Journal of Theoretical and Applied Information Technology
28th February 2023. Vol.101. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

SPEECH EMOTION RECOGNITION SYSTEM PERFORMANCE ANALYSIS WITH OPTIMIZED FEATURES USING DIFFERENT CLASSIFICATION ALGORITHMS

KOGILA RAGHU[1,2*], MANCHALA SADANANDAM[1]

[1] Department of CSE, Kakatiya University, Warangal 506001, TS, India
[2] Department of CSE, Geethanjali College of Engineering and Technology, Hyderabad 501301, TS, India

Corresponding Author Email: kraghu.cse@gcet.edu.in

ABSTRACT

Affective computing is becoming increasingly significant in the interaction between humans and machines. Emotion recognition of spoken language is a hot topic in Human Computer Interaction (HCI). Understanding a person's physical and mental state could be greatly aided by learning to identify the emotions conveyed in their speech. There are a number of practical uses for emotion recognition from Speech and it has much interest research domain in recent years. Many of the current options, however, are not yet suitable for use in production environments. The system is divided into three phases: features extraction, features selection/dimensionality reduction and classification. The first step involves extracting a wide variety of features, including prosodic and spectral components, Speech and glottal-waveform signals are used to construct long-term statistics. To tell apart associated emotions is a major challenge for SER systems. These features improve the SER's capacity to distinguish between emotions in speech. Inevitably, this high-dimensional feature vector will contain some repetition. In the second phase, the dimensionality of the feature vectors is reduced through the application of feature selection method such as Auto-Encoder technique described by the authors. Next, several classifiers, including K-Nearest Neighbour (K-NN), Logistic Regression (LR) , Support Vector Machines (SVM), Multi-Layer Perceptrons (MLP) and Convolutional Neural Networks (CNNs) are used to the optimize feature vector in stage three. Two widely-cited datasets serve as the basis for experimental evaluations of method efficacy. The Database for Emotions in Telugu Language (DETL) of a native Telugu Language and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) of English Language are two such collections. Experimentation is carried out with the different Feature Extraction methods such as MFCC, MFCC+ΔMFCC+ΔΔMFCC, MFCC+ΔMFCC+ΔΔMFCC+Pitch+ZCR (41-dimension) and Optimized Features (30-dimension) along with the different Classification Methods. From the results, it is observed that the suggested method has potential for use in speech emotion recognition. Furthermore, when compared to other methods in the context, our approach is clearly superior in terms of classification performance accuracy rate. Here CNN model got highest accuracy with Optimized Features for RAVDESS 95.42% and for DETL 98.37%.

Keywords: *Affective Computing, HCI, Features, K-NN, LR, SVM, MLP, CNN, RAVDESS, DETL.*

1. INTRODUCTION:

Speaking is the most efficient and fundamental way for humans to communicate with one another. Vibrations of the vocal cords are produced by air passing via the trachea from the lungs to the larynx, thereby generating speech signals [1]. In recent years, researchers have paid a lot more attention to the study of recognizing emotions in people's voices through speech. The automatic recognition and assessment of human emotions is one of the cutting-edge study areas in disciplines as diverse as biomedical sciences, psycho-physiology, computer sciences, and machine intelligence [2]. To a big extent, an observer may make sense of an individual's emotions and many other significant attitudes by simply seeing their behavior. Human emotions are constructed from the ground up by all of the physical actions that make up a person, including speech, facial expressions, and physical movement [3]. A

person's identity and emotional condition are transmitted through the speech to those surrounding them [4]. People's reliance on smart technologies is growing, and with that comes the need for faster and more efficient data processing, which has led to a rise in the prevalence of emotional recognition in human-computer interactions.

Fundamentally, autonomous speech emotion identification systems replicate human emotions on a computer, and then use spectrum-based features to match aspects of the speech signal, such as intonation, emphasis, and pauses, to the desired emotions. The three main components of any speech emotion detection system are the following: (1) speech emotion corpus, (2) emotions features extraction, and (3) emotions classification [5]. Emotions may display inherent diversity and Variation caused by people's physical conditions and external circumstances, as will be discussed in the course of the speech. Therefore, two essential elements related to emotion identification are a robust architecture and emotion of speech features relevant to knowledge. E-learning, Voice surveillance, clinical investigations, deception detection, computer games, entertainment and call centers are just few of the application areas where speech emotion recognition would be useful. However, there is no powerful machine learning algorithm is found to recognize the emotions. There have been a lot of studies and developments in emotion identification in recent years, but the best approach is still unknown. Emotional subjectivity is to blame for this predicament. When we say that emotions are subjective, we imply that various people can identify the same feelings in different ways, which can lead to confusion when trying to establish a universally accepted classification scheme. Also, figuring out which emotional features are the most useful is difficult task. Additionally, feature sets are not allocated, which is preset for emotional recognition [6]. Real-world noise in sound recordings can affect a machine learning model's performance [7]. Features characterizing the acoustic content of speech are retrieved in traditional methods of emotion recognition in speech is also important while conducting Speech Emotion Recognition (SER). Several different machine learning methods are used to make sense of the connections between the speech data collected and the emotions that were assigned to them in advance. SVM, HMMs and neural networks are popular methods used in SER. While neural networks and hidden Markov models offer superior predictions, they are difficult to build and train. SVMs are an appealing middle ground. Furthermore, a lot of time and processing power are needed. Face recognition, voice recognition for IoT devices, and emotion recognition in speech are just some of the recognition challenges that have been solved with deep learning models in recent years [8–10]. SER (Speech Emotion Recognition) has been implemented using a variety of feature vectors and different methodologies.

A growth in the amount of studies on this topic has been attributed to the fact that emotion recognition detection techniques can be applied in numerous contexts. The following systems serve as illustrations of how and where these studies are put to use.

