Data and Metadata. 2025; 4:715
doi: 10.56294/dm2025715

ORIGINAL

Enhanced Speech Emotion Recognition Using Audio Signal Processing with CNN Assistance

Reconocimiento mejorado de las emociones del habla mediante el procesamiento de señales de audio con ayuda de CNN

Chandupatla Deepika[1,2], Swarna Kuchibhotla[1]

[1]Koneru Lakshmaiah Education Foundation (Deemed to be University. Vaddeswaram, Guntur, Ap, India.
[2]Department of CSE, Vidya Jyothi Institute of Technology (VJIT). Hyderabad, India.

Cite as: Deepika C, Kuchibhotla S. Enhanced Speech Emotion RecognitionUsing AudioSignal Processing with CNN Assistance. Data and Metadata. 2025; 4:715. https://doi.org/10.56294/dm2025715

Submitted: 02-07-2024 **Revised:** 10-11-2024 **Accepted:** 12-04-2025 **Published:** 13-04-2025

Editor: Dr. Adrián Alejandro Vitón Castillo

Corresponding Author: Chandupatla Deepika

ABSTRACT

The important form human communicating is speech, which can also be used as a potential means of human-computer interaction (HCI) with the use of a microphone sensor. An emerging field of HCI research uses these sensors to detect quantifiable emotions from speech signals. This study has implications for human-reboot interaction, the experience of virtual reality, actions assessment, Health services, and Customer service centres for emergencies, among other areas, to ascertain the speaker's emotional state as shown by their speech. We present significant contributions for; in this work. (i) improving Speech Emotion Recognition (SER) accuracy in comparison in the most advanced; and (ii) lowering computationally complicated nature of the model SER that is being given. We present a plain nets strategy convolutional neural network (CNN) architecture with artificial intelligence support to train prominent and distinguishing characteristics from speech signal spectrograms were improved in previous rounds to get better performance. Rather than using a pooling layer, convolutional layers are used to learn local hidden patterns, whereas Layers with complete connectivity are utilized to understand global discriminative features and Speech emotion categorization is done using a soft-max classifier. The suggested method reduces the size of the model by 34,5 MB while improving the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) and Interactive Emotional Dyadic Motion Capture (IEMDMC) datasets, respectively, increasing accuracy by 4,5 % and 7,85 %. It shows how the proposed SER technique can be applied in real-world scenarios and proves its applicability and efficacy.

Keywords: Neural Networks (NN); Artificial Intelligence (AI); Emotion Recognition (ER); Noise Reduction (NR).

RESUMEN

La forma más importante de comunicación humana es el habla, que también puede utilizarse como medio potencial de interacción persona-ordenador (HCI) con el uso de un sensor de micrófono. Un campo emergente de la investigación en HCI utiliza estos sensores para detectar emociones cuantificables a partir de señales del habla. Este estudio tiene implicaciones para la interacción persona-ordenador, la experiencia de la realidad virtual, la evaluación de acciones, los servicios sanitarios y los centros de atención al cliente para emergencias, entre otros ámbitos, para conocer el estado emocional del hablante a partir de su habla. En este trabajo presentamos contribuciones significativas para (i) mejorar la precisión del Reconocimiento de Emociones del Habla (SER) en comparación en los más avanzados; y (ii) disminuir la naturaleza computacionalmente complicada del modelo SER que se está dando. Presentamos una arquitectura de red neuronal convolucional

(CNN) de estrategia simple con apoyo de inteligencia artificial para entrenar características prominentes y distintivas de los espectrogramas de la señal de voz se mejoraron en rondas anteriores para obtener un mejor rendimiento. En lugar de utilizar una capa de agrupación, se emplean capas convolucionales para aprender patrones ocultos locales, mientras que las capas con conectividad completa se utilizan para comprender características discriminativas globales y la categorización de la emoción del habla se realiza mediante un clasificador soft-max. El método sugerido reduce el tamaño del modelo en 34,5 MB al tiempo que mejora los conjuntos de datos Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) e Interactive Emotional Dyadic Motion Capture (IEMDMC), respectivamente, aumentando la precisión en un 4,5 % y un 7,85 %. Se muestra cómo la técnica SER propuesta puede aplicarse en escenarios del mundo real y se demuestra su aplicabilidad y eficacia.

Palabras clave: Redes Neuronales (NN); Inteligencia Artificial (IA); Reconocimiento de Emociones (RE); Reducción de Ruido (RN).

INTRODUCTION

The Speech Emotion Recognition (SER) is most efficient natural and biological means about communication in addition exchange in between people and computers. It is crucial to real-time applications of HMI. Since speech signals include further information than uttered words, researchers are actively studying how to use them to identify speakers' qualitative emotional states utilizing speech signals produced by sensors for SER.(1) A lot of Researchers are engaged in this field to create the machine smart enough to decipher a speaker's emotional state from their voice and assess or identify it. Selecting and extracting the prominent and discriminative features in SER is a difficult undertaking.(2) To retrieve concealed data and train different Consolation Nural Network CNN models to increase performance and reduce the SER's computational complexity for human behaviour evaluation, researchers are currently attempting to identify the reliable and most important characteristics for SER by utilizing artificial intelligence and deep learning methods.(3,4,5) Because of the large quantity of social media users, inexpensive price, and the quick Internet speed in this period of research, SER have encountered numerous difficulties and limitations. Semantic gaps arise because of the usage of social media and the inexpensive internet. Researchers and Scholars have attempted to cover and present new ways to take out to most noticeable aspects from trained models and voice signals to properly understand an emotion of the speaker when speaking in order in order to bridge the conceptual divide in this area, the technology is always evolving to give academics new, adaptable platforms on which to introduce novel applications of artificial intelligence.

The improvement of Human Computer Interaction (HCI), including acknowledgment of an emotions of sense depends heavily on technological advancements, skill development, and the application of artificial intelligence as well as deep learning methods. SER is an effective field in Human Computer Interaction (HCI) with numerous applications in real time. For example, it is usable in make a call centres to measure customer satisfaction; in human-reboot interactions to measure Human feelings; in emergencies contact centres to determine a user's emotional state so that a suitable response can be given; and in virtual reality. Fiore et al.(6) SER were implemented for the car panel system to identify drivers' emotional or emotional states to properly action could be taken to protect the passengers. Additionally, SER contributes to automatic translation systems and the knowledge of crowd dynamics that could lead to aggressive or destructive behaviour that is difficult to accomplish manually.(7) To illustrate the efficacy of the SER healthcare care centres, Badshah et al.(8) used the quantified SER services and resources to explain approaches using with CNN method structