IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Analysis of Classification Techniques for Categorizing Emotions in Human Speech

[1]Mr. Sachin Jadhav, [2]Mrudula Kulkarni, [3]Kaustubh Narkhede, [4]Kishor Shivsharan

[1]Assistant Professor, [2]U.G Student, [3]U.G Student, [4]U.G Student
[1]Department of Information Technology
[1]Pimpri Chinchwad College of Engineering, Pune, India

Abstract : Speech Emotion Recognition (SER) is a term used to describe a system that identifies the human emotion and related feeling states from the speech. This is aided by the evidence that the speech frequently expresses basic emotion through pitch and tone. In recent years, emotion identification has become a rapidly growing research subject. Machines, in contrast to individuals, lack the ability to see and express emotions. However, by using automated emotion recognition, human-computer interaction can be enhanced, eliminating the need for human assistance. Emotional speech signals are used to examine basic emotions such as disgust, calm, fear, happy, and so on. We focused on studying Human Speech Emotion using a feature extraction technique called Mel Frequency Cepstrum Coefficients (MFCC), as well as anticipating emotions using three different classifiers: Convolutional Neural Network (CNN), Support Vector Machine (SVM), and Multilayer Perceptron (MLP). It has been discovered that every one of these strategies gives the best accuracy with a specific arrangement of features separated from the input voice. This paper analyzes the accuracy of SVM, CNN, and MLP with their own set of features.

IndexTerms **- Speech emotion recognition, mel cepstral coefficient, convolutional neural network, multi-layer perceptron, support vector machine.**

I. INTRODUCTION

Speech is the most basic and direct way of sharing information. It can communicate complex feelings and emotions through the emotions it carries and depicts them in reaction to objects, scenarios, or events, and it can store a large amount of data. Because automatic emotion detection systems may be utilised for a variety of reasons in a variety of fields, there has been a considerable growth in the number of studies on the issue. As an illustration of the areas in which these studies are employed and their intended usage, consider the following systems [1]:

- **Education:** It is difficult to understand the mental state of students learning through online platforms. SER can help to change the approach of teaching to make learning effective.
- **Automobiles:** The driver's mental state and driving performance are usually linked with emotions. Analyzing emotions while driving can help to reduce road accidents.
- **Security:** By recognizing severe emotions like fear and anxiety, they can be employed as assistance tools in public areas.
- **Communication:** It can help to enhance and improve customer service in call centres when integrated with interactive voice assistants and can also help in mobile voice assistants.
- **Health:** People with autism can use SER to understand their feelings and accordingly adjust their social behaviour.

Speech Emotion recognition is a sophisticated and demanding task that requires a high degree of accuracy. Numerous studies were conducted on various aspects that have a significant influence on emotion. Despite this, there is no optimal feature set for correctly classifying emotion.This is because the presence of different speakers with varying speaking patterns, speaking speeds, and distinct languages affects the recognition rate. Due to age, gender, society, and surroundings, each speaker's speaking style are distinct [2].

When compared to performed speech, emotions expressed through natural speech appear to be more difficult to detect. Spectral characteristics such as Linear Predictive Coding (LPC), Linear Prediction Cepstral Coefficients (LPCC), Mel-Frequency Spectrum Coefficients (MFCC) and its derivatives, as well as prosodic features like energy and pitch, are widely utilized features [2]. Along with feature extraction, various classification techniques are used to classify the emotion from human speech. Primarily, this paper discusses three techniques: Multilayer Perceptron (MLP Classifier), Convolutional Neural Network (CNN), Support vector machines (SVM).

Another classic network, the convolutional neural network (CNN), has been shown to capture compact representations from task-specific knowledge extracted from annotated input. CNN is being used in image classification, object identification, speech acoustic modelling, text classification, and some other fields because of its advantage. Several academics have looked into the use of CNNs to learn emotional representation from voice signals in the area of speech emotion recognition.

AI and ML approaches such as Multilayer Perceptron (MLP classifier) is used to organise the data into nonlinearly segregated groups. The MLP classifier is set up using mel-frequency cepstrum coefficients (MFCC), chroma, and mel characteristics. MLPs (Multi-Layer Perceptron) are increasingly being employed for voice recognition and other speech processing applications. The most popular method of converting a voice signal captured by a phone into a string of characters is speech recognition. They can also contribute to further phonetic processing in order to achieve speech comprehension [11].

