A New Approach for Speech Emotion Recognition using Single Layered Convolutional Neural Network

Mannar Mannan. J[1], V Vinoth Kumar[2], Shivakumara Palaiahnakote[3], Surbhi Bhatia Khan[3], Ahlam Almusharraf[4]

[1]Dept. of Information Science and Engineering, CMR Institute of Technology, Bengaluru. Email: man.endeavour6381@gmail.com

[2]School of Computer Science Engineering & Information Systems (SCORE), Vellore Institute of Technology, Email: vinothkumar.v@vit.ac.in

[3]School of Science, Engineering and Environment, University of Salford, United Kingdom, Email: s.palaiahnakote@salford.ac.uk, s.khan138@salford.ac.uk

[4]Department of Business Administration, College of Business and Administration, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia, Email: Aialmusharraf@pnu.edu.sa

Corresponding Author: Shivakumara

Abstract

Creating a computational device to identify human emotions via voice analysis represents a notable achievement in the sector of human-computer interaction, especially within the healthcare domain. We propose a new light-weight model for addressing challenges of emotions recognition. The model works based on CNN with change of kernel processing. The proposed model performs a direct matching to recognize speech emotions of different eight categories using a statistical model named Analysis of Variance (ANOVA) as kernel for features extraction and Cosine Similarity Measurement (CSM) as activation function for CNN model. This proposed model contains eight-folded single-layered intermediate neurons, and each neuron can segregate speech emotion pattern using CSM from the voice convergence matrix to explore a part of the solution from the whole solution. Experiment results demonstrates that the proposed model outperforms compared with multiple layered existing CNN methods in identifying the emotional state of a speaker.

Keywords: Analysis of Variance; Speech Emotion Recognition; Deep Learning; CNN; Cosine-similarity measurement.

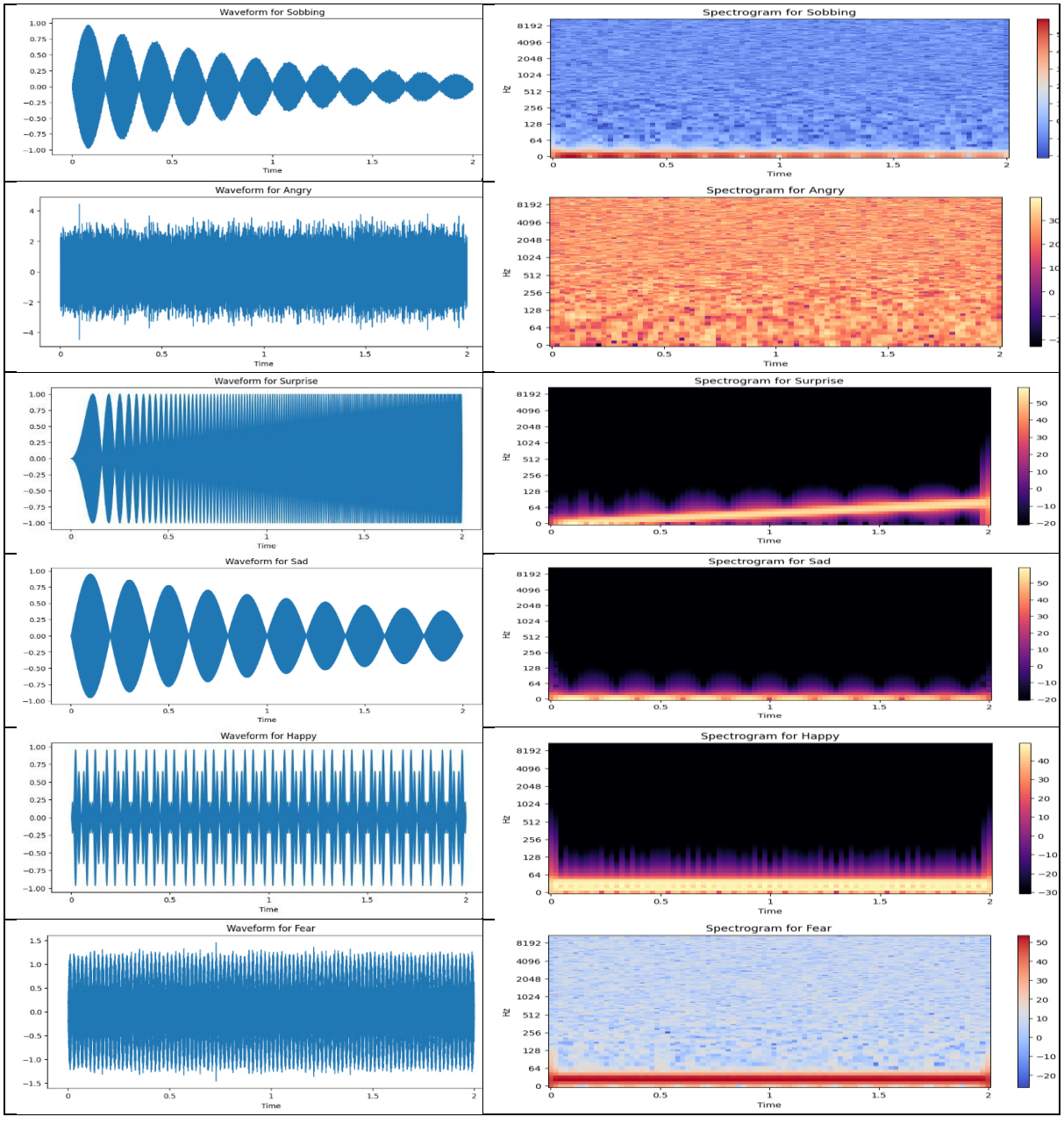
1. Introduction

Speech Emotion Recognition (SER) is a crucial task in medical field, child health and customer satisfaction analysis in business. The outcome of the last three decades of research and many successful voices have been processed, even though the performance of the Speech Emotion Recognition (SER) system could not be equalized to human performance. Dozens of algorithms had come up with different parameters for SER, however speech emotion recognition still not yet reached the high end. SER is a challenging issue since the machine can fine-tune its performance by learning itself from the history to recognize emotion states of a speaker either on live or recorded audio stream, and it is an alternative to face emotion recognition system for recognizing state of an individual. Unlike human learning, machine learning models need a vast input data to provide more accurate results from the experience. ML model for detecting voice discrimination to classify manifold features requires huge data processing which is unrealistic in SER. A new SER model is required for SER to replace the large data processing and time-consuming ML based SER model.

SER can be used to identify the mental state of a child or old-age people to understand their mental state or diagnosis mental illnesses. The advent of the Fourier transform [28] and the inception of deep

learning in computer science have catalyzed the advancement of Speech Emotion Recognition (SER). This involves training on a "belief network" with multiple hidden layers to enhance performance. Among the spectrum of speech emotions, excitement levels in individuals can be discerned from their voice, correlating with arousal and valence in music data. Diverse selections of popular songs and music albums from various countries have been uniformly analyzed using regression theory [1] for recognizing emotional cues in music, employing a sampling rate of 22050Hz and 16 bits. Data dimension reduction and non-negative factorization [2] has been used for SER. Alternatively, Gaussian mixture model [26,27] for Speech Recognition (SR) is replaced by deep belief network [3], which uses pre-trained multiple layers of spectral feature vectors for better performance. An alternative to the pitch, duration, and frequency, the coefficient for speech recognition [35], the autoregressive model has been proposed [4] for better SER, which includes reflection coefficient along with linear prediction coefficient. Common feature extraction techniques for speech include Mel-frequency cepstral coefficients (MFCCs), filter banks, pitch, and energy. These features capture the spectral and temporal characteristics of speech signals and can be used as input to the CNN.

