Volume: 11 Issue: III Month of publication: March 2023

DOI: https://doi.org/10.22214/ijraset.2023.49674

A Review on Recognizing of Positive or Negative Emotion Based on Speech

Dr. P. Kiran Kumar[1], K. Sudha Rani[2], K. Praveen Sai[3], N. Rakesh Krishna[4], S. Chaitanya Naga Aditya[5]
Computer Science and Engineering, Sasi Institute of Technology andEngineering

Abstract: Emotions are the feelings of a person and the reaction to a situation. People can verbally or nonverbally convey their feelings. In the literature ofSpeech emotion recognition, many of the techniques have used traditional ways to detect emotions, this paper present different way of recognizing emotions, by speech signals are processed by CNN to extract features, the extracted features are then used to input SVM. The SVM outputs the predicted emotions. From this approach we improved accuracy by testing and training the models based on the input audio data.

Keywords: Speech emotion detection, Speech signals, SVM, Noise Reduction, emotion classification.

I. INTRODUCTION

The method of detecting human emotions is called emotion recognition. The ability to understand other people's emotions differs widely amongst individuals. Technology's usage to help people recognize their emotions is a relatively recent area of research . A crucial stage of study for bettering the interface between people andcomputers is emotion recognition. The acquisition task is made more difficult by the complexity of emotion. There are different types of emotion recognition fields. They are Video, Image, Speech, Text, and Conversation. Emotion recognition in conversation is concerned with collecting emotions from conversations between two or more people. Datasets in this category are often derivedfrom free samples obtained from social media networks. However, many obstacles remain in this subject, such as the existence of sarcasm in a discourse, the emotional change of the interlocutor, and conversational-context modelling[1].

A. Emotion Recognition Through Speech

Humans' four fundamental emotions are happiness, sadness, anger, and fear, according to most experts . Among other things, our faces, words, gestures, blood pressure, temperature, and heart rate may all be used to detect our emotional states. In particular, speech is essential for conveying human emotions. The connection between speech and emotion has been extensively studied. Williams andStevens discovered in 1972 that emotions had numerouseffects on a speech's fundamental frequency contour.

Murray summarized a number of previous studies onemotions in 1993.

Pitch, length, intensity, and tone of voice are now the most regularly quoted vocal elements in emotion studies[2].

Determine a speaker's emotional state for a number of purposes. In human-machine interactions, the computer maybe programmed to produce more suitable responses if the individual's emotional state can be correctly identified. To enhance the precision of spoken word detection, most modern automatic speech recognition systems include actuallanguage comprehension.

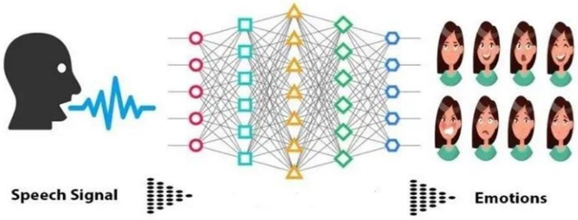

Fig.1 Speech Emotion Detection (MLA)

Such language comprehension can be increased far more if the speaker's emotional state can be retrieved, whichwill strengthen the device's reliability. In general, transcription is essential when communicating across multiple languages. The semantic element of speech is the focus of current machine translation techniques. If the emotional state of the speaker could be discovered and conveyed, it would provide additional relevant information to the communicating entities, especially in non-face-to-face Adjacent sources. Tutoring, alerting, and other uses of automatic emotion recognition systems and entertainment[3].