Another classification technique that is believed to be best for classification is SVM. SVM is commonly used for classification problems because of its good recognition rate with a limited amount of training data and its simplicity. The SVM classifier is based on the idea of determining the most effective hyperplane for separating data points from different classes. This maximizes the margin between the two classes. The hyperplane separates data points linearly in space which is not possible when the data is non-linearly separable. For this, SVM uses kernel functions that convert original feature space into a high-dimensional feature space that allows data points to be separated linearly. A single SVM is a classification algorithm for binary category data. There are often several emotion classes in speech emotion recognition. Therefore, SVM should be generalised to answer multi-class problems.

II. LITERATURE SURVEY

2.1 Convolutional Neural Network (CNN):

The identification of emotion-specific characteristics is a key difficulty with a robust Speech Emotion Recognition(SER) system. Deep learning models can be used rapidly since there is so much information available for training. Because of their ability to remove exceptional highlights in photos without the need for human assistance, CNNs were initially used as the core of computer vision systems. A spectrogram is a graphical portrayal of sound impulses, these neural structures are now utilized to identify and classify key examples in sound documents. When compared to typical feedforward neural networks, CNNs have two distinct advantages: parameter sharing and dimension reduction. CNN has fewer trainable properties as a result of these guideline features, reducing deep network training time even more. The model will retrieve the emotion-specific properties once the organization begins back-propagation [3].

Researchers have used a variety of methods in the literature to increase CNN performance in Speech Emotion Recognition, and they have outperformed numerous commonly used Speech Emotion Recognition features. PCA (principal component analysis) was employed in a method to extract meaningful speech data using spectrograms in a CNN architecture. And for five emotional classes, on the IEMOCAP database, the classification accuracy was 40.02%. For final feature extraction, two distinct descriptors have been combined using CNN architectures. On IEMOCAP and Berlin EmoDB, the classification accuracies are 86.36% and 91.78%, respectively [3].

In [4], a CNN-based emotion identification system was introduced that does not require any input data stream preprocessing and is computationally efficient. This work's singularity is that it has achieved supremacy in terms of exactness over other concurrent and comparable works.

2.2 Multilayer perceptron Classifier (MLP):

A Multilayer Perceptron is a classification system designed to mimic the action of neurons in the nervous system. The data is delivered as an audio input, which is then processed and put through a series of steps before being used to develop a model. Audio characteristics are retrieved from this data utilising different methods such as framing, hamming, windowing, and so on. We have 24 speakers in the RAVDESS database (12 male and 12 female). The model is trained after feature extraction, and MLP is utilised for classification. This classifier is used to categorise and classify the numerous categories in the dataset into emotions [11].

MLPs are able to adequately estimate any continuous values which are non-linear on a small period. There are a number of applications of Multilayer Perceptron such as regression and classification [12]. The neurons in MLP are organised in the layers, which start with the input layer and move forward via hidden layers and finally to the last layer i.e output layer. The network is feedforward of multiple because the two nearby levels may join. It is made by organising various neurons which are tightly linked with each other that may be employed concurrently in order to find a solution to a specific problem.

When it is used to solve non-linear optimization problems with several minimas that are local, it may not perform well. It may also overfit small datasets, making it sensitive to data size. On data that is not scaled, MLP gives 65.1% and 53.3% for training and testing data respectively. Furthermore, MLP achieved 100% accuracy for the training dataset and 75% accuracy for the testing dataset when using standard scaled data [14].

2.3 Support vector machine (SVM):

The Support Vector Machine is a technique for supervised learning. Simplicity along with good accuracy with limited training data makes SVM a widely used technique. SVM classifier finds a hyperplane that linearly separates data points of different classes. Between two classes, the hyperplane selected by the classifier must maximize the margin. There are two types of data points accessible for classification: Non-linearly and linearly separable. For Linearly separable data, to maximize the margin between classes, the normal vector to the hyperplane should be minimized. For non-linearly separable data, SVM makes use of kernel functions that convert the original input feature space into a high-dimensional feature space that makes data linearly separable. There exist different kernel functions that can be used with SVM. The paper [7] studies the performance of different kernel functions. It shows that Linear, RBF, Polynomial and Sigmoid kernel function gives an accuracy of 64.77%, 79.55%, 78.41% and 78.41% respectively when used for multiclass SVM.