• Education: Distance education course systems can identify bored students and adjust the style or level of material presented, as well as provide emotional incentives, or compromises, to keep them engaged.

• Automobile: The driver's emotional condition and driving performance are often intertwined. The driving experience and driving performance are both enhanced by using these systems.

• Security: Capable of detecting strong emotions like fear and anxiety, they can be useful as safety nets in public places.

• Communication: An interactive voice response system, automatic emotion recognition can assist call center's provide better customer service.

• Health: Autism People, generally wear gadgets with portability because they can't explain their own emotions and feelings. SER kind of systems may be able to change their social conduct accordingly.

The system's performance depends on the choice of feature vectors and models for representing features. Following is a brief summary of this paper's significant findings and contributions:

- Acoustic features such as MFCC, ΔMFCC and ΔΔMFCC spectral features, as well as prosodic features

Pitch, ZCR are used to create new feature vectors (MFCC+ΔMFCC + ΔΔMFCC + Pitch + ZCR) of 41-dimension, and Optimized features of 30-dimension using Auto-Encoders during the feature extraction step. These characteristics improve the ability to recognize the same emotions expressed by various speakers. In addition, by reducing overlap between features, they will help enhance speech/text categorization accuracy.

- Auto-Encoders are employed at the feature selection step, also known as the dimension reduction stage, to pick out the most salient features of Original Data. Here Optimized Features are extracted with 30-dimension.

- K-NN, LR, SVM, MLP, and CNN are only few of the classifiers utilized in the classification phase.

Our most recent findings suggest that, when compared to existing state-of-the-art systems, the proposed Speech Emotion identification system excels on the RAVDESS dataset and DETL dataset with Optimized Features. Here are the sections of the article: The related study on speech emotion recognition is presented in Section 2. The system we envision is outlined in Section 3. Section 4 summarizes the findings of the experiment. Finally, in section 5, we come to an end to the study.

2. RELATED WORK:

As a result of numerous investigations, it is possible to identify emotional states from aural data. After identifying the features of an emotional speech corpus that has been selected or developed, it is then used to categorize the emotions based on those extracted characteristics. Effectiveness in classifying emotions is extremely parasitic on the quality of character traits that are extracted. In the Human-Computer Interaction (HCI) field, speech emotion identification is regarded as a difficult problem. There have been numerous methods and corpora proposed in the past [10–12]. In the aboriginal stages of SER research, classical machine learning models were trained using handcrafted speech characteristics and low-level feature descriptors. The researchers analyzed a wide range of factors, including the prosodic features, spectral features and the hybrid combination of these traits, in their investigation. There has been a recent uptick in curiosity of deep neural networks (DNNs). But there are two fundamental problems with DL methods: (1) an adequate amount of speech data with labels. (2) Emotional content can be extracted from sounds. Multiple approaches have been investigated to deal with the lack of training data. This problem can be tackled in one of three ways. (a) Obtaining and annotating new information. Creating a large enough dataset, on the other hand, is both costly and time consuming. (b) The addition of new information to existing records. This is the most widely used method in DL research [14]-[15]. (c) Apply what you've learned. This is a well-known DL research subject that aims to store and then apply the knowledge gained from training one model to another. A variety of fields [16]–[19] have benefited from its implementation. However, the lack of improvement in the SER system's accuracy is due to a mismatch between the datasets. Many new DL algorithms have been presented in modern era in an attempt to improve the extraction of features. SER tasks have also seen the introduction of auto encoders following their success in the DL industry. In order to rebuild the input with the least amount of reconstruction error possible, an unsupervised learning model known as a "auto encoder" is utilized. All three layers of an auto encoder are contained in one device. Using nonlinear mapping, an auto encoder maps the input vector to the optimal latent representation, and from this representation, the input vector may be reconstructed. If the network has more than one hidden layer, it is considered deep. Basic auto encoders' latent representation has been used extensively in earlier SER studies.

Based on the study of visual and aural information, Noroozi et al. developed a comprehensive emotion identification system. His study used 88 features (Mel Frequency Cepstral Coefficient (MFCC), filter bank energies (FBEs)) Using PCA to extract features reduces the dimension of previously derived features. [20].

Bandela et al. used the Berlin Emotional Speech database to detect five unique emotions [21].For the Logistic Model Tree approach (LMT), Zamil et al. employed the 13 MFCC generated from audio data, with an accuracy rate

of 70 percent, to categorise the seven emotions [22]. Ignored features have been ignored throughout all of this work. Using these methods, the accuracy can't go above 70%, which can have an impact on the ability to discern emotional content in speech. However, the Features of energy like filter bank energies (FBE) and the Teager Energy Operator (TEO) are not widely accepted. Many researchers agreed MFCC,ΔMFCC,ΔΔMFCC, TEO and ZCR are the most crucial auditory characteristics to discern emotions [23].

For instance, in [24], a multilayer perceptron SER with deep auto encoder was proposed. Pal and Baskar [25] suggested a multilayer perceptron based on a deep dropout auto encoder. For the purpose of dimensionality reduction, [26] and [27] also used an auto encoder to identify the bottleneck features.

Finally, to classify emotions mostly Machine Learning Algorithms like K-NN, LR, SVM, Multi Layer Perceptron (MLP) and Convolution Neural Networks (CNN) were used. MFCCs with their derivation, Pitch and ZCRs are often used features in emotion detection systems.

3. PROPOSED METHODOLOGY:

In this paper, we describe an Emotion Recognition System from Speech Utterances that uses 41 features from MFCC 39, Pitch, ZCR. We initially use an Auto-Encoder (AE) to extract the essential parameters from previously extracted parameters before using different classifiers to compare the performance of SER systems. Fig. 1 depicts the planned architecture.

