. 60, no. 1, pp. 1499-1507, 2021.
[24] H. M. M. Z. Shadi Langari, "Efficient speech emotion recognition using modified feature extraction," Informatics in Medicine Unlocked, vol. 20, pp. 1-11, 2020.
[25] S. L. a. S. S. Upadhya, "Music player based on emotion recognition of voice signals," in International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, India, 2017.
[26] S. J. V.V. Nanavare, "Recognition of Human Emotions from Speech Processing," Procedia Computer Science, vol. 49, pp. 24-32, 2015.

©IJRASET: All Rights are Reserved

[27] S. D. B. C. A. B. K. S. R. Shashidhar G. Koolagudi, "Recognition of Emotions from Speech using Excitation Source Features," Procedia Engineering, vol. 38, pp. 3409-3417, 2012.
[28] E. G. N. Md. Zia Uddin, "Emotion recognition using speech and neural structured learning to facilitate edge intelligence," Engineering Applications of Artificial Intelligence, vol. 94, pp. 1-11, 2020.
[29] H. Y. Z. G. a. Z. L. T. Kexin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition," in 2019 Chinese Automation Congress (CAC), Hangzhou, China,, 2019.
[30] Y. K. a. L. Li, "Speech emotion recognition of decision fusion based on DS evidence theory," in 2013 IEEE 4th International Conference on Software Engineering and Service Science, Beijing, China, 2013.
[31] R. L. a. P. G. O. Lahaie, "Influence of audio bandwidth on speech emotion recognition by human subjects," in 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, Canada, 2017.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)