The accuracy of emotion recognition using SVM like other techniques depends on features extracted from input speech data and the speaker's style of speaking. Along with this there exist different methods to implement SVM for multiclass classification which also contributes to the performance of the classifier. Different approaches to implement speech emotion recognition systems using SVM are developed by considering different combinations of features set, databases and methods of implementing SVM.

Features considered for implementation are of two types: Prosodic and spectral features. Both contain useful information of emotion associated with the speech. Energy, Mel-energy spectrum dynamic coefficients (MEDC), Pitch, Fundamental Frequency (F0), Mel-frequency spectrum coefficients (MFCC) are among some commonly used features. When spectral and prosodic features are utilized to build the classifier yields a higher recognition rate and reduction in error rate [8].

SVM with Linear kernel function and extracted features set as MFCC, Pitch, Energy provided an accuracy of 95.83% for self-built Malayalam language database, 73.75% for EMO-DB database and 61.25% for SAVEE database in [2]. The SVM classifier in paper [9] used fundamental frequency, energy and MFCC as feature sets and showed an accuracy of 89.80% for EMO-DB, 93.57% for Japanese Emotional Speech Database and 98.00% for the Thai Emotion database. Paper [10] extracted MFCC, MEDC and energy features and the SVM classifier showed a recognition rate of 91.3% and 95.1% for the Chinese emotional database and the EMO-DB database respectively.

III. RESEARCH METHODOLOGY

3.1 Feature Extraction

3.1.1. MFCC

Among various audio features, Mel-frequency cepstral coefficients (MFCC) proves to be the most prominent one. The cepstrum of a windowed short-time signal is obtained from the signal's FFT in this depiction of voice signal. Following that, the signal is transformed to the mel frequency scale's frequency axis that is decorrelated using only a modified Discrete Cosine Transform preceded by a log-based transform. Some of the techniques involved in obtaining MFCC features are DTC, Mel filter-bank, frame blocking and windowing, log energy, FFT magnitude and pre-emphasis. MFCC employs the mel-scale, which is calibrated to the frequency response of the human ear. As a result, MFCC is shown to be useful in the domain of speech recognition and is merged with emotion recognition. Spectral audio characteristics like MFCC, are considered ideal for an N-way classifier[11]. The block diagram of MFCC is illustrated in Figure 3.1.1.

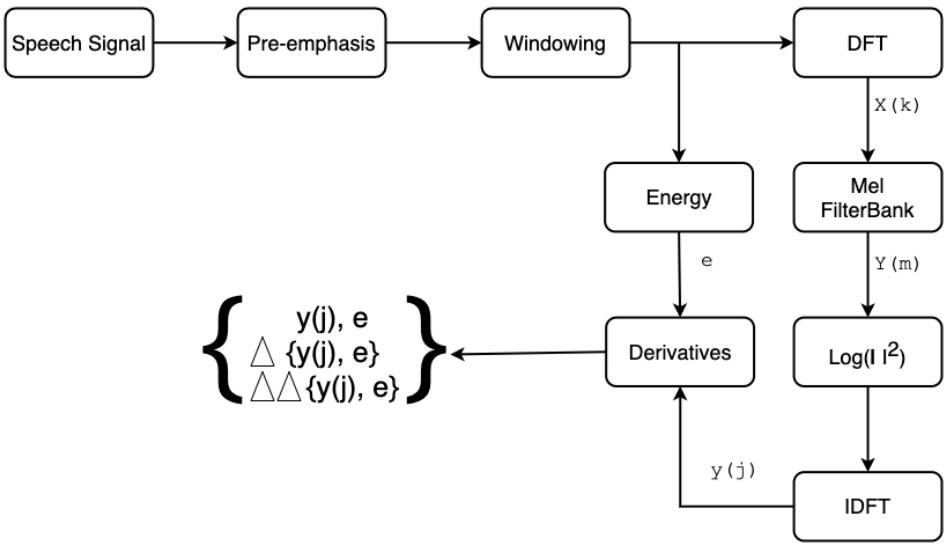

fig. 3.1.1: mel-frequency cepstral coefficients block diagram

3.1.2. Pitch

The mechanical movement of the human vocal cords describes pitch, which is the basic frequency found in human speech. The auto-correlation technique of pitch extraction is used to retrieve it [2].

3.1.3. Energy

Another prosodic element that communicates change in emotion is energy. An input speech signal's energy is a depiction of its amplitude fluctuations. It gives a description of the intensity based on the enthusiasm of emotions. Each frame determines the transient energy, as well as the logarithm of the mean energy's first and second derivatives. Then, for the entire speech, the maximum, mean, minimum, standard deviation, and range of energy are computed. The spoken signal is framed with a 20 ms length and a 10 ms overlap to extract energy. Then, to retain the consistency of the start and final points in the frame, each frame is multiplied by a hamming window, windowing is achieved. Then, using the expression, energy is determined [2].

3.2. Convolutional Neural Network (CNN)

According to [4], the following methodology can be used to prepare CNN Model for Speech Emotion Recognition.

3.2.1. Data Preprocessing

Two methodologies have been looked at for this reason. In the feature extraction-based process, modulation spectral features and Mel-frequency cepstral coefficients (MFCC) have been chosen to extract the emotional features as mentioned in [5]. The Fourier transform as well as the energy spectrum for each frame have been calculated and matched onto the Mel-frequency scale. Modulation Spectral Features (MSF) are calculated by duplicating the human auditory system's spectro-temporal (ST) processing and taking regular acoustic frequency and modulation frequency into account. An auditory filter bank decomposes voice signals first (19 filters in total). Modulation spectra are the spectral contents of modulation signals, and as a result the proposed characteristics are called Modulation Spectral Features (MSFs). MFCC and ST computations can be done as mentioned in [5]. There is no need to preprocess data while using the Deep Convolutional Neural Network (CNN) approach. Raw audio data is fed directly into the neural network model to train the model.The Recursive Feature Elimination method is used to remove the least important features by using the basic Linear Regression [4].

3.2.2. Model Architecture of Deep Convolutional Neural Network

Except for the last convolutional layer, the CNN model has complete one-dimensional convolutional layers, each followed by a max-pooling layer having a pool size 2 preceded by batch normalisation layer. The model was fed raw audio with an 8-second length. Zero-padded the audio document with a duration of less than 8 seconds. Convolutional Layers have 32, 64, 128, 256, 512, 1024, and 1024 filters, respectively. Kernel sizes of those filters are 21, 19, 17, 15, 13, 11, and 9 accordingly. A global max-pooling layer follows the last one-dimensional convolutional layer, after that a dense layer with 128 nodes is added. 'Relu' is the actuation function for all dense and convolutional layers so far. The model's last layer is a dense layer with 7 nodes (due to the dataset's total number of excited classes being 7) and the activation function 'softmax' [4]. The model architecture of CNN is illustrated in Figure 3.2.2.

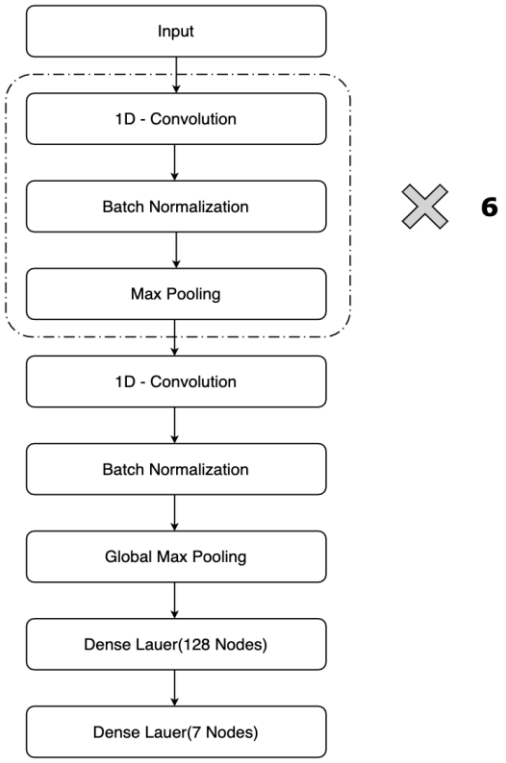

fig. 3.2.2: convolutional neural network

With a 0.001 initial learning rate, a 0.9 beta 1 value, and a 0.999 beta 2 value, the Adam [6] optimizer was used to train the model. 400 epochs were used to train the model.