Nonverbal cues such as hand movements, facial expressions, and speech tone are used to express feelings and feedback in human communication, on behalf of spoken language. However, new human-computer interface trends, which have developed from the traditional keyboard and mouse to automatic speech recognition systems and special interfaces designed For physically disabled people, do not fully exploit these valued language abilities, resulting inless-than-natural interactions. If computers could understand these emotional inputs, they could provide users with more personalized and suitable assistance that is more in line with their requirements and preferences[4]. In the area of Human Computer Interaction, automatic speech emotion recognition is a hot issue with several applications. It can be utilized in an in-car board system to convey information about the driver's mental state in order to initiate his or her safety. It is utilized in an autonomous remote call centre to detect consumer unhappiness in real- time. In the realm of e-learning, recognizing students' emotions early and providing appropriate therapy can improve the performance of instruction[5]. In today's E-learning environment, professors and students are frequently separated in space and time, which might result in a lack of emotional interactions. Furthermore, the teacher is unable to adapt his or her teaching technique and content in response to the pupils' emotions. When students are actively engaged in a topic, they will be dynamic and active in an online group discussion, and they will demonstrate their positive feeling. If they get into difficulty or are uninterested in it, on the other hand, they will exhibit the opposite reaction. We can help the teacher change the teaching plan and enhance learning efficiency by detecting emotional data and providing helpful feedback. Intelligent Human-Computer Connection (HCI), which makes use of body gestures, eye contact, voice, facial expressions, cognitive modelling, and other methods, aims to enhance human-machine interaction. Perceptual recognition, machine learning, cognitive modelling, and emotional computing have all seen an increase in interest recently. Facial expressions can be utilized to detectemotions quickly and easily, making HCI easier[6].

Facial expression detection and classification can be utilized as a natural approach for man and computer to interact. But each person's facial expressions differ in intensity based on their age, gender, face shape, size, and size, among other things. Even a person's expressions change throughout time, even if they are the same person. As a result, developing a generalized facial emotion analysis system is a difficult endeavor. Specialized techniques, on the other hand, can effectively interpret an individual user's facial features[7]. When two people engage with one other, they may quickly detect the underlying emotion in the other person's speech. The emotion recognition system's goal is to imitate human perception systems. Speech emotion recognition is applicable in various fields. In making decisions, emotion can hold a crucial impact. Diverse biological signs can also be used to detect emotion. A system can operate appropriately if emotion can be identified correctly from speech. Medical science robotics engineering, contact centreapplications, and other fields can benefit from an effective emotion recognition system. A human can quickly detect the speaker's emotion. Since interpersonal relationships are crucial for making sane and knowledgeable decisions. It aids in understanding and comprehending the ideas of many individuals by identifying feelings and offering comments toward others. Research shows that emotions significantly shape how people connect with one another. A lot may be learned about someone's mental health through their emotional behavior. A relatively recent area of study, automated emotion identification, is concerned with comprehendingand retaining desirable feelings. [4].

There are numerous modes of communication, but the speech signal is one of the quickest and also most normal ways to communicate between humans. As a result, speech can be a quick and efficient method of collaboration between humans and machines. Living beings have the natural skill to use all of their accessible senses to gain the most appreciation of the text msg. People can detect their communication partner's emotional state using all of their senses. Emotion detection is instinctive for humans, but it is a challenging task for machines.

II. GENERAL ARCHITECTURE OF SPEECHEMOTION DETECTION

The several phases that make up the general architecture of the speech emotion detection are illustrated in Fig.2.In the first speech signals are taken as input and thespeech is extracted from the speech signals and byclassifying the words of that speech the emotion of theperson is detected in the final step. The emotions are like Angry, Sad, Happy, Surprise, Anxiety, Joy , etc..,.