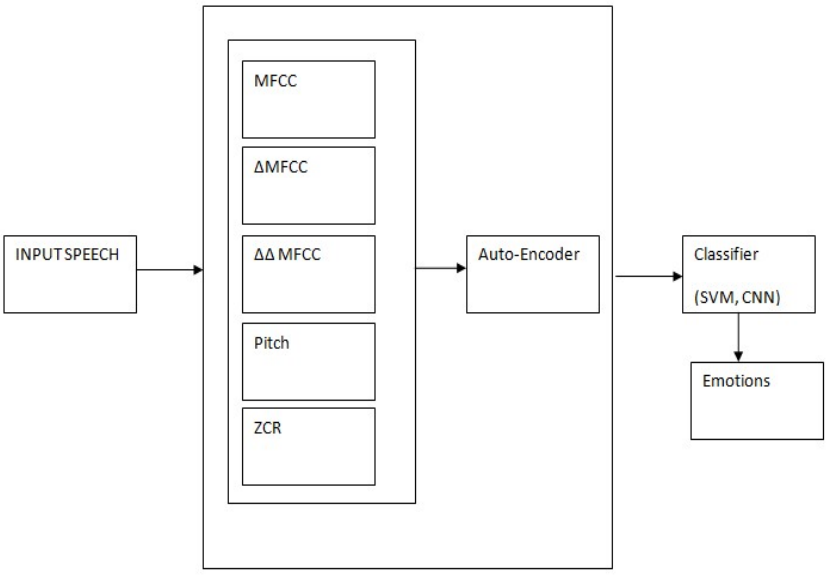

Figure 1: Proposed System Architecture

Spectral Feature MFCC and their derivatives, the other prosodic features like ZCR and Pitch, were utilized to extract feature vectors as feature extraction. Algorithms use feature vectors to classify data emotions, Classification Algorithms used are SVM and CNN. This necessitates the usage of auto-encoder just like feature selection or dimensionality reduction method for emotion recognition in terms of improving the performance of classifiers outputs.

Feature Extraction:

The proposed work uses 41 Features as a total, in that 39 MFCC (12 MFCC + energy, 12 Δ MFCC + energy and 12 ΔΔ MFCC + energy) and ZCR.

Mel-Frequency Cepstral Coefficient (MFCC)

MFCC coefficients are calculated by using the logarithm of the Fast Fourier Transform (FFT) module and a Mel-scale filter and the Inverse Fast Fourier Transform (IFFT).

Journal of Theoretical and Applied Information Technology
28th February 2023. Vol.101. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

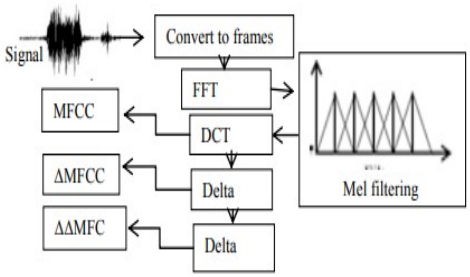

Figure 2: MFCC Process

To locate MFCCs, simply follow the methods outlined below. In the figure, these steps are depicted (Fig.2.) The first step is to perform a FFT on the signal. Then the spectrum's power is mapped on to the Mel scale. In the next step, Every Speech signal Mel Frequency logarithm power is calculated. Then Mel log power banks are taken for Discrete Cosine Transformation. The log Mel spectrum is reverting back with the time domain in this final phase. The Mel frequency cepstrum coefficients are the name given to the resulting data set (MFCC). The formula for calculating the MFCC coefficients is as follows:

$$\Delta Cep(i) = \alpha \sum_{j=1}^{2} j\left(Cep(i+j) - Cep(i-j)\right)$$

Where α is a constant ≈ 0.2 and Cep denotes the MFCC coefficients.

The ΔΔMFCC is calculated with the formulae:

$$\Delta\Delta Cep\ (i) = \Delta Cep\ (i+1) - \Delta Cep\ (i-1)$$

Zero Crossing Rate: ZCR (Zero Crossing Rate) is a remarkable quantity in speech recognition systems. Zero crossings are counted and distributed by the number of samples in a given region, and this is what the term implies.

$$z(m) = \sum_{n=-\infty}^{\infty} \left\|\operatorname{sgn}(s(n)) - \operatorname{sgn}(s(n-1))\right\| w(m-n)$$

$$\operatorname{sgn}(s(m)) = \begin{cases} 1 & if\ s(m) \geq 0 \\ -1 & if\ s(m) \prec 0 \end{cases}$$

Feature Dimensionality Reduction:

Reduced feature dimensions are achieved by retaining most prominent features with feasible while reducing the feature dimensions. Two methods exist for reducing feature dimensions: selection and extraction. As part of a feature selection process, this paper uses an auto-encoder (AE). Like the multi-layer ANN, this method uses a non-recursive feed forward method to train. This work uses AE to learn a more succinct data representation. There are input, output, and hidden layers [28]. To see this, look at Figure 3 where the number of nodes in the output layer (x1') is equal to the input layer (x1). AE absorbs the weight vector (W, W') assuming that the output layer vector is same as the input layer vector. An auto-parameters encoder's include the number of hidden layers, units per layer, weight regularization, and iterations.

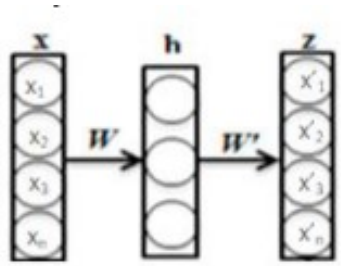

Figure 3: Auto-Encoder Basic Architecture

In the proposed methodology, SER involves two phases namely Feature Extraction and designing a reference model using various Classification Algorithms.