The following Table 3.2.2 shows the recognition table for the Deep Convolutional Neural Network approach using test data [4]

table 3.2.2: test data

Class	Precision	Recall	F1-Score
Surprise	0.84	0.87	0.86
Happiness	0.83	0.87	0.85
Fear	0.85	0.80	0.82
Disgust	0.82	0.81	0.82
Anger	0.85	0.73	0.79
Sadness	0.83	0.90	0.86
Total	0.84	0.84	0.84

3.3. **Multilayer perceptron Classifier (MLP)**

The Multilayer Perceptron Classifier (MLP Classifier) is a classification algorithm that divides data input into categories. The speech signals are divided into mel-frequency cepstrum coefficients (MFCC), chroma, and mel highlights, which are then utilised to create the MLP classifier. The collected features, together with the emotion category to which they belong, should be saved in respective arrays as data input to the model so that the classifier may find patterns, connections, and finally categorise the input [13].

MLP recognises several classes in the inputs and categorises them into various emotions. The trained model is used to recognise the categories of speech attribute values that correspond to different emotions. If we provide the model an unknown test dataset as an input, it will extract the parameters and predict the emotion based on the values in the training dataset [11]. The architecure of MLP is illustrated in Figure 3.3.

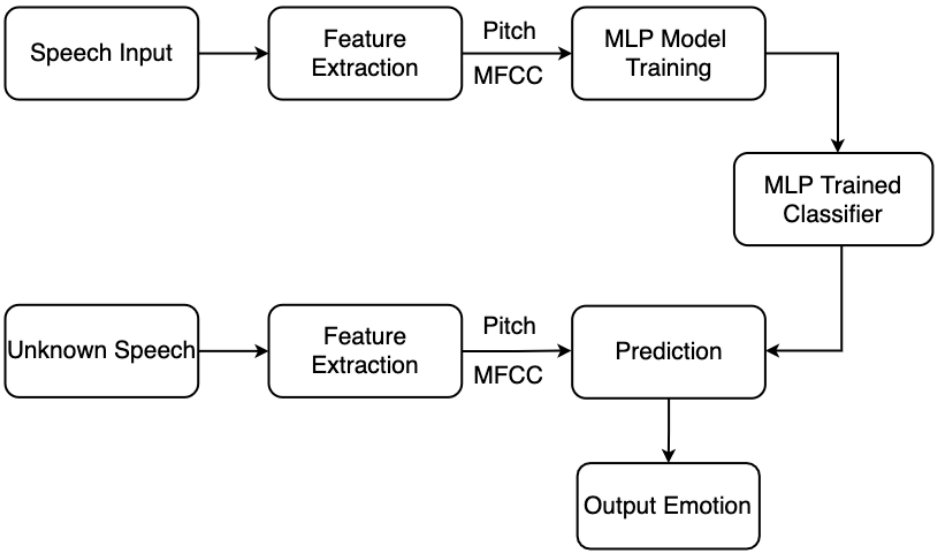

fig. 3.3: Multilayer perceptron Classifier

As mentioned in [11], Speech emotion recognition using MLP can be accomplished by following the step:

3.3.1. Pre-processing:

To prepare your data for a neural network, you should first set it up and it should be numeric. If you have unambiguous data, such as a gender feature with the parameters "male" and "female," or emotion features such as happy, sad, angry, disgust, fear, and so on, you may convert it to a genuine form known as a OneHot encoding.

3.3.2. Model Training

The characteristics retrieved, together with the emotion class to which they correspond, should be saved in corresponding arrays as input to the model so that the classifier may find patterns, connections, and ultimately categorise the data. This training aids the model in determining which emotions have which variety of characteristics. As a result, it will be able to compare and predict emotion when given unknown data as an input.

3.3.3. Prediction

A neural network may be used to make various predictions after it has been trained. You may evaluate the model's performance on new dataset by making predictions on test data.

3.4. Support Vector Machine (SVM)

The architecture of human speech emotion recognition as mentioned in [2] composed of fundamental building blocks: Speech input, Feature extraction, Feature labelling, and SVM classifier. This architecture is illustrated in Figure 3.4.