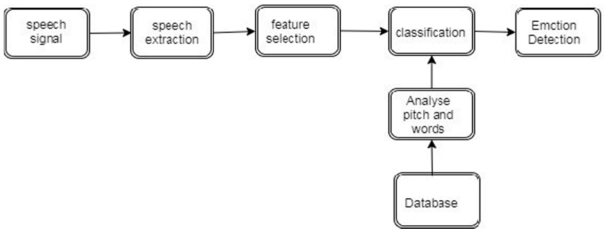

Fig.2: General architecture diagram for emotion detection

III. LITERATURE SURVEY

Sudarsana reddy kadiri et al. (2020) [1] have proposed amodel that used the Kullback-Leibler (KL) distance to compare the feature distributions of emotional and neutral speech. By comparing the KL distance between a test speech and an utterance spoken by the same speaker under neutral conditions, the approach assesses the emotion. Both the IIIT-H Telugu emotional speech database and the Berlin emotional speech database were used in this investigation (EMO-DB).Without submitting the databases to emotional speech training, they used these two databases and were ableto recognize emotion. They employed KL distance and took input as a speech signal to use this database without training with emotional speech to automatically recognize the emotion from the speech. A technique for automatically identifying emotions in speech that makes use of excitation properties taken from GCIs. The system calculates the difference between a test utterance of emotional speech anda reference utterance generated in a neutral condition using the KL distance. According to the suggested method's results, emotional speech may be distinguished from neutral speech primarily by its excitation properties. The accuracy rates for Neutral vs. Emotional are 83% and 77%, respectively.

Didik Muttaqin et al. (2020) [2] designed a system for categories those seven emotions, A Mel-Frequency Cepstral Coefficient (MFCC), which is frequently used to create specific coefficients, is a determinant parameter for HMM. They had taken the speech as input signal up to 240 utterances and used MFCC and HMM and developed a model and check for slower and faster execution of thespeech signal. By this model they detected the Emotion of the speech in a better way. Emotion is detected with both faster execution time and also smaller execution time. But slower execution time gives better accuracy for detection of emotion. Faster execution time accuracy for 3 scenarios is: 73.57%, 77.67%, 82.01% and slower execution time accuracy for 3 scenarios are: 73.80%, 80.50%, and 81.65%.

Zeenat Tariq et al (2019) [3] Offer a technique for real-time emotion recognition utilizing voice signals andInternet of Things (IoT)-based deep learning for the care of elderly residents in nursing homes. They created the Speech Emotion Detection (SED) integrated deep learning model using 2D convolutional neural networks (CNN). Using deeplearning techniques, they provided a novel method for speech emotion identification in this study. This essay is divided into three parts. A cutting-edge fusion strategy was first used to enhance and normalize data pre-processing. Using cutting-edge methods such as Root Mean Square (RMS), Peak value identification, and EuropeanBroadcasting Unit, we have developed a realistic method fornormalizing and eliminating noise from audio data (EBU). Second, to improve classification, they created a deep learning model dubbed Speech Emotion Detection employing data normalization and augmentation pre- processing techniques (SED). A live speech system that can instantly anticipate the speaker's emotional state was developed using the SED model. Psychologists working in the medical area may employ this technique to assess patients' emotional states while they go about their daily lives, such as elderly patients receiving care in nursing homes. The technique may be used by systems that offer recommendations based on speech-based analysis and emotional prediction.

Omar ahmad mohammad et al. (2021) [4] created a system that uses recordings of real people's voices portraying different emotion classes to recognise and detect emotions in Arabic speech. The three primary stages of the suggested technique are the signal pre-processing stage for noise removal and signal band width reduction, the feature extraction stage utilizing linear predictive coding, and the final stage (LPC). The recommended method starts with the speech signal, smooths it, divides it in half with DWT, and then gets rid of the quiet energy nodes. After STE signal segmentation, the 14 LPC coefficients are extracted from the segmented original signal. To train the 25-coefficient logistic regression ML model, the LPC and PPSD coefficients are merged. The feeling was eventually recognized.

Utkarsh Garg et al (2020) . [5] proposed a system which discern emotion from audio speech signals, a speech emotion recognition (SER) system has been created. The classifier used to identify emotions and the type of features retrieved both affect how effectively an emotion identification system works. The MFCC, MEL, and Chroma techniques are those suggested for this investigation of the paper. They classified different emotions using classifiers and machine learning techniques. There are two steps to our work. First, from acoustic signals, we have retrieved characteristics. Then, using a dataset to train a variety of classifiers on, we compared their results to show the impact of these characteristics on SER. We also looked at the outcomes when the model and MFCC were trained using additional characteristics like Chroma and MEL.

Dario Bertero et al. (2017) [6] For the objective of recognizing speech emotions, a real-time convolutional neural network model has been proposed. In this study's article, the proposed method (CNN) was a convolutional neural network. Our network was trained using standardback propagation with an initial learning rate of 104 and momentum set at 0.9. Every 15 epochs, the learning rate was cut in half, and the training was terminated when an increase in error on the development set was seen. We selected the rectified linear function as a non-linear function since it performed better than tanh.