Figure 1 demonstrates all 8 types of emotions spectrogram as well as their waveform.



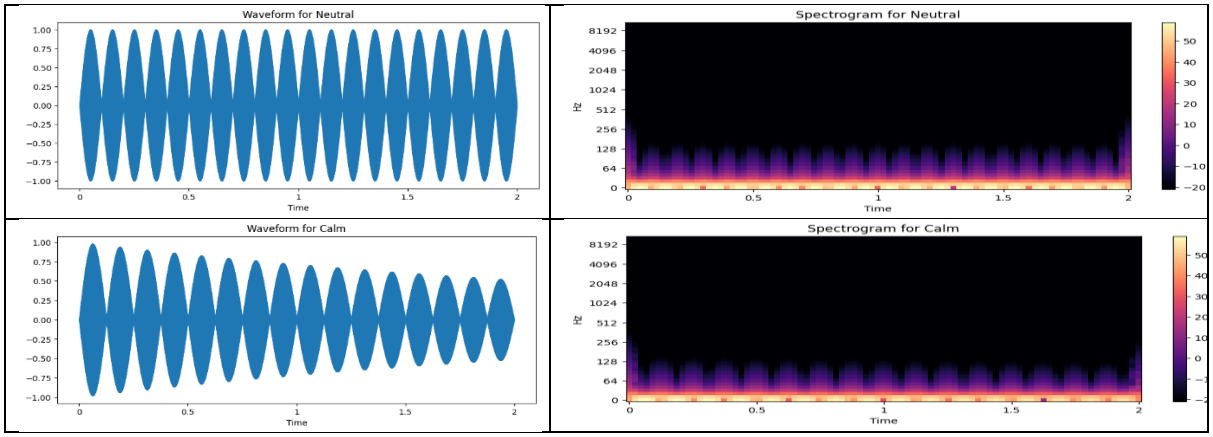


Figure 1: Spectrogram and Waveform of Emotions

Apart from all implementations for SER, design a realistic SER model is not yet attained due to huge amount of data and large computing requirements. To enhance the performance of the SER system to achieve SER, a direct method for recognizing speech emotions is required within the machine learning model. Modifying the functional model of CNN to compute the pattern of the different emotions directly that are already stored on the eight different low dimension matrix (LDM). The objective of the proposed research is to change the kernel function to recognize a speech emotion of a speaker. Here, we extracted eight different patterns of emotions irrespective of gender using ANOVA, and CSM technique is used as activation function for CNN to reduce the number of processing step for speedy response without compromise the performance.

The paper's structure is outlined as follows: Section 2 provides a comprehensive review of previous studies, serving as the Literature Review. Section 3 delves into the Proposed work. The findings and analysis are detailed in Section 4, and the concluding remarks are presented in the final Section 5.

2. Literature Review

The predecessor of SER is a sound, and sound analysis begins from environmental sound recognition by obtaining audio time-frequency features [7], and progressed towards voice segregation and classification [23], voice activity detection [24], speech diarization and now reached into SER using deep learning. Various automatic diarization methods [22] have been analyzed using clustering, classifying, and other mathematical models on voice data to recognize and convert into text form. ML algorithms can be used for classification, regression or clustering based on the identical characters of parameters of any data model. DL plays a major role in feature extraction from the voice data for human-computer interaction to recognize the psychological state of a speaker. Speech into text conversion and speech recognition are basic platform for SER. The performance of any speech recognition system is associated with phonetics of the corresponding language [32]. Japanese speech recognition has been evaluated [16] using an acoustic model. Computing machines can speak like as human from the text [12], vocabulary-speech recognition [5], voice activity detection using deep belief network for SER using multiple nonlinear hidden layers [6], SER using RNN [8], and the comprehensive review for SER using deep learning [9], speech motion recognition for human-machine(robot) interaction [29] are notable achievements in speech processing; since speaking while crying is untouched.

2.1 ML for SER

To enhance the performance of Speech Emotion Recognition (SER), a hybrid model [10] has been introduced, amalgamating the strengths of various models to achieve superior results. This model incorporates deep neural network (DNN), convolutional neural network (CNN), and recurrent neural network (RNN) to determine the optimal combination for SER. Integrating different methodologies for SER enables the aggregation of diverse features, thereby enhancing performance compared to individual SER methods with varying parameters. By aggregating DNN, CNN, and RNN, the model

aims to identify four emotions: anger, happiness, neutrality, and sadness [25]. A unified framework is proposed to leverage the interdependent features of DNN, CNN, and RNN, resulting in generalized features for SER. The model has attained a weighted accuracy of 57.1% and an unweighted accuracy of 58.3% over the IEMOCAP dataset [10].The empirical ML models for speech recognition are using CNN [25], multichannel hidden CNN-BLSTM for speech emotion recognition [34], Markov and Bayesian learning [17], which leads us to move towards SER. The contemporary part of speech recognition for speaker verification using cosine kernel estimation and cosine similarity [18], have achieved the best result on Linear Discriminant Analysis (LDA). It is an SVM based channel estimation technique with Join Factor Analysis (JFA). Speaker recognition under noisy conditions using pipelined manners [19], is subject to computer vision techniques using CNN for confirming the speaker using facial recognition on the videos. Speech Emotion Recognition (SER) is approached from various perspectives, incorporating techniques such as gated recurrent unit (GRU) and multi-head attention [37-39]. These methods have been applied to the IEMOCAP and Emo-DB corpora, resulting in improved performance. Furthermore, they have demonstrated enhanced efficacy on CH-SIMS and MOSI datasets. Speech recognition can be sped up by accent recognition. CNN has been used on five different language data sets such as Spanish, Mandarin, French, English, and Arabic to train CNN. The complete review of various SER and performance over the different data set is presented [40]. The performance of the SER using different ML is tabulated [41].

2.2 MFCC for SER

Numerous algorithms are available for Speech Emotion Recognition (SER), each with distinct parameters and features. For instance, the RNN classifier utilizing Mel frequency cepstral coefficients (MFCC) and modulation spectral features (MSFs) has shown superior performance on Berlin and Spanish datasets compared to Multivariate Linear Regression (MLR) and Support Vector Machine (SVM) [11]. The vast temporal voice data in digital form are represented in matrix takes more time to process SER. To reduce computing time of SER, Non-negative Matrix Factorization (NMF) can be used for matrix decomposition [30]. To improve performance, supervised model of NMF with join discrimination ability and similarity measurement has been introduced [13]. Learning local invariant features in (LIF) the two stage CNN model [14], uses unlabeled samples, and LIF is used as input to learn discriminative features through feature extractor. Generally, machine learning performance will increase when the data set grows and takes more time to process since there is more data. Instead of training individual speaker models, a public trained data set can be synchronized with an independent model to enhance the source speaker to sound like that of the target to ensure linguistic meaning [15]. Mel-Frequency Cepstral Coefficient (MFCC) is the best method to classify accents compared to Spectrogram, Chromogram and all the other methods [20]. The design and implementation of a deep learning model for extracting emotional patterns from speech data play a crucial role in feature extraction [31]. The IEMOCAP dataset [33] is utilized in Speech Emotion Recognition (SER), encompassing emotions such as anger, happiness, sadness, neutrality, frustration, excitement, fear, surprise, disgust, and others. In the IEMOCAP dataset, these emotions are typically classified into five major categories, comprising 1103 samples for anger, 1636 samples for happiness, and so forth. Mannan et al. [44] proposed a model for human emotions recognition based on convolutional neural network and Mel frequency cepstral coefficients. This approach uses CNN for capturing facial expression and MFCC us used to extract features using voice signals. Finally, the work fuses the facial expression features and audio features for emotions recognition. However, the performance of the method depends on two modalities. The gap between two different modalities sometime leads to poor performance compared to single modality as proposed work.