Feature Vectors play a prominent role in SER for increasing accuracy and reducing time complexity. In this work, spectral features are extracted from shortest duration of speech signals of huge speech training corpus. Windowed speech signals are used to generate 13-dimensional MFCC features, 13-dimensional ΔMFCC features, and 13-dimensional ΔΔMFCC feature vectors in this study. Single spectral (13 MFCC) and its derivative features (13 ΔMFCC, 13 ΔΔMFCC) are extracted in the Feature Extraction Process. Figure.2: shows how acoustic and prosodic features can be used to create new feature vectors.

Then these MFCC and its derivative features are concatenated to form 41-dimensional

Feature vector (13 MFCC + 13 ΔMFCC + 13 ΔΔMFCC + Pitch + ZCR).

In second phase of SER system, a reference model with various Classification algorithms like K-NN, LR, SVM, MLP and CNN are designed with training feature vectors of 41-dimension and 30-dimension.

Optimized Features Speech Emotion Recognition

By reducing the number of features in each feature vector using various dimensionality reduction techniques, we can expedite training and improve the model's capacity to recognise the emotions of any language. Because there are so many feature vectors, the model takes a long time to train and test. The size of the feature vector can be decreased using techniques like PCA and rough set theory, it has been established.

Additionally, the Auto-Encoder (AE) is effective at minimizing the number of dimensions. The 41-dimensional new feature vector (MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR) is reduced to 30 dimensions using Auto Encoder in this study. Figure.4 shows the Optimized Feature Extraction Process.

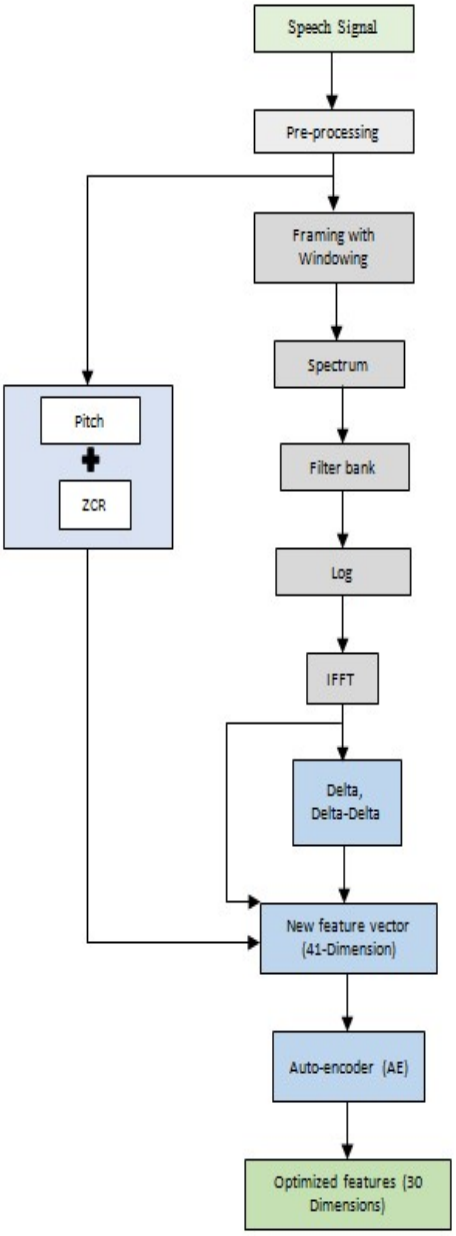

Figure.4: Optimized Feature Extraction

4. RESULTS AND DISCUSSION:

In this research work, experiments are carried out on two databases RAVDESS (Ryerson Audio Visual Database of Emotions in Speech and Song) and DETL (Database of Emotions in Speech and Song). Ryerson University, Canada, created RAVDESS, a simulated speech emotion database for English language speech [29]. There are 1440 .wav files containing emotive speech samples, with 60 utterances in each category. This speech features

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

24 actors, 12 of them are male and 12 of whom are female. A sample rate of 44.1 kHz is used for the recordings. DETL is a database of the native Telugu language that we have produced in this work [30]. Telugu speakers aged 20 to 70 years old have contributed emotional speech samples to the database. The Speech Utterances were gathered from a variety of sources, such as You Tube, movies, news readings, spontaneous speech, and others. For the most part, this database is made up of five categories of emotions: Anger, Fear, Happy, Sad and Neutral.

Table 1: Performance Comparison Of SER System Using Different Models With New Features (41-Dimension) And Optimized Features (30-Dimension) For RAVDESS

Classification Model	**Accuracy (%)**	
	MFCC+ΔMFCC+ΔΔMFCC+Pitch+ZCR (New Features)	**MFCC+ΔMFCC+ΔΔMFCC+Pitch+ZCR (Optimized Features)**
K-NN	59.87	61.32
LR	70.23	73.16
SVM	71.11	73.43
MLP	77.23	79.27
CNN	93.31	95.42

From the above results, it is noticed that the performance of SER system for RAVDESS Database by using the K-NN model got accuracies of 59.87%, 61.32%, by using LR model got accuracies of 70.23%, 73.16%, by using the SVM model got accuracies of 71.11%, 73.43%, by using MLP model got the accuracies of 77.23%,79.27% and the CNN model got accuracies of 93.31%, 95.42% with New Features of 41-dimension, Optimized Features of 30-dimension respectively in each Classification model.