.

The CNN was implemented using the THEANO toolkit. Due to the small size of the corpus, we randomly selected 10% of the test set and 80% of the training set for each round of a three-fold cross validation. The INTERSPEECH 2009 emotion challenge feature set served as the baseline for our three-fold cross-validation, three-fold cross-training of a linear- kernel SVM. In contrast to the SVM average result of 63.0%, CNN offers an average accuracy of 66.1%. The CNN performs better overall and for the "Angry" and "Happy" classes, whereas the SVM performs better for the "Sad" class.

Sabrine Dhaouadi et al. (2019) [7] developed a method where the key difficulty is determining how directly the model performance is impacted by the audio file quality, actor emotion expression, and amount of files supplied into the training process. Support Vector Machines (SVMs) and Artificial Neuron Networks are the recommended techniques (ANN). The suggested method was implementedby integrating two separate speech corpora: Emo-DB, whichis in German, and SAVEE, which is in British English. These classification investigations mainly focused on four classes: wrath, pleasure, melancholy, and neutral mood. Each audio segment produced a 39*N extracted feature vector, 39 cepstral coefficients, and statistical components based on the temporal trend of the energy method, where N is the number of frames in each database. SVM and ANNare two supervised machine learning models that have been proposed and are often used for identifying emotions from audio signals. They were investigated and contrasted. The results of the trials suggest that, under our conditions and forthe parameters we fixed, the ANN model outperformed the SVM. The MLP classifier won the research because to its superior accuracy and favourable f1-score in the three submitted datasets. The calibre of the audio files, the performers' capacity to convey their emotions clearly, and the volume of audio files utilised during training all have a direct bearing on the model performance.

Pavol Harár et al. (2017) [8] Convolutional, pooling, and fully connected layers make up the Deep Neural Network (DNN) architecture used in this study's technique for voice emotion identification (SER). Three class subsets from the German Corpus were used (angry, neutral, and depressed). The proposed model was created by optimising the binary cross-entropy loss function log-loss and trainingit with the stochastic gradient descent technique at a fixed learning rate of 0.11. Over numerous epochs, the DNN received input data in batches of size 21. Each sample featured a segment from a distinct successful class, and eachbatch contained exactly the same number of segments from each class (e.g. angry, neutral, sad, angry, neutral, sad, angry ...).The final batch may be smaller if necessary sinceit includes the last few parts needed to finish an era, but all prior equality rules were upheld. The sequence in which the data were sent to DNN changed at each epoch. Since batch sizes might vary, vectors should start with "None" rather than a scalar. We set the waiting time to 15 minutes in order to prevent over-fitting. In other words, if the validation loss had not improved after more than 15 training epochs, the experiment was considered to have failed. The statistics demonstrated that the 38th training cycle produced the best outcomes. By leveraging the Keras and Theano libraries to create the DNN model, we were able to speed the GPU training (NVIDIA GeForce GTX 690). The 38-epoch training was completed in 4.12 minutes. The outcomes of the validation were taken into consideration while changing each hyper parameter. Using the trained model, we can forecast the probabilities for each class of each segmentfrom the validation and testing sets. To represent the anticipated class, we selected the highest number from the estimated likelihood. The goal of this study was to identify an individual's emotional state from a brief vocal recording that was divided into 20-millisecond segments. On test data, our technique obtained a 96.97% accuracy rate and an average file prediction confidence of 69.55%.

A. Yousaf et al. (2021) [9] had done an analysis of seven machine learning algorithms for classifying the tweets as happy or unhappy. Using a voting classifier thatcombines Stochastic Gradient Descent and Logistic Regression, the tweets are automatically categorised (LR- SGD) with TF-IDF, providing an accuracy of 79% without requiring personal reading of the tweets. There are 99,989 records in the collection. Women's e-commerce apparel evaluations and sentiment analysis of hate speech detection on Twitter are the additional datasets utilised for stability. Voting Classifier is the technique used to determine thetext's emotional tone (LR-SGD). The voting classifier is a mix of stochastic gradient descent and logistic regression. The text is classified using a logistic regression classifier into happy or sad, top or bottom, positive or negative, and black or white. Stochastic Gradient Descent is used as a high-level iterative strategy for perfection in target work. Using the phrases term frequency and inverse document frequency, one may determine how often a word appears ina document. TF determines how many times a certain phraseappears in the text. IDF is a word that appears often throughout several texts. It can be applicable in Twitter textual tweets classification, women’s e-commerce clothing reviews, and to identify the reviews of the products.

Wang et al. (2021) [10] a deep learning model was put up to extract emotional connections from material written in natural language. Previous artificial intelligence research concentrated on detecting sentiment rather than investigating why sentiments are not or are identified erroneously. Using natural language content from internet news articles, the gap between emotion detection and emotion correlation mining is filled. The news articles use a variety of lengthy and short, as well as long and short,subjective and objective language. It came from a popular social media platform.

A Chinese article on Sina News served as the source of the case study's data. The CNN- LSTM2 and CNN-LSTM2-STACK deep neural network models are suggested as two ways to infer emotion correlation from emotion detection in text. Three criticalphases comprise the computing process. It involves analysing features in part 1, calculating emotions in part 2, and paying attention to original features in part 3. Applicable for affective interaction like social media comments, interactions between humans and computers, andidentifying sentiment in the public networks.

Sundaram et al. (2021) [11] used an approach to identify emotion in the text based on TF-IDF. The accuracy is increased from 80% to 85% by building a system that incorporates additional algorithms and classifiers including LSA, Support Vector Classifier, and Random Forest Classifier. The dataset consists of 16000 phrases separated into six different categories. Information retrieval and text analysis may both be done using latent semantic analysis. Texts can be organised and unstructured. To discriminate between various regression problems, SVM is a powerful machine learning model. A number of trees make up RF, and each tree has a weight. The words are analysed using a method known as Phrase Frequency-Inverse Document Frequency in order to balance each term and establish its principal usage. The system is mainly used for Suicidalprevention and depression detection by automatically identifying the text and for the product recommendationsystem by the reviews.

Rashid et al. (2020) [12] developed a system using Aimens system. The system is used for detecting emotions from textual tweets conversation by taking input as word2vec and doc2vec embeddings. Aimens system scores about 71% to improve human-computer interaction. The model used is Long Short-Term Memory (LSTM), Convolutional Network, and Pooling to improve accuracy. Mainly GridCV is used for tuning hyper parameters. NLP library, textbolb is used for spell correction. The dataset used mainly is a 3-turn conversation. It is retrieved from the Twitter tweets dataset. SemEval 2019 provides a dataset thatis used with 15k tweet conversation pairs. An LSTM is a long short-term memory. In order to detect hyper parameters during model training, LSTM leverages deep learning, which speeds up and enhances learning. It uses an RNN strategy (Recurrent Neural Network). It may be used toidentify handwriting as well. A convolutional network is a deep learning technique. It is mostly used to recognise faces in images and determine the expressions on those faces.Pooling is used for accuracy improvement. Pooling is used for computational time reduction. The system is applicableto Twitter conversations and any social media conversations.

Nirag T et al. (2020) [13] have used support vector machine, naive bayes, and random forest machine learning classifiers to do sentiment analysis to identify whether phrases are positive or negative. Using different features such as Unigram, TF-IDF, and Bags of word compared withclassifiers to get high accuracy. The sentiment analysis dataset is used for analyzing texts as positive or negative. All the social media sites have been utilized for the extraction of comments as datasets. Support Vector Machine is a classification algorithm with a linear model that classifies data into classes. It depicts new data elements for the calculation of margin. Naive Bayes is a Theorem-based machine learning classification technique that is used for calculating probabilities that are often known as the Zero frequency technique. Text classification is one of the basic prediction techniques of Naive Bayes. The applications of the proposed system are low cost and the fastest way to collect data from the customer.

Juyan Islam et al. (2020) [14] developed a system for emotion recognition. The research is conducted from English-language microblogs and the same principle thatmay be applicable to microblogs in other languages. Their research takes into account four basic emotional subcategories (i.e., sad, happy, angry, and love.) Future research can integrate more accurate emotional states (such as disgust and surprise) into the recognition system.Additionally, employing the LSTM approach, a system witha large dataset may provide output that is more realistic.

Tatsuki Akahori et al. (2021) [15] with a 98% accuracy developed a useful method for Japanese twitter emotion datasets utilising sentence-final phrases. This study offers a useful method for gathering literary works in Japanese that communicate the original experience through emotional expressions and sentence-final phrases. The suggested technique uses the deep learning-based language model BERT to find tweets that exhibit the first emotion. There are2,234 tweets in the dataset that express a variety of emotions.

The developed algorithm will categorise data from a Japanese dataset that will detect the levels of happiness and sadness. A deep learning-based language model outperforms standard methods when applied to the Bag of Words model.

Siyuan Zheng et al. (2021) [17] To address this problem, they present in this work an acoustic segment model (ASM)-based technique for speech emotion recognition (SER).This research suggests a brand-new SER paradigm based on ASM. Topic models like LSA in the field of information retrieval can process the relationship between a document and a word. The method presented in this study is useful for gathering Japanese literature that expresses the original sentiment through emotive expressions and sentence-final clauses.

The suggested approach uses the BERT language model, which is based on deep learning, to identify tweets that exhibit the first emotion. 2,234 tweets from the sample show a variety of moods.

The built-in technology will classify information from a Japanese dataset that will detect the levels of happiness and sorrow. An advanced language model based on deep learning surpasses standard methods using the Bag of Words paradigm.

Nancy Semwal et al. (2017) [18] developed a method to infer an emotional state from speech cues. The proposed technique displays a variety of acoustic characteristics, such as energy, zero crossing rate (ZCR), fundamental frequency, and Mel frequency cepstral coefficients (MFCCs).Theexperiments below were carried out with the chosen method:

1) SVM classification of one emotion vs one feeling.
2) SVM and LDA-based multiclass classification.

The algorithms LDA and SVM classifiers must first be taught to make an accurate classification, and then they mustbe evaluated to determine their correctness. As a result, we need to divide the training and testing datasets so that thereis no overlap. Using the K-fold cross validation approach, which was described in Section, one may divide data effectively. Through repeated 10-fold cross validation, the correctness of the experiments in our implementation has been evaluated. This results in a more accurate assessment of the classifier's performance as opposed to only running k- fold cross validation once. They demonstrated a technology that automatically detects speech emotions.

By integrating the mean, median, maximum, and standard deviation of LLDs from the temporal, spectral, and cepstral domains, global characteristics are created instantly.

Kun-Yi Huang et al. (2019) [16] created a technique for identifying emotions in face-to-face interactions using both spoken and nonverbal sounds present in an utterance. Convolutional neural networks and a long short-termmemory (LSTM)-based sequence-to-sequence model are among the recommended techniques (CNNs). The construction of the emotion model, the extraction of embedding features, and the segmentation of verbal and nonverbal sound were the three crucial elements in the training phase of the suggested system architecture.

The initial methods for analysing speech data were prosodic- phrase segmentation, verbal/nonverbal segment recognition, and silence detection. Second, employing sound/speech segments of verbal and nonverbal sounds, the essential CNN-based models for extracting the general features of emotion and sound were trained.

CNN models that were trained for emotion/sound type classification were employed as feature extractors by removing the output layer. The CNN sound and emotion qualities are finally combined to produce the representative feature vector for each segment. Using the feature vector sequence and accounting for time-dependent emotional changes, the results of each segment's emotion identification were achieved. The results showed how well emotion identification performed utilising CNN-based features and an LSTM-based emotion model.

Safa Chebbi et al. (2018) [19] proposed a method in which the study is to assess how pitch-related characteristicsaffect the ability to recognise the emotion of dread in voice signals.K-nearest neighbours, the SVM algorithm, the decision tree method, and the subspace discriminant algorithm. Each emotion statement yields a mean value, and pitch values are computed for each emotion frame by frame.

The average values for each emotional category may then be calculated, as well as the imitation sequences produced by the stationary speaker. division of the pitch distribution into emotional categories. The dots on the graph represent various mean pitch levels for a sampling of words said by this speaker.

The mean of that pitch makes it easy to discriminate between various mood groups. The other emotions' mean pitch values encompass a wider range, from 100 to 300 Hz, , but the fear category is easily identified as falling between 200 and 320 Hz. In contrast to other classes the neutral classes mean values are dispersed throughout a more condensed range between 140Hz and 200Hz.

On the the neutral class' mean values are dispersed throughout a more condensed range between 140Hz and 200Hz. On the second level, dread's mean pitch value rises to 320Hz, exceeding the other negative emotions' mean pitch value of 280Hz. They conclude by confirming that pitch can serve as a distinguishing feature between various states and the intended emotion of dread. A research that examined how pitch-based factors affected the ability to recognise fear. It implies that pitch-related acoustic variables have a comparatively high level of discriminating between emotional states. Thus, employing the k-nearest neighbour approach, the greatest accuracy rate in this investigation was78.7%.

Table1: Comparision Table

Ref no.	objective	Input type	Technique used	Accuracy	Limitations
Sudarsana reddy kadiri[1]	Emotion detection	Speech Input	Kullback-Leibler (KL) distance	For Neutral vs. Emotional are 83% and 77%	Classified only neutral or Emotional but not specific emotion.
Didik Muttaqin[2]	Emotion detection	Speech Input And Text input.	Mel-Frequency Cepstral Coefficient (MFCC)	Faster execution time accuracy for 3 scenarios is: 73.57%, 77.67%, 82.01% and slower execution time accuracy for 3 scenarios are: 73.80%, 80.50%, and 81.65%.	Detection is observed more on slow time execution but slow time execution takes more amount of time to detect the emotion.
Zeenat Tariq[3]	Emotion detection	Speech Input	deep learning model (CNN).	-	Proposed a speech emotion recognizer but few emotions are detected.
Dario Bertero[6]	Emotion detection	Speech Input	CNN	66.1%.	False detections are made in detecting emotions and also detected one emotion i.e, sad emotion.
Utkarsh Garg[5]	Emotion detection	Speech Input and text input	MFCC, MEL, and Chroma	-	Detected the emotion but to less number of speeches.

IV. ANALYSIS AND DISCUSSION

The analysis of emotion detection based on the literaturesurvey is included in this part. we have taken into account that includes accuracy based and objective based and datasetbased analysis.

V. CONCLUSION

This study includes about the speech emotion detection based on different techniques used and other emotion detection methods. We have included the parameters in emotion detection are Speech input, Database, Accuracy. Most of the speech emotion detection methods did not included the noise reduction in the speech and we included CNN with SVM for improving accuracy and performance of the model. Computational complexity ,ability of work and speed of the model are the challenges which are addressed in further.

REFERENCES

[1] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8,pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.

[2] Z. Tariq, S. K. Shah and Y. Lee, "Speech Emotion Detection using IoT based Deep Learning for Health Care," 2019 IEEE International Conference on Big Data (Big Data), 2019, pp. 4191-4196, doi: 10.1109/BigData47090.2019.9005638.

[3] U. Garg, S. Agarwal, S. Gupta, R. Dutt and D. Singh, "Prediction of Emotions from the Audio Speech Signals usingMFCC, MEL and Chroma," 2020 12th International Conference on Computational Intelligence and Communication Networks (CICN), 2020, pp. 87-91, doi: 10.1109/CICN49253.2020.9242635.

[4] K. -Y. Huang, C. -H. Wu, Q. -B. Hong, M. -H. Su and Y. -H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp.5866-5870, doi: 10.1109/ICASSP.2019.8682283.

[5] D. Bertero and P. Fung, "A first look into a Convolutional Neural Network for speech emotion detection," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 5115-5119, doi: 10.1109/ICASSP.2017.7953131.

[6] P. Harár, R. Burget and M. K. Dutta, "Speech emotion recognition with deep learning," 2017 4th International Conference on Signal Processing and Integrated Networks(SPIN), 2017, pp. 137-140, doi: 10.1109/SPIN.2017.8049931.

[7] X. Wang, L. Kou, V. Sugumaran, X. Luo and H. Zhang, "Emotion Correlation Mining Through Deep Learning Modelson Natural Language Text," in IEEE Transactions on Cybernetics, vol. 51, no. 9, pp. 4400-4413, Sept. 2021, doi: 10.1109/TCYB.2020.2987064.

[8] D. Muttaqin and S. Suyanto, "Speech Emotion Detection Using Mel-Frequency Cepstral Coefficient and Hidden Markov Model," 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2020, pp. 463-466, doi: 10.1109/ISRITI51436.2020.9315433.

[9] O. A. Mohammad and M. Elhadef, "Arabic Speech Emotion Recognition Method Based On LPC And PPSD," 2021 2nd International Conference on Computation, Automation andKnowledge Management (ICCAKM), 2021, pp. 31-36, doi: 10.1109/ICCAKM50778.2021.9357769.

[10] S. Dhaouadi, H. Abdelkrim and S. B. Saoud, "Speech Emotion Recognition: Models Implementation & Evaluation," 2019 International Conference on Advanced Systems and Emergent Technologies (IC_ASET), 2019, pp. 256-261, doi: 10.1109/ASET.2019.8871014.

[11] A.Yousaf et al., "Emotion Recognition by Textual TweetsClassification Using Voting Classifier (LR-SGD)," in IEEE Access, vol. 9, pp. 6286-6295, 2021, doi: 10.1109/ACCESS.2020.3047831.

[12] V. Sundaram, S. Ahmed, S. A. Muqtadeer and R. Ravinder Reddy, "Emotion Analysis in Text using TF-IDF," 2021 11th International Conference on Cloud Computing, Data Science

& Engineering (Confluence), 2021, pp. 292-297, doi: 10.1109/Confluence51648.2021.9377159.

[13] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi and S. K. Shahzad, "Emotion Detection of Contextual Text using Deep learning," 2020 4th International Symposiumon Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, pp. 1-5, doi:10.1109/ISMSIT50672.2020.9255279.

[14] Nirag T. Bhatt, Asst. Prof. Saket J. Swarndeep," Sentiment Analysis using Machine Learning Technique: A Literature Survey", in International Research Journal of Engineering andTechnology (IRJET), Volume: 07 Issue: 12 | Dec 2020.

[15] Juyana Islam; Sadman Ahmed; M. A. H. Akhand; N. Siddique; (2020). Improved Emotion Recognition from Microblog Focusing on Both Emoticon and Text. 2020 IEEE Region 10 Symposium (TENSYMP), (), –. doi:10.1109/TENSYMP50017.2020.9230725.

[16] Tatsuki Akahori; Kohji Dohsaka; Masaki Ishii; Hidekatsu Ito; (2021). Efficient Creation of Japanese Tweet Emotion DatasetUsing Sentence-Final Expressions. 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech), (), –. doi:10.1109/lifetech52111.2021.9391800.

[17] S. Zheng, J. Du, H. Zhou, X. Bai, C. -H. Lee and S. Li, "Speech Emotion Recognition Based on Acoustic Segment Model," 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021, pp. 1-5, doi: 10.1109/ISCSLP49672.2021.9362119.

[18] N. Semwal, A. Kumar and S. Narayanan, "Automatic speech emotion detection system using multi-domain acoustic featureselection and classification models," 2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA), 2017, pp. 1-6, doi: 10.1109/ISBA.2017.7947681.

[19] S. Chebbi and S. Ben Jebara, "On the use of pitch-based features for fear emotion detection from speech," 2018 4th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2018, pp. 1-6, doi: 10.1109/ATSIP.2018.8364512.

[20] https://medium.com/@raihanh93/speech- emotionrecognition- using-deep-neural-network-part-i-68edb5921229