2.3 Other Models for SER

Clustering method for speech recognition [21] involved three steps such as 1) iteration for unknown word learning, 2) multi-probability normalization, and 3) filtering. To classify the emotion state of the speaker, the pattern of various emotions state found in the data should be identified and kept in LDM

as trained model. An approach is suggested that involves multi-level acoustic feature cross-fusion, with the objective of addressing the absence of information across different features by compensating for missing data. Alternatively, SER attempted using a graph-based approach. A mode speech data is analysed for SER [42]. Different SER with ML model is completely analysed from the different parameters [43].

In summary, knowledge revealed from the review of literature for SER models are broadly classified into two categories, such as SER with Neural Network (DNN, CNN, RNN and hybrid model), and MFCC, LDA, and cosine kernel estimation models. These two branches of SER included a pros and cons over the accuracy on various dataset taken for experiment. A common milestone yet to be achieved is SER. Large data processing and multiple layers on neural network-based models is not provided an optimal solution to SER. On the other side, computing model of MFC and LDA is not given accuracy of detecting corresponding emotion state of a speaker. To address these drawbacks, a new sophisticated direct SER model is needed to fine-tune the performance of the SER system to implement for SER. The proposed model fusions ML model for fine-tune each emotion pattern and utilizing CSM for recognize SER. This proposed model can be used in SER because of its reduced computing steps hens proved with multiple test cases.

3. Proposed Methodology

The objective of the proposed model is to fill up the gap in ML models for recognizing SER by developing a knowledge base to distinguish the discriminative and similarity information to SER system for detecting emotion patten of the speaker. This model synthesized a common difference between the normal voice and emotion voice by extracting the pattern using ANOVA as filter/kernel for CNN, and stored the features into the eight different LDM after maximum pooling. The entire architecture of the proposed Speech Emotion Recognition (SER) system is illustrated in Figure 3. The preliminary process of this model is to divides the audio data into smaller chunks of 128-bit equal sized blocks. Each block examines with different trained data set stored in the eight-folded LDM. The trained emotion patterns are stored in eight contagious blocks, and each emotion pattern kept on the different block.

Initially, analog voice data converted into digital form using sampling process shown in Fig. 2. To extract emotions pattern, two sets of analog voice data, such that normal speech data and various emotions voice data have taken. Both data sets are converted into digital form at 44.1 kHz or 44,100 samples per second and stored in two different low dimensional matrixes.

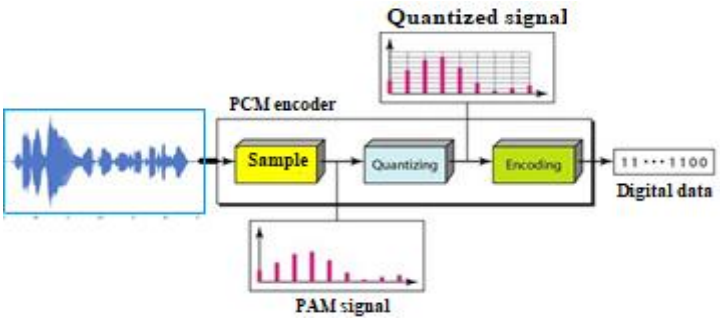


Fig. 2. Audio into digital conversion

3.1. Realtime emotion detection

The speech data is divided into continuous equal sized blocks. We have generated the trained set based on emotions and identified the pattern where the region of different emotions scattered on the vector space of speech data. Consider the three different patterns for low is P1, medium is P2, and high is P3. The voice of the speaker has classified into any one of the three-classification based on modulation calculated in Hz. The overall structure of the proposed model is portrayed in Figure 3.

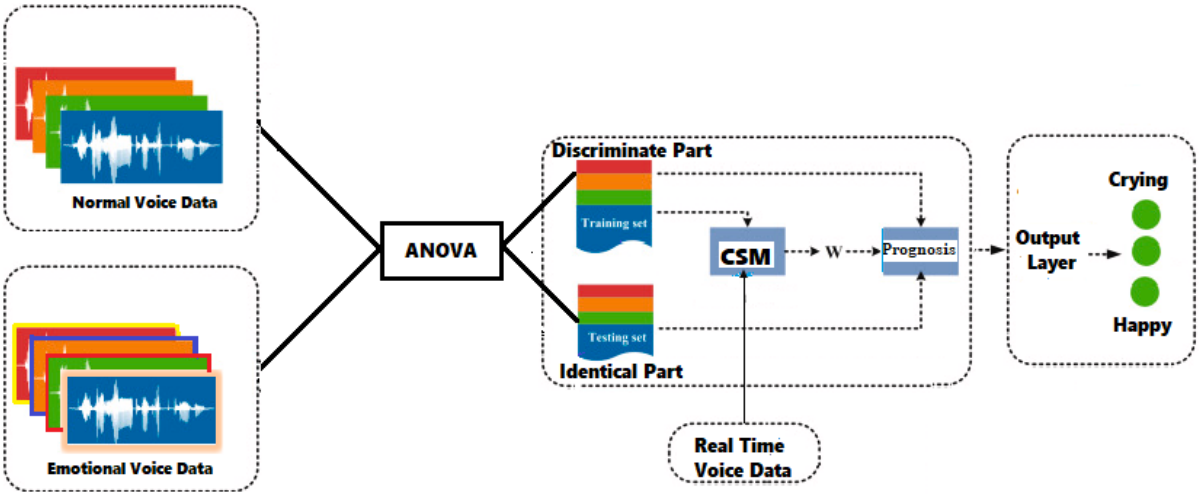


Fig. 3. Proposed SER System Architecture

Realtime speech data is divided into equal sized chunks referred as speech data blocks and any given speech dataset block *'ss[n]'*, can include the neutral region sp[n] and along with emotion region for depiction s[n] on the LDM.

$$ss[n] = sp[n] + s[n] \tag{1}$$

Supering neutral region *'sp[n]'* and extracting emotion region *'s[n]'* from the LDM of speech dataset is accomplished using ANOVA. We have synthesized a common difference between normal voice pattern with emotional voice pattern. This proposed model retrieves the discriminative and similarity features from the samples of LDM of neutral and emotional speech datasets. The mean difference between these two LDMs is computed using ANOVA and the variance of result has been stored in another LDM which is treated as trained data set to recognize emotion state of a speaker, and it is a threshold values for recognizing emotion patterns on the speech data. Similarly, eight different emotion patterns and the corresponding regions on the LDM are retrieved and stored on eight different LDM. Each LDM is a knowledge base for the labelled emotion.

$$Speech\ Emotion\ State\ Region\ (SESR) = ss[n] * ss'[m] \tag{2}$$

Where *ss[n]* – is the *'n'* number of neutral speech vector stored in ss[LDM] and *ss'* – is the *'m'* labelled speech emotion vector stored in ss'[LDM].