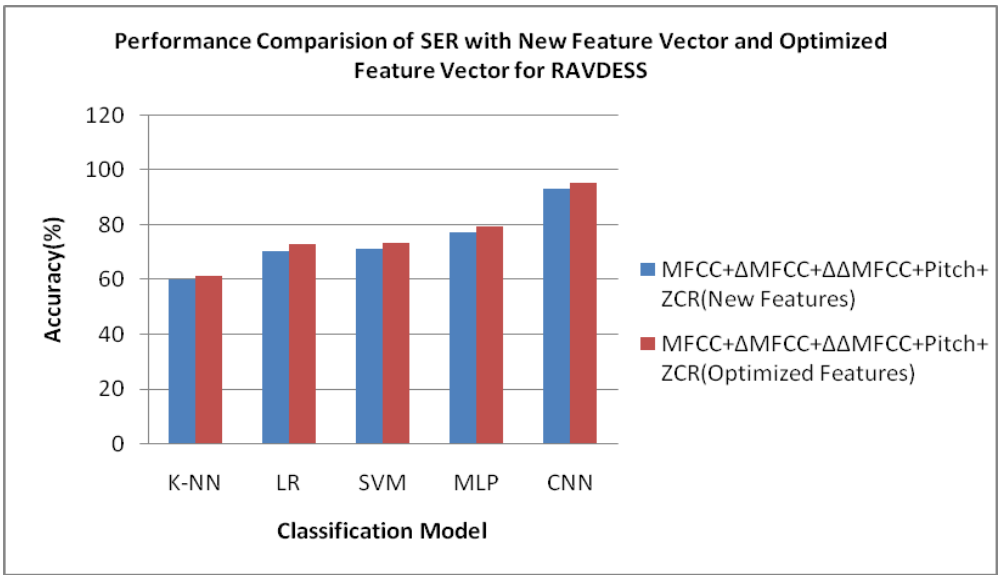

Figure.5: Performance Comparison Of SER System With New Features And Optimized Features For RAVDESS

From the Figure.5, it is noticed that the CNN model has got highest Accuracy than K-NN, LR, SVM and MLP, as illustrated. CNN model has 95.42% for RAVDESS Database with Optimized MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR feature extraction.

Table 2: Performance Comparison Of SER System Using Different Models With New Features (41-Dimension) And Optimized Features (30-Dimension) For DETL

Classification Model	**Accuracy (%)**	
	MFCC+ΔMFCC+ΔΔMFCC+Pitch+ZCR (New Features)	**MFCC+ΔMFCC+ΔΔMFCC+Pitch+ZCR (Optimized Features)**
K-NN	61.54	64.92
LR	75.27	77.36
SVM	76.24	77.54
MLP	83.61	88.35
CNN	96.78	98.37

From the above results, it is noticed that the performance of SER system for RAVDESS Database by using the K-NN model got accuracies of 61.54%, 64.92%, by using LR model got accuracies of 75.27%, 77.36%, by using the SVM model got accuracies of 76.24%, 77.54%, by using MLP model got the accuracies of 83.61%, 88.35% and the CNN model got accuracies of 96.78%, 98.37% with New Features of 41-dimension, Optimized Features of 30-dimension respectively in each Classification model.

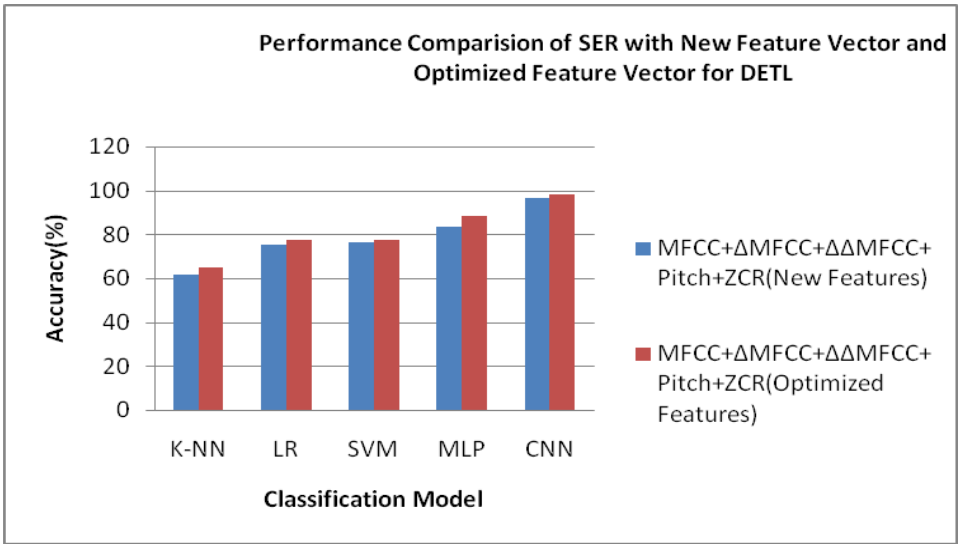

Figure 6: Performance Comparison Of SER System With New Features And Optimized Features For DETL

From the Figure 6, it is noticed that the CNN model has got highest Accuracy than K-NN, LR, SVM and MLP, as illustrated. CNN model has 98.37% for DETL Database with MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR feature extraction.

Table 3: Performance Comparison Of Different Models In SER With Different Features

Model	Feature Vector	Accuracy (%)	
		RAVDESS	DETL
K-NN	MFCC+ΔMFCC +ΔΔMFCC	53.66	56.13
	New Feature Vector	59.87	61.54
	Optimized Feature Vector	61.32	64.92
LR	MFCC+ΔMFCC +ΔΔMFCC	68.68	73.23
	New Feature Vector	70.23	75.27
	Optimized Feature Vector	73.16	77.11
SVM	MFCC+ΔMFCC +ΔΔMFCC	70.73	74.36
	New Feature Vector	71.11	76.24
	Optimized Feature Vector	73.43	77.54
MLP	MFCC+ΔMFCC +ΔΔMFCC	75.54	81.32
	New Feature Vector	77.23	83.61
	Optimized Feature Vector	79.27	88.35
CNN	MFCC+ ΔMFCC +ΔΔMFCC	91.64	95.26
	New Feature Vector	93.31	96.78
	Optimized Feature Vector	95.42	98.37

From the results, the performance of K-NN, LR, SVM, MLP and CNN model with optimized feature vectors is 61.32%,73.16%,73.43%,79.27% and 95.42% for RAVDESS Database respectively, Whereas 59.87%,70.23%,71.11%,77.23% and 93.31% with new features for RAVDESS only.