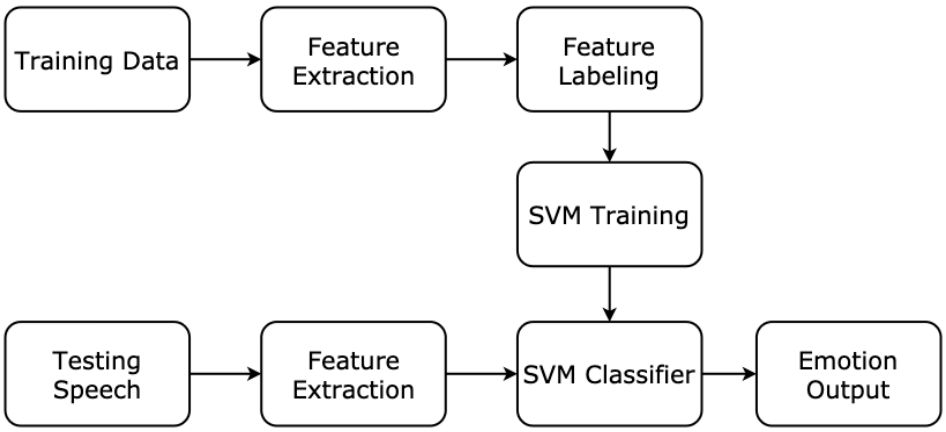

fig. 3.4: Support Vector Machine

3.4.1. Input

The audio file of a speech sample from the database serves as the system's input. These samples are divided into two groups: testing and training set. The SVM classifier is trained by utilizing the samples in the training set by extracting emotion features from them and labelling them by the corresponding emotion class. After training the SVM classifier, samples from the testing set are given as input to the system from which features are extracted and provided to train the classifier which predicts the emotion present in the sample as output.

3.4.2. Feature Extraction

This is the important stage of the system, in this stage, a set of features containing emotion are extracted from the input sample. Emotion in speech is identified from the features associated, selection of optimum features which contributes in the recognition of emotion most should be selected for higher accuracy.

3.4.3. Feature Labelling

After the extraction of the feature vector for the input sample, they are labelled to their corresponding emotion class. This information is stored in the database which will be used for training SVM.

3.4.4. SVM Classifier

SVM is a supervised learning algorithm that requires labelled data points as input. Labelled data points created in the previous step satisfies this requirement. While building the classifier these inputs are used to train the classifier and while testing or using the classifier these inputs are used to produce output i.e., to classify the speech. To use SVM for multiclass classification, various methods are implemented such as One versus the rest, One against one and Binary tree. Also, SVM using linear kernel function and RBF are tested for accuracy.

Recognition accuracy produced by this methodology as mentioned in [2] is mentioned in Table 3.4.4.

table 3.4.4: recognition accuracy

Database	Kernel	Classification Strategy	Percentage Accuracy(%)
Self-Built Malayam Language Database	Linear	One against one	95.83
Emo-DB(General)	Linear	One versus the rest	73.75
SAVEE	Linear	Binary Tree	61.25

3.5 Databases

To build speech emotion recognition systems various databases used are mentioned in the following table 3.5.

table 3.5: databases for speech emotion recognition

Database	Language	Type	Size	Emotions
SAVEE(SurreyAudio-Visual Expressed Emotion) [2]	English	Acted	7 Emotions, 4 male Speakers, 120 utterances per speaker [4]	Surprise,sadness, fear, disgust, happiness, and anger.
RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [11]	English	Acted	7 emotions, 24 speakers (12 males, 12 females).	Calm, happy, sad, angry, fearful, disgust and surprise.
EMO-DB (Berlin Emotional Database) [2]	German	Acted	7 emotions, 10 utterances, 10 speakers (5 male and 5 female)	Disgust, anger, happiness, boredom, neutral, sadness, fear
Audiovisual Thai Emotion Database. [9]	Thai	Acted	6 emotions, 6 speakers	Surprise, sadness, anger, disgust, fear, happiness

IV. Results and Discussions

Table IV shows a quick comparison of standard classifiers and deep learning classifiers that have been built in recent years.

table 4: comparison table

Classification Methods	Dataset	Feature Extraction method	Accuracy
Convolutional Neural Network (CNN) [4]	SAVEE(SurreyAudio-Visual Expressed Emotion.[4]	Mel-frequency cepstral coefficients (MFCC), Modulation Spectral Features(MSF)	83.61 %
Multilayer Perceptron (MLP) [11]	RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) [24]	Mel-Frequency Cepstral Coefficients (MFCC)	75 %
Support Vector Machine (SVM)[2]	EMO-DB (Berlin Emotional Database[19]), SAVEE(SurreyAudio-Visual Expressed Emotion)[2]	Mel-Frequency Cepstral Coefficients (MFCC), Pitch, Energy	89.8 %