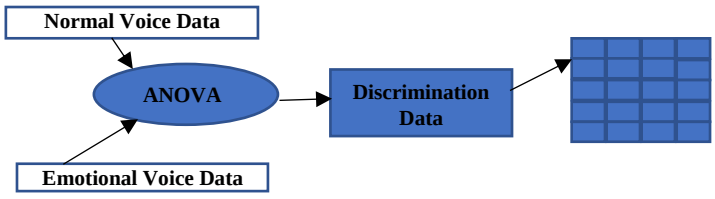


Fig. 4. Extracting sobbing pattern and stored in LDM

The mean value of the overall the vector samples can be calculated using the equation 3.

$$M = \frac{\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\right]^2}{N_{total}} Yij \tag{3}$$

Figure 3 shows the trained data set extraction. Numerically, one way ANOVA is a generalization of the two-sample '*t'* test. In this model, *'F'* statistics used to compare the variability between the two groups of LDM to identify the variability within the groups.

Degree of freedom $df = n - K$ and mean squares between $(MSB) = \frac{SSB}{dfb}$ (4)

The ANOVA 'F' can be calculated using the equation 5 given below

$$F = \frac{MST}{MSE} \tag{5}$$

The sum of squares between (SSB) the vectors can be calculated using the equation 6.

$$SSB = \sum_{j=1}^{k} (\bar{X}j - \bar{\bar{X}}j)^2 \tag{6}$$

$\bar{X}$ – is a mean of individual LDM and $\bar{\bar{X}}$ overall mean among LDM. The SSB along with SSW used to measure the significant difference among the mean values of several LDM of speech data. The Mean Sum of Squares (MST) attributed to treatment and the Mean Sum of Squares (MSE) attributed to error can be computed for the two distinct groups of LDM (Linear Discriminant Model) of speech vectors using the equations provided as 7 and 8, respectively.

$$MST = \frac{\sum_{i=1}^{k}(T_i^2/n_i) - G^2/n}{K-1} \tag{7}$$

$$MSE = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n} Y_{ij}^2 - \sum_{i=1}^{k}(T_i^2/n_i)}{n-k} \tag{8}$$

N_x – collections of *'A'* neutral(normal) speech vector blocks.

$$N = \sum_{i=1}^{n} A_i^2 \tag{9}$$

S_i – is a collections of *'A'* different labelled emotional speech vector. Both the data sets are stored in two different matrixes (LDM).

$$S = \sum_{i=1}^{n} B_i^2 \tag{10}$$

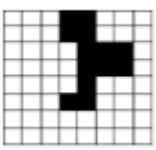
$$SVD\ (Emotional\ Voice\ Data) = F = \sum_{i=1}^{n} ANOVA(N,S) \tag{11}$$

The overall mean value of two vectors can be calculated using the equation 3 and MSB, MST can be calculated using the equation 5 and 6 respectively.

'N_i' and *'S_i'* substitutes on equation 5 to identify the variance between the two groups of vectors using ANOVA. Figure 5 shows the variance between the neutral and labelled emotion LDM vector. Each vector represents the corresponding labelled emotions and these vectors considered as trained set for real time speech emotion recognition.

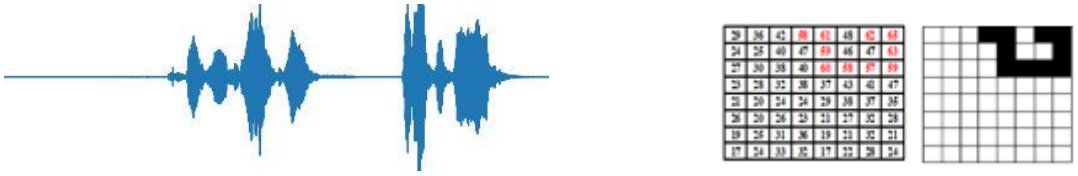
1. Crying

LDM Pattern



2. Angry

3. Fear



4. Sad

5. Happy

6. Surprise

7. Normal

8. Calm

Fig. 5(a). Eight Folded LDM with different emotion patterns

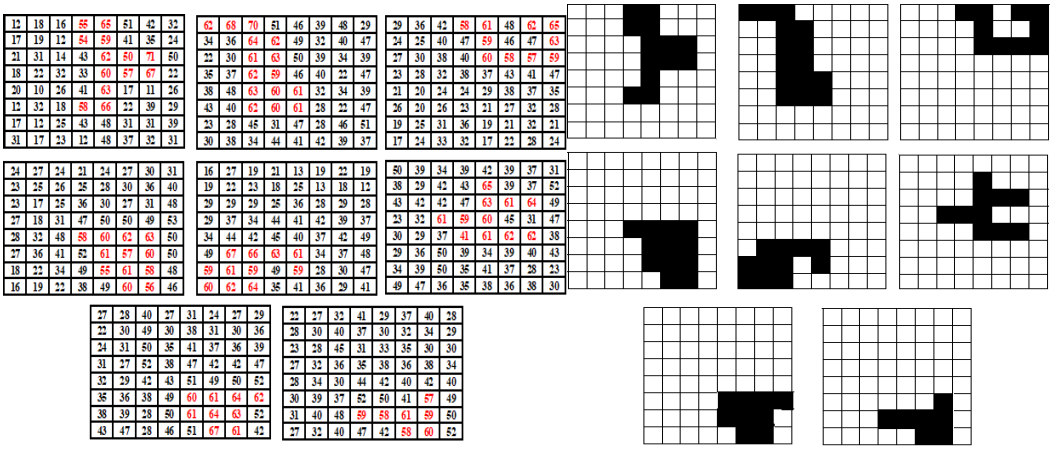


Fig. 5(b). Eight Folded LDM with different emotion patterns

The eight different emotion patterns stored in LDMs is a trained dataset for CNN. The similarity between the real time speech data and the trained set are compared using CSM in CNN as the activation function.

3.2. Cosine Similarity

Cosine similarity is a measure used to determine the similarity or dissimilarity between two vectors in a multi-dimensional space. Using cosine similarity measurement technique, the angle of similarity between the real time speech data blocks and the trained emotion data blocks which is stored in LDM is compared and measured similarity.

$$\boldsymbol{Sim}(\mathrm{N},\mathrm{S}) = \frac{N\ .\ S}{|N| * |S|} = \frac{\sum_{i=1}^{n} Ni\ X\ Si}{\sqrt{\sum_{i=1}^{n} N_i^2}\ \ X\ \sqrt{\sum_{i=1}^{n} S_i^2}} \tag{12}$$

Here, $|N|$ is a set of trained emotion vector, and $|S|$ vector is a real time speech data. The similarity angle or ratio between these two vectors shows that the speaker's emotions while speaking and wider angle is treated as dissimilar.

Each block of speaker data is compared with trained data set to detect emotion pattern in a pipelined manner. Here, initially, dot product has taken between the trained LDM with input LDM magnitude of both the vector calculated. At the end, the cosine similarity calculated by dividing dot product by the magnitude equation. The highest similarity LDM is delivered as final output from the CSM.