From the results, the performance of K-NN, LR, SVM, MLP and CNN model with optimized feature vectors is 64.92%,77.11%,77.54%,88.35% and 98.37% for DETL Database respectively, Whereas 61.54%,75.27%,76.24%,83.61% and 96.78% with new features for DETL only.

The corresponding graphs are depicted in Figure 7 and Figure 8 for RAVDESS and DETL respectively.

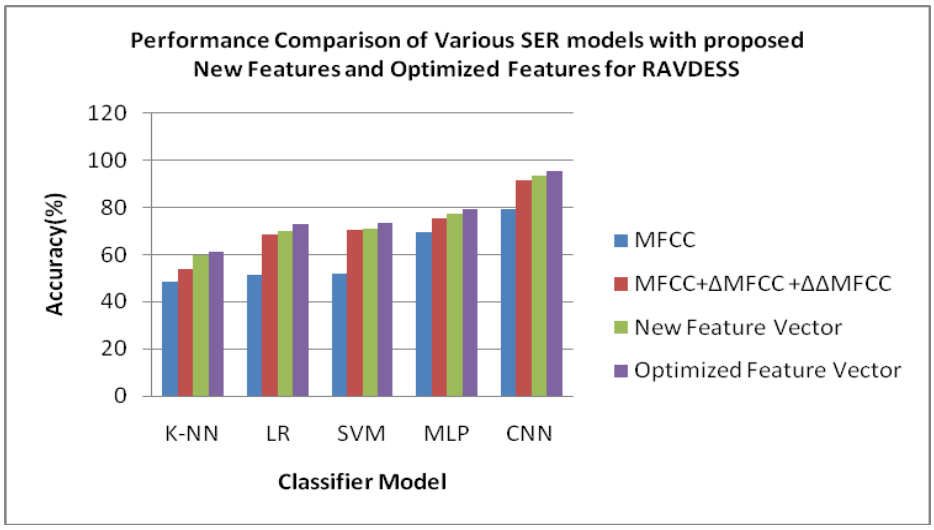

Figure .7: The Performance Comparison Of K-NN, LR, SVM, MLP And CNN Using Proposed New Features And Optimized Features For RAVDESS

From Figure.7 , it is observed that performance of CNN is impressive with optimized features. Overall, CNN model with Optimized (30-dimentsional) MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR features provided good performance for English language RAVDESS Database with 95.42%.

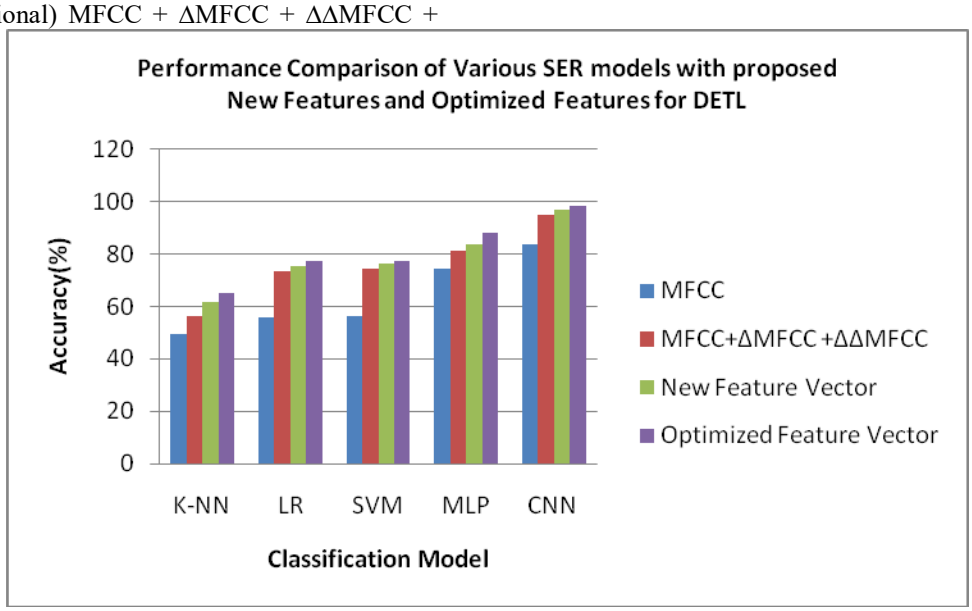

Figure 8: The Performance Comparison Of K-NN, LR, SVM, MLP And CNN Using Proposed New Features And Optimized Features For DETL.

From Figure 8, it is observed that performance of CNN is impressive with optimized features. Overall, CNN model with Optimized (30-dimentsional) MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR features provided good performance for Telugu language DETL Database with 98.37%.

Table 4: Comparison Of Proposed Method With Recent Published Work

No	Reference	Year	Database	Methodology	Accuracy (%)
1.	Deng et. Al [30]	2013	Emo-DB	SVM	51.62
2.	Deb et. Al [31]	2016	Emo-DB	SVM	74.64
3.	Zhang et. Al [32]	2016	RAVDESS	Binary Classifier	64.32
4.	Deng et. Al [33]	2017	GeWEC	Universem AE	59.27
5.	Mirsamadi et.al [34]	2017	IEMOCAP	RNN	63.47
6.	Tomba et. Al [34]	2018	RAVDESS	SVM CNN	78.68 89.21
7.	Jannat et. Al [35]	2018	RAVDESS	SVM	66.47
8.	Bhavan et.al [36]	2019	RAVDESS	SVM	75.69
9.	Hadhami et.al [37]	2021	RAVDESS	SVM	74.07
10.	Husam et.al [38]	2021	RAVDESS	NN	86.1
11.	Zhu-Zhou et.al [39]	2022	Berlin	Logistic Regression	91
12.	Proposed Method (New Features- 41 Dimension)	2022	RAVDESS DETL	MLP	77.23 83.61
13.	Proposed Method (New Features- 41 Dimension)	2022	RAVDESS DETL	CNN	93.31 96.78
14.	Proposed Method (Optimized Features-30 Dimension)	2022	RAVDESS DETL	MLP	79.27 88.35
15.	Proposed Method (Optimized Features- 30 Dimension)	2022	RAVDESS DETL	CNN	95.42 98.37

It is observed that the proposed model with optimized feature vectors which are derived from Optimized Features 30-dimension (MFCC + ΔMFCC + ΔΔMFCC + Pitch + ZCR) is performed well compared to existing works and even New Feature vector 41-dimension also produced good results compared to existing works.

5. CONCLUSION:

In this research, Speech Emotion Recognition (SER) system based on the hybrid Feature Extraction with fusion using 39 MFCC (13 MFCC + 13 ΔMFCC + 13 ΔΔ MFCC), ZCR and Pitch as a total of 41 Feature Vector is used then Robust Optimized Features are selected by applying Auto-Encoders Technique, which reduces dimensionality and selects prominent features. After that, a model is built using different Classification Algorithms like K-Nearest Neighbour (K-NN), Logistic Regression (LR), Support Vector Machines (SVM), Multi-Layer Perceptron (MLP) and Convolution Neural Network (CNN) to recognize the emotions of test speech samples. This work used two Speech Emotion Databases i.e, RAVDESS, an English Speech Emotion Database and DETL, a native Telugu Speech Emotion Database. It is obvious from findings of these systems that with Auto-Encoders perform better than without Auto-Encoders using different Classification Techniques.

Here Auto-Encoder Technique reduces the dimensions and enhances the accuracy rate as showed in the Results section. And also, Comparison study of proposed system and Existing Related System work mentioned. In Future, the SER system may consider other kinds

of Features and can apply the proposed technique to bigger database and also it may use other methods for dimensionality reduction and selection. The features or descriptors from speech and facial expression images can be used in this situation to be benefit from the emotions recognition by using audio-visual databases also in near future.

REFERENCES:

[1]. P. Schlegel, S. Kniesburges, S. Dürr, A. Schützenberger, and M. Döllinger, ''Machine learning based identification of relevant parameters for functional voice disorders derived from endoscopic high-speed recordings,'' Sci. Rep., vol. 10, no. 1, p. 10517, Jun. 2020, doi: 10.1038/s41598-020- 66405-y.

[2]. M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, ''Recognizing emotions induced by affective sounds through heart rate variability,'' IEEE Trans. Affect. Comput., vol. 6, no. 4, pp. 385–394, Oct. 2015, doi: 10.1109/TAFFC.2015.2432810.

[3] H. Gunes and M. Piccardi, ''Bi-modal emotion recognition from expressive face and body gestures,'' J. Netw. Comput. Appl., vol. 30, no. 4, pp. 1334–1345, Nov. 2007, doi: 10.1016/j.jnca.2006.09.007.

[4] D. Połap, ''Model of identity verification support system based on voice and image samples,'' J. Univers. Comput. Sci., vol. 24, pp. 460–474, Jan. 2018.

[5] G. Lu, L. Yuan, W. Yang, J. Yan, and H. Li, ''Speech emotion recognition based on long short-term memory and convolutional neural networks,'' J. Nanjing Univ. Posts Telecommun., vol. 38, no. 5, pp. 63–69, Nov. 2018, doi: 10.14132/j.cnki.1673-5439.2018.05.009.

[6] V. Garg, H. Kumar, and R. Sinha, ''Speech based emotion recognition based on hierarchical decision tree with SVM, BLG and SVR classifiers,'' in Proc. Nat. Conf. Commun. (NCC), Feb. 2013, pp. 1–5, doi: 10.1109/ncc.2013.6487987.

[7] K. Han, D. Yu, and I. Tashev, ''Speech emotion recognition using deep neural network and extreme learning machine,'' in Proc. Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 1–5.

[8] S. Mittal, S. Agarwal, and M. J. Nigam, ''Real time multiple face recognition: A deep learning approach,'' in Proc. Int. Conf. Digit. Med. Image Process. (DMIP), 2018, pp. 70–76, doi: 10.1145/3299852.3299853.

[9] H.-S. Bae, H.-J. Lee, and S.-G. Lee, ''Voice recognition based on adaptive MFCC and deep learning,'' in Proc. IEEE 11th Conf. Ind. Electron. Appl. (ICIEA), Jun. 2016, pp. 1542–1546, doi: 10.1109/iciea.2016.7603830.

[10] K. R. Malik, M. Ahmad, S. Khalid, H. Ahmad, F. Al-Turjman, and S. Jabbar, ''Image and command hybrid model for vehicle control using Internet of Vehicles,'' Trans. Emerg. Telecommun. Technol., vol. 31, no. 5, p. e3774, 2019, doi: 10.1002/ett.3774.

[11] Issa, D., Demirci, M.F. and Yazici, A., 2020. Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, p.101894.

[12] X. Zhao, S. Zhang, and B. Lei, ''Robust emotion recognition in noisy speech via sparse representation,'' Neural Comput. Appl., vol. 24, nos. 7–8, pp. 1539–1553, Jun. 2014, doi: 10.1007/s00521-013-1377-z.

[13] T. Özseven, ''A novel feature selection method for speech emotion recognition,'' Appl. Acoust., vol. 146, pp. 320–326, Mar. 2019, doi: 10.1016/j.apacoust.2018.11.028.