V. Conclusion

The fundamental purpose of Speech Emotion Recognition (SER) is to plan capable and vigorous techniques to predict emotions. In this paper, we have primarily studied three classification techniques used for SER systems: Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Multilayer Perceptron (MLP). Various Feature extraction methods used for SER systems are Energy, Modulation Spectral Features (MSF), Pitch, Mel-frequency cepstral coefficients, Fundamental Frequency (F0), Mel-energy spectrum dynamic coefficients (MEDC). It was found that implementation of SVM using Linear kernel function and feature set as MFCC with Energy and pitch, fundamental frequency or with MEDC gives the best performance. Also if spectral and prosodic features used together can reduce the error rate. Convolutional Neural Networks (CNN) classifier is considered to be a singular-based neural network speech-emotion recognition procedure and probably a better alternative to other traditional methods. It can classify 7 types of emotions with greater accuracy considering the SAVEE database. Also, it is discovered that for emotion recognition, MLPs are incredible classifiers with vocal-dependent execution that gives better classification in song contrasted with speech channel.

REFERENCES

[1] Hadhami Aouani and Yassine Bed Ayed Learning Salient Features for Speech Emotion Recognition Using CNN. 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia).

[2] M. S. Sinith, E. Aswathi, T. M. Deepa, C. P. Shameema and S. Rajan, "Emotion recognition from audio signals using Support Vector Machine," 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 2015, pp. 139-144, doi: 10.1109/RAICS.2015.7488403.

[3] Speech Emotion Recognition Using Convolution Neural Networks. In International Conference on Artificial Intelligence and Smart Systems (ICAIS-2021). IEEE Xplore Part Number: CFP21OAB-ART; ISBN: 978-1-7281-9537-7.

[4] Alif Bin Abdul Qayyum, Asiful Arefeen, Celia Shahnaz Convolutional Neural Network (CNN) Based Speech-Emotion Recognition.2019IEEE International Conference on Signal Processing, Information, Communication & Systems(SPICSCON).28-30 November 2019, Dhaka, Bangladesh.

[5] Kerkeni, Leila and Serrestou, Yousse fand Raoof, Kosaiand CIMer, Catherine and Mahjoub, Mohamed andMbarki, Mohamed, ”Automatic Speech Emotion Recognition Using Machine Learning;’ March 2019.

[6] Kingma, Diederik and Ba, Jimmy, "Adam: A method for stochastic optimization," International Conference On Learning Representations, December 2009.

[7] Vaishali M. Chavan, V.V. Gohokar, 2012, ”Speech Emotion Recognition by using SVM-Classifier”, International Journal of Engineering and Advanced Technology, IJEAT, Vol. 1, Issue 5.

[8] Y. Zhou, Y. Sun, J. Zhang and Y. Yan, "Speech Emotion Recognition Using Both Spectral and Prosodic Features," 2009 International Conference on Information Engineering and Computer Science, 2009, pp. 1-4, doi: 10.1109/ICIECS.2009.5362730.

[9] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using Support Vector Machines," 2013 5th International Conference on Knowledge and Smart Technology (KST), 2013, pp. 86-91, doi: 10.1109/KST.2013.6512793.

[10] Yixiong Pan, Peipei Shen and Liping Shen, “Speech Emotion Recognition Using Support Vector Machine”, International Journal of Smart Home, Vol. 6, No. 2, April, 2012

[11] Sanjita. B. R, Nipunika. A, Rohita Desai, “Speech Emotion Recognition using MLP”, International Journal of Engineering Science and Computing, Vol 10, Issue 5, May 2020.

[12] S. K. Pal and S. Mitra, "Multilayer perceptron, fuzzy sets, classification" 1992.

[13] U. Raghu Vamsi, B. Yuvraj Chowdhary, M. Harshitha, S. Ravi Theja, Divya Udayan J, “Speech Emotion Recognition(SER) using Multilayer Perceptron and Deep learning techniques”, High Technology Letters, Volume 27, Issue 5, 2021.

[14] Behzad Javaheri, “Speech \& Song Emotion Recognition Using Multilayer Perceptron and Standard Vector Machine”, Department of Computer Science, City University of London, London, UK.