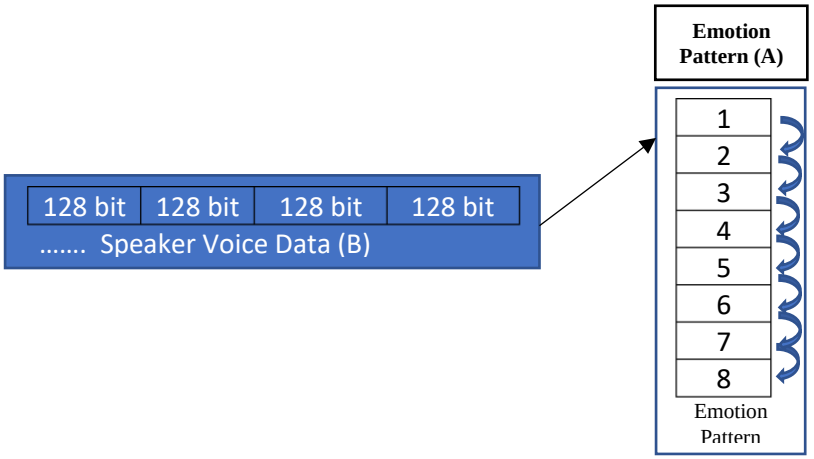


Fig. 6. Cosine Similarity Measurement for Emotion Pattern Detection

3.3. CNN Model for Emotion Detection

Convolutional Neural Networks (CNNs) are a type of deep learning model that has demonstrated significant success across various tasks, such as image recognition, natural language processing, and audio processing. In the context of speech emotion recognition, CNNs are utilized to extract features and classify emotional states from speech data.

We have collected different speech related datasets such as MINDSET, RAVDESS, SAVEE, and TESS along with thirty our own datasets. These datasets are grouped into two broad categories. First data set is neutral/normal speech dataset, and the second one is labeled dataset of speech recordings along with corresponding emotion labels. The datasets employed for training, validation, and testing the CNN model are diverse, encompassing a broad spectrum of emotions, speech styles, and speakers. This diversity ensures that the model can effectively generalize across various scenarios and conditions. The statistical method ANOVA have used for feature extraction from voice data stored on the LDM to process speech signals using CNN.The proposed CNN architecture comprises a single convolutional layer succeeded by pooling layers, aimed at capturing local patterns within the features. Following this, fully connected layers are utilized for recognizing global patterns. The fully connected layer comprises eight neurons, each corresponding to different emotions, facilitating emotion recognition. An activation function, such as CSM, is applied to classify the emotions, resulting in the final output.

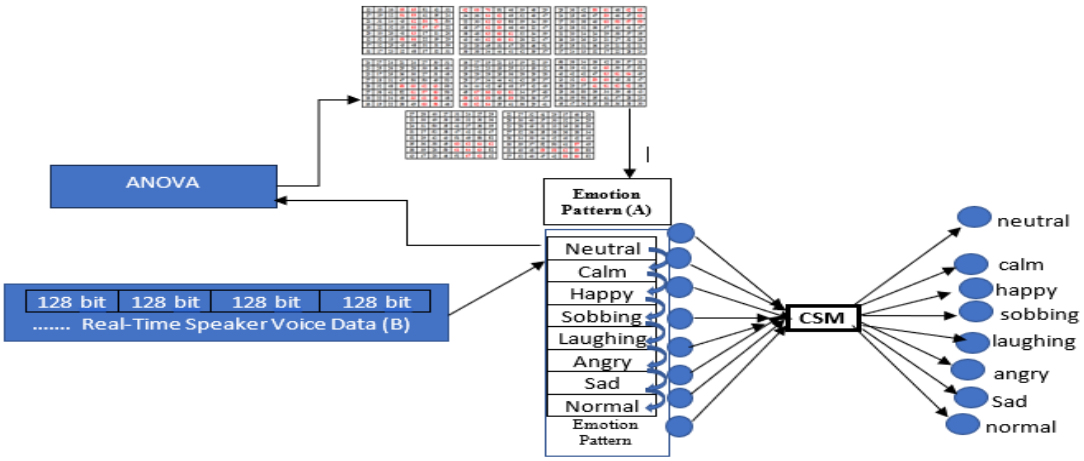


Fig. 7. Kernel function of modified CNN – Direct method for SER

The CNN model is trained using a labeled dataset, wherein input features with ANOVA are inputted into the CNN. Through this process, the model learns to associate these features with the respective emotion labels, thereby mapping input features to their corresponding emotions. During training, the model is optimized using an appropriate loss (error) function, such as cross-entropy between neutral speech input matrix and dataset contains emotions, and the weights are updated using gradient descent.

In this context, the input LDM is denoted by 'f' and 'h', while the emotional pattern is represented by 'm' and 'n'. The resulting matrix is designated as G[m,n].

$$\boldsymbol{G[m,n] = (f * h)(m,n) = \sum_{j} \sum_{k} j[j,k]\, f[m-j,\ n-k]} \qquad (\mathbf{13})$$

Once the emotion patterns are positioned, each value from the pattern is multiplied pairwise with the corresponding values from the input voice dataset. The dimension of the output matrix, considering padding and stride, can be calculated using the provided equation.

$$\boldsymbol{n_{out} = \left[\frac{n_{in} + 2p - f}{S} + 1\right]} \qquad (\mathbf{14})$$

The resulting matrix for emotion patterns is computed using Equation 15. This process involves classifying the speech data and comparing it with the labeled emotion patterns stored in the LDM (Linear Discriminant Model) to finalize the emotion detections.

$$dA+= \sum_{m=0}^{n_h} \sum_{n=0}^{n_w} W.dZ[\,m,n\,] \tag{15}$$

The developed CNN model is designed to classify the emotional state of a speaker, distinguishing between whether the speaker is speaking normally or emotionally.

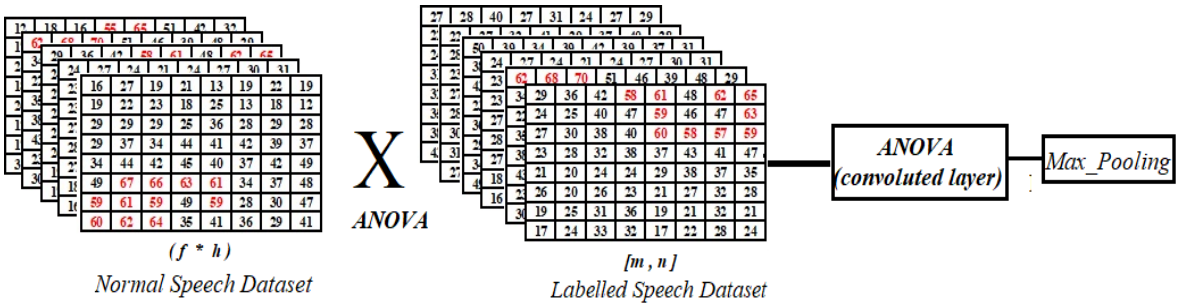


Fig. 7. ANOVA-Convoluted Layer

The output matrix confirms that the proposed CNN model fulfills the design objectives by accurately classifying the emotional state of a speaker. The Cosine Similarity Measure (CSM) is employed to scrutinize the fully connected layer, assessing the similarity at the global maxima to determine the final emotion state. By selecting the highest similarity, the CSM identifies the emotion state of a speaker.

4. Results and Discussions

To experiment the proposed model, we have taken IEMOCAP as major dataset along with three categories of sobbing emotion samples, and 10 labelled datasets were handled for each category with python code. *'statsmodels'* python library along with NumPy, SciPy, and other required modules added.

Table 1. Male/Female Speaker Distribution

Dataset	Total Recordings	Male Recordings	Female Recordings
RAVDESS	480	480	0
SAVEE	1920	960	960
TES	2800	0	2800
Total	5200	1440	3760

Table 1 displays the datasets collected, categorized by male and female recordings, while Table 2 presents the performance accuracy of various models across different datasets [10]. The proposed approach, integrating ANOVA with CSM, effectively reduces the number of intermediate layers between the input and output of the CNN model.

Table 2. Accuracy Comparison with state-of-art studies

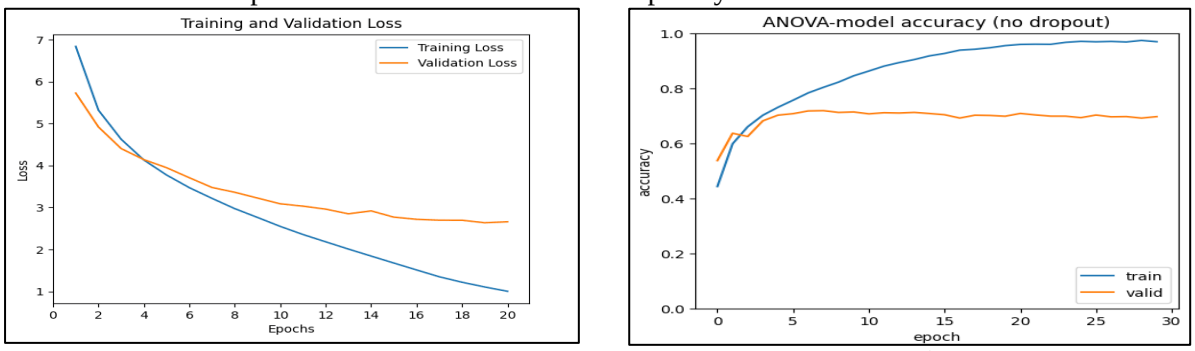
Sl.No	Dataset	CNN %	MFCC %	CNN+Attention %	ANOVA + CSM +CNN (Proposed Model)%
1	IEMOCAP	76.36	77.30	76.18	79.57
2	MINDSET	79.12	67.21	72.67	79.32
3	RAVDESS	68.21	53.08	77.8	81.28
4	SAVEE	70.89	72.66%	89.23	88.14
6	TESS	67.02	49.48	87	88.31

The degree of variance among the labelled emotion shows in Figure 3. The confusion matrix measured using CMS between the labelled, trained LDM and neutral LDM over the IEMOCAP dataset.

Table 3. Confusion matrix for the IEMOCAP using CSM

Calm	52.34%	46.19%	12.32%	28.26%	47.57%	31.35%	49.19%	96.07%
Neutral	42.00%	33.18%	53.37%	21.25%	61.78%	33.17	89.62%	49.19%
Fear	61.31%	23.12%	27.25%	37.42%	18.21%	83.01%	33.17	31.35%
Happy	32.06%	18.34%	68.78%	12.15%	89.64%	18.21%	61.78%	47.57%
Sad	74.82%	42.28%	20.47%	92.47%	12.15%	37.42%	21.25%	28.26%
Surprise	23.12%	37.54%	90.95	20.47%	68.78%	27.25%	53.37%	12.32%
Angry	55.63%	91.14%	37.54%	42.28%	18.34%	23.12%	33.18%	46.19%
Sobbing	92.13%	55.63%	23.12%	67.32%	32.06%	69.31%	42.00%	23.44%
	Sobbing	**Angry**	**Surprise**	**Sad**	**Happy**	**Fear**	**Neutral**	**Calm**

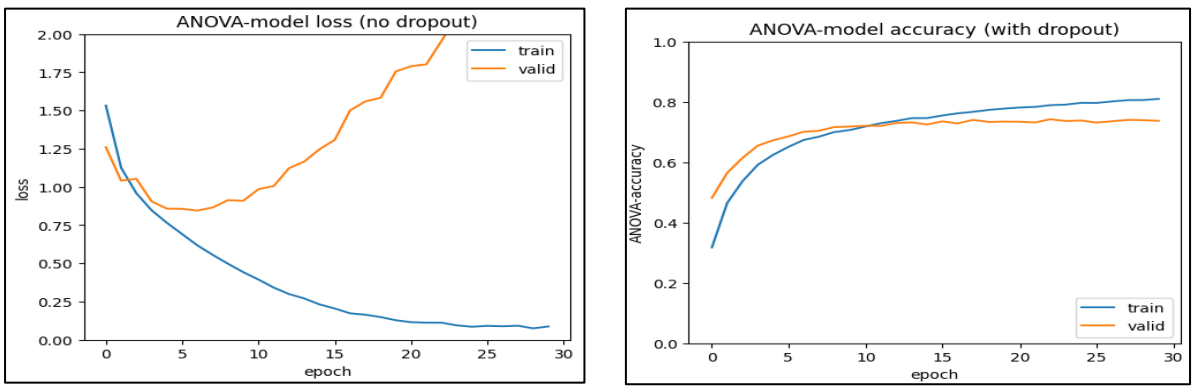
The cross-tabulated matrix ensures similarity among various emotion patterns, facilitating the classification of the speaker's emotional state at the output layer in the CNN.



(a) Validation Loss **(b)** Validation of ANOVA

Fig. 8. Epoch graph for training and validation of ANOVA-Convolution Layer

Fig. 8(a) overall deviation between the CNN-ANOVA training and the validation on the IEMOCAP dataset, and the Fig. 8(b) show the accuracy of predicting and segregating the features from the given two LDM.



(a) Training & Validation with dropout (b) Training & Validation without dropout

Fig. 9. Loss of CNN-ANOVA model with and without dropout

In CNN, the CSM classify the labelled emotions before the output layer during global pattern recognition by comparing input LDM with eight different labelled motion LDM. Among the eight LDM who mean difference is less than 30 is dropout due to its contribution is not impact on the feature selection. Figure 9(a) displays the loss and accuracy of the CNN-ANOVA model with dropout, while Figure 9(b) illustrates the same for the dropout model.

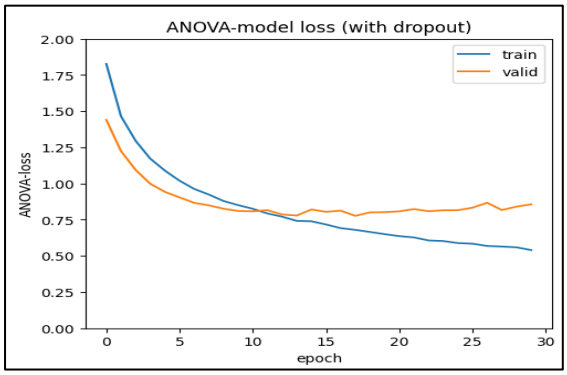


Fig. 10. CSM loss on classfication

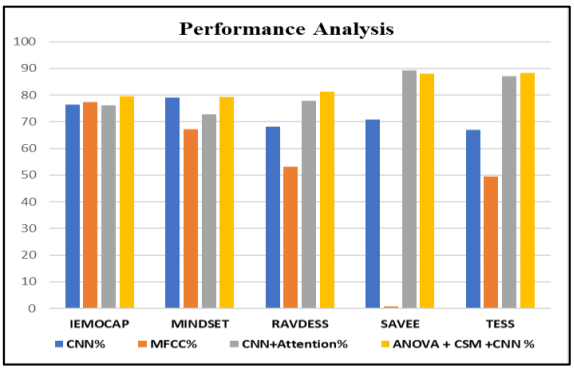


Fig. 11. Overall perfomance of the the ANOCA+CSM+CNN

Table 2 shows the accuracy various models for recognizing the speech emotion on the standard dataset. Each model has its own pros and cons to recognize the speech emotions over the different dataset. MFCC shows the better accuracy results on IEMOCAP compare to CNN and CNN with attention. Eventually, the performance of CNN model progressing on the MINDSET corresponding to the MFCC and CNN + Attention. Alternatively, CNN+Attention's accuracy is better compared with other models tabulated in table 2.

In CNN model, the loss can be reduced using the dropout. Fig. 10 shows the variation between the training and testing on CSM part in the CNN. The accuracy enhanced with 79.57 with dropout and achieved overall accuracy 88.31 on TESS dataset. Figure 11 illustrates the overall performance of the proposed system compared to state-of-the-art existing models, demonstrating better accuracy with a single convoluted layer. Figure 12 shows the misclassified instances.

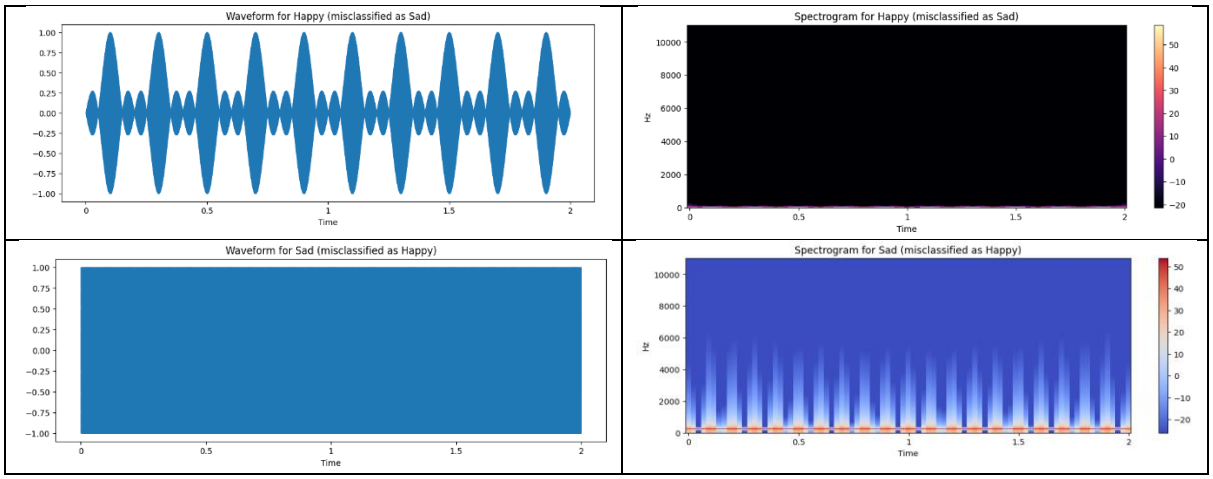


Figure 12: Misclassified Instances

The misclassification of emotional audio signals has occurred due to overlapping acoustic features between different emotions. For instance, if a 'Happy' sound is slowed down or has its pitch lowered, it might resemble a 'Sad' sound, leading to misclassification. Similarly, 'Sobbing' and 'Angry' sounds might both have high-intensity and erratic patterns, causing confusion between the two. In the case of

'Surprise' and 'Fear,' both might share sudden onset and high-pitched elements, making them difficult to distinguish. Such misclassifications highlight the importance of distinct and nuanced features in accurately identifying emotions from audio signals.

The misclassification of emotional audio signals has occured due to overlapping acoustic features between different emotions. For instance, if a 'Happy' sound is slowed down or has its pitch lowered, it might resemble a 'Sad' sound, leading to misclassification. Similarly, 'Sobbing' and 'Angry' sounds might both have high-intensity and erratic patterns, causing confusion between the two. In the case of 'Surprise' and 'Fear,' both might share sudden onset and high-pitched elements, making them difficult to distinguish. Such misclassifications highlight the importance of distinct and nuanced features in accurately identifying emotions from audio signals.

5. Conclusion and Future Work

The proposed approach presents a promising direction for speech emotion recognition using CNN, with potential applications in healthcare, human-computer interaction, and emotional well-being assessment. In this model, speech emotion recognition using ANOVA for CNN and CSM shows promise in detecting emotions while speaking, which can have significant applications in the healthcare sector. With a well-curated dataset and the implementation of advanced techniques, such as ANOVA for threshold determination, the proposed approach aims to accurately classify speech signals into emotional or normal speech categories. The ANOVA and CSM directly reduces the number of convolutional steps in CNN to decrease response time and improve performance. The experimentation results demonstrated that the proposed approach performs effectively in identifying the emotional state of a speaker, showcasing the potential of CNN-based models in speech emotion recognition tasks. Continued research and development in this field hold the potential to advance speech-based emotion recognition technology further. Such advancements could pave the way for integrating this technology into remote voice emotion assessment for practical healthcare applications.

Compliance with Ethical Standards

Disclosure of potential conflicts of interest: We declare that, no conflict of interest among all authors.

Research involving Human Participants and/or Animals: No animals or Human involved in this research.

Informed consent: Not applicable

Data Availability: The data used for the findings will be shared by the corresponding author upon request.

References:

[1] Yi-Hsuan Yang, Yu-Ching Lin, Ya-Fan Su, and Homer H. Chen, “A Regression Approach to Music Emotion Recognition”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 16, No. 2, February 2008. Pp. 448-457. Digital Object Identifier 10.1109/TASL.2007.911513

[2] Mixiao Hou, Jinxing Li, and Guangming Lu, “A supervised non-negative matrix factorization model for speech emotion recognition”, Speech Communication, 124 (2020) 13–20.

[3] Abdel-rahman Mohamed, George E. Dahl, and Geoffrey Hinton, “Acoustic Modeling Using Deep Belief Networks”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No. 1, pp. 14-22, Jan 2012.

[4] A. Milton, S. Tamil Selvi, “Class-specific multiple classifiers scheme to recognize emotions from speech signals”, Computer Speech and Language Vol. 28, pp. 727–742, 2014.

[5] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, “Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition”, IEEE Transaction on Audio, Speech, and Language Processing, Vol. 20, No. 1, Jan 2012.

[6] Xiao-Lei Zhang and JiWu, “Deep Belief Networks Based Voice Activity Detection”, Ieee Transactions on Audio, Speech, and Language Processing, Vol. 21, No. 4, pp. 697-710, April 2013.

[7] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo, “Environmental Sound Recognition with Time–Frequency Audio Features”, Ieee Transactions on Audio, Speech, and Language Processing, Vol. 17. No. 6, pp. 1142-1158, Aug. 2009.

[8] Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton, “Speech Recognition with Deep Recurrent Neural Networks”, ICASSP 2013.

[9] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jain, Mohammad Haseeb Zafar, and Thamer Alhussain, “Speech Emotion Recognition Using Deep Learning Techniques: A Review”, Vol. 7, 2019.

[10] Zengwei Yao, Zihao Wang, Weihuang Liu, Yaqian Liu, Jiahui Pan, “Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN”, Speech Communication, Vol. 120, pp-11-19, 2020.

[11] Basheer, Shakila, M. Anbarasi, Darpan Garg Sakshi, and V. Vinoth Kumar. "Efficient text summarization method for blind people using text mining techniques." International Journal of Speech Technology 23 (2020): 713-725.

[12] Yusuke Yasuda, Xin Wang, and Junichi Yamagishi, “Investigation of learning abilities on linguistic features in sequence-to-sequence text-to-speech synthesis”, Computer Speech & Language 67 (2021) 101183.

[13] Mixiao Hou, Jinxing Li, and Guangming Lu, “A supervised non-negative matrix factorization model for speech emotion recognition”, Speech Communication, 124, pp-13-20, 2020.