[14] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, ''Bagged support vector machines for emotion recognition from speech,'' Knowl.-Based Syst., vol. 184, Nov. 2019, Art. no. 104886, doi: 10.1016/j.knosys.2019. 104886.

[15] S. Li and L. Xu, ''Research on emotion recognition algorithm based on spectrogram feature extraction of bottleneck feature,'' Comput. Technol. Dev., vol. 27, no. 5, pp. 82–86, 2017.

[16] P. Tzirakis, J. Zhang, and B. W. Schuller, ''End-to-end speech emotion recognition using deep neural networks,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp. 5089–5093, doi: 10.1109/icassp.2018.8462677.

[17] K. Khanchandani and M. Hussain, ''Emotion recognition using multilayer perceptron and generalized feed forward

neural network,'' J. Sci. Ind. Res., vol. 68, pp. 367–371, Apr. 2009.

[18] D. Gharavian, M. Sheikhan, and F. Ashoftedel, ''Emotion recognition improvement using normalized formant supplementary features by hybrid of DTW-MLP-GMM model,'' Neural Comput. Appl., vol. 22, no. 6, pp. 1181–1191, May 2013, doi: 10.1007/s00521-012-0884-7.

[19] Er, M.B., 2020. A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. IEEE Access, 8, pp.221640-221653.

[20] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., &Anbarjafari, G. Audio-visual emotion recognition in video clips. IEEE Transactions on Affective Computing,2017.

[21] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.

[22] Zamil, AdibAshfaq A., et al. "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST).IEEE, 2019.

[23] D. Kaminska, T. Sapi ´ nski, and G. Anbarjafari, "Efficiency of chosen ´ speech descriptors in relation to emotion recognition," EURASIP Journal on Audio, Speech, and Music Processing, 2017.

[24] N. E. Cibau and M. L. Enrique Albornoz Hugo Rufiner, "Speech emotion recognition using a deep autoencoder,'' Anales de La XV Reunion de Procesamiento de la Informacion y Control, vol. 16, pp. 934–939, May 2013.

[25] A. Pal and S. Baskar, "Speech emotion recognition using deep dropout autoencoders,'' in Proc. IEEE Int. Conf. Eng. Technol. (ICETECH), Mar. 2015, pp. 1–6.

[26] K.-Y. Huang, C.-H.Wu, T.-H.Yang, M.-H.Su, and J.-H. Chou, "Speech emotion recognition using autoencoder bottleneck features and LSTM," in Proc. Int. Conf. Orange Technol. (ICOT), Dec. 2016, pp. 1–4.

[27] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang, and S. Shang, "Research on speech emotion recognition based on deep auto-encoder," in Proc. IEEE Int. Conf. Cyber Technol. Autom., Control, Intell. Syst. (CYBER), Jun. 2016, pp. 308–312.

[28] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoisingautoencoders," in Proceedings of the 25th International Conference on Machine Learning, ser. ICML '08. New York, NY, USA: ACM, 2008, pp. 1096–1103.

[29] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

[30] Raghu, K., Sadanandam, M. (2021). A perspective study on speech emotion recognition: Databases, features and classification models. Traitement du Signal, Vol. 38, No. 6, pp. 1861-1873. https://doi.org/10.18280/ts.380631.

[31] Deng J, Zhang Z, Marchi E, Schuller B (2013) Sparse autoencoder based feature transfer learning for speech emotion recognition. In: 2013 Humaine association conference on affective computing and intelligent interaction, pp 511–516.

[32] Deb S, Dandapat S (2016) Emotion classification using residual sinusoidal peak amplitude. In: 2016 International conference on signal processing and communications (SPCOM), pp 1–5.

[33] Zhang B, Provost EM, Essl G (2016) Cross-corpus acoustic emotion recognition from singing and speaking: a multi-task learning approach. In: 2016 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 5805–5809.

[34] Deng J, Xu X, Zhang Z, Frühholz S, Schuller B (2017) Universum autoencoder-based domain adaptation for speech emotion recognition. IEEE Signal Process Lett 24(4):500–504.

[35] Mirsamadi S, Barsoum E, Zhang C (2017) Automatic speech emotion recognition using recurrent neural networks with local attention. In: 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp 2227–2231.

[36] Tomba K, Dumoulin J, Mugellini E, Khaled OA, Hawila S (2018) Stress detection through speech analysis. In: Proceedings of the 15th International joint conference on e- Business and telecommunications—Volume 1: ICETE, INSTICC, SciTePress, pp 394–398. https://doi.org/10.5220/0006855803940398 .

[37] Jannat R, Tynes I, Lime LL, Adorno J, Canavan S (2018) Ubiquitous emotion recognition using audio and video data. In: Proceedings of the 2018 ACM international joint conference and 2018 international symposium on pervasive and ubiquitous computing and wearable computers, association for computing machinery, New York, NY, USA, UbiComp'18, pp 956–959. https://doi. Org/10.1145/3267305.3267689.

[38] A. Bhavan, P. Chauhan, Hitkul, and R. R. Shah, "Bagged support vector machines for emotion recognition from speech," Knowl.-Based Syst., vol. 184, Nov. 2019, Art. no. 104886, doi: 10.1016/j.knosys.2019. 104886.

[39] Husam Ali Abdulmohsin a, Hala Bahjat Abdul wahab b ,Abdul Mohssen Jaber Abdul hossen c, "A new proposed statistical feature extraction method in speech emotion recognition", Computers and Electrical Engineering 93 (2021) 107172, https://doi.org/10.1016/j.compeleceng.2021 .107172 .

[40] Zhu-Zhou, F.; Gil-Pita, R.; García-Gómez, J.; Rosa-Zurera, M. Robust Multi- Scenario Speech-Based Emotion Recognition System.Sensors 2022, 22, 2343.https://doi.org/ 10.3390/s22062343.