[14] Qirong Mao, MingDong, Zhengwei Huang, and Yongzhao Zhan, “Learning Salient Features for Speech Emotion Recognition Using Convolutional Neural Networks”, IEEE Transactions on Multimedia, Vol. 16, No. 8, December 2014

[15] Mingyang Zhang, Berrak Sisman, Li Zhao, and Haizhou Li, “DeepConversion: Voice conversion with limited parallel training data”, Speech Communication 122 (2020) 31–43.

[16] Norihide Kitaoka, Daisuke Enami, and Seiichi Nakagawa, “Effect of acoustic and linguistic contexts on human and machine speech recognition”, Computer Speech and Language, Vol. 28, pp.769-787, 2014.

[17] Li Deng and Xiao Li, “Machine Learning Paradigms for Speech Recognition: An Overview”, IEEE Transactions of Audio, Speech, and Language Processing, Vol. 21, No. 5, May 2013.

[18] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-End Factor Analysis for Speaker Verification”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 4, pp. 788-798, May 2011.

[19] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, “VoxCeleb: Large-scale Speaker Verification in the Wild”, Computer Speech & Language, 2019.

[20] Yuvika Singh, Anban Pillay, and Edgar Jembere, “Features of Speech Audio for Accent Recognition”, International Conference on Advances in Big Data, Computing and Data Communication Systems (icABCD), Durban, South Africa, 2020.

[21] Ashokkumar Palanivinayagam and Sureshkumar Nagarajan, “An optimized iterative clustering framework for recognizing speech”, International Journal of Speech Technology, July 2020.

[22] Sue E. Tranter and Douglas A. Reynolds, “An Overview of Automatic Speaker Diarization Systems”, Ieee Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 5, Sep. 2006.

[23] Sait Melih Dogan and Ozgul Salor, “Music/singing voice separation based on repeating pattern extraction technique and robust principal component analysis”, ICEEE, 3-5 May 2018.

[24] Xiao-Lei Zhang and Ji Wu, “Deep Belief Networks Based Voice Activity Detection”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 21 Iss. 4, April 2013.

[25] Umamaheswaran, S., Ravi Lakshmanan, V. Vinothkumar, K. S. Arvind, and S. Nagarajan. "New and robust composite micro structure descriptor (CMSD) for CBIR." International Journal of Speech Technology 23 (2020): 243-249.

[26] Rashid Jahangir, Ying Wah The, Nisar Ahmed Memon, Ghulam Mujtaba, Mahdi Zareei, Uzair Ishtiaq, Muhammad Zaheer Akhtar, and Ihsan Ali, “Text-Independent Speaker Identification Through Feature Fusion and Deep Neural Netowrk”, IEEE Access, Vol. 8, 12 Feb. 2020: 10.1109/ACCESS.2020.2973541.

[27] Ajinkya N. Jadhav, Nagaraj V. Dharwadkar, " A Speaker Recognition System Using Gaussian Mixture Model, EM Algorithm and K-Means Clustering", International Journal of Modern Education and Computer Science (IJMECS), Vol.10, No.11, pp. 19-28, 2018.DOI: 10.5815/ijmecs.2018.11

[28] Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, “Speech Emotion Recognition Using Fourier Parameters”, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, VOL. 6, NO. 1, JANUARY-MARCH 2015.

[29] Javier G. R´azuri, David Sundgren, Rahim Rahmani, Aron Larsson, Antonio Moran Cardenas, and Isis Bonet, “Speech emotion recognition in emotional feedback for Human-Robot Interaction”, (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 4, No.2, 2015.

[30] Ravi Kumar Kandagatla and Venkata Subbaiah Potluri, “Performance analysis of neural network, NMF and statistical approaches for speech enhancement”, International Journal of Speech Technology, 2 September 2020. doi.org/10.1007/s10772-020-09751-6.

[31] Praveen Sundar, P. V., D. Ranjith, T. Karthikeyan, V. Vinoth Kumar, and Balajee Jeyakumar. "Low power area efficient adaptive FIR filter for hearing aids using distributed arithmetic architecture." International Journal of Speech Technology 23, no. 2 (2020): 287-296.

[32] Petr Cerva, Lukas Mateju, Jindrich Zdansky, Radek Safarik, Jan Nouza, “Identification of related languages from spoken data: Moving from off-line to on-line scenario”, Computer Speech & language 68, 2021. 101184.

[33] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim ,Jeannette N. Chang , Sungbok Lee, and Shrikanth S. Narayanan, “IEMOCAP: interactive

emotional dyadic motion capture database", Lang Resources & Evaluation (2008) 42:335–359. DOI 10.1007/s10579-008-9076-6.

[34] G. A. Prabhakar, B. Basel, A. Dutta and C. V. R. Rao, "Multichannel CNN-BLSTM Architecture for Speech Emotion Recognition System by Fusion of Magnitude and Phase Spectral Features using DCCA for Consumer Applications," in IEEE Transactions on Consumer Electronics, doi: 10.1109/TCE.2023.3236972.

[35] S. Nakagawa, L. Wang and S. Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 4, pp. 1085-1095, May 2012, doi: 10.1109/TASL.2011.2172422.

[36] Jageet Singh, Lakshmi Babu Saheer, and Oliver Faust, "Speech Emotion Recognition Using Attention Model", International Journal of Environmental Research and Public Health, Vol. 20, 2023.doi: 10.3390/ijerph200665140

[37] Shahzadi, A., Ahmadyfard, A., Yaghmaie, K., & Harimi, A. (2013). Recognition Of Emotion In Speech Using Spectral Patterns. *Malaysian Journal of Computer Science*, *26*(2), 140–158. Retrieved from https://ejournal.um.edu.my/index.php/MJCS/article/view/6767.

[38] SUSMITHA VEKKOT, DEEPA GUPTA, MOHAMMED ZAKARIAH, AND YOUSEF AJAMI ALOTAIB, "Emotional Voice Conversion Using a Hybrid Framework with Speaker-Adaptive DNN and Particle-Swarm-Optimized Neural Network", IEEE Access, May 4, 2020. DOI:10.1109/ACCESS.2020.2988781

[39] Al Moteri, M., Khan, S. B., & Alojail, M. (2023). Economic growth forecast model urban supply chain logistics distribution path decision using an improved genetic algorithm. Malaysian Journal of Computer Science, 76-89.

[40] A Agrawal et al., "Comparative Analysis of Speech Emotion Recognition Models and Technique," 2023 International Conference on Computational Intelligence, Communication Technology and Networking (CICTN), Ghaziabad, India, 2023, pp. 499-505, doi: 1109/CICTN57981.2023.10141044.

[41] Huan Zhao, Nianxin Huang & Haijiao Chen (2024) Knowledge enhancement for speech emotion recognition via multi-level acoustic feature, Connection Science, 36:1, DOI: 10.1080/09540091.2024.2312103

[42] Pentari, A., Kafentzis, G. & Tsiknakis, M. Speech emotion recognition via graph-based representations. Sci Rep 14, 4484 (2024). https://doi.org/10.1038/s41598-024-52989-2

[43] Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, Sandra L. Schneider,Speech emotion recognition using machine learning — A systematic review,Intelligent Systems with Applications, Volume 20, 2023. https://doi.org/10.1016/j.iswa.2023.200266.

[44] Mannar Mannan, J., Srinivasan, L., Maithili, K., & Ramya, C. (2023). Human Emotion Recognize Using Convolutional Neural Network (CNN) and Mel Frequency Cepstral Coefficient (MFCC). *Seybold Report Journal*, *18*(4), 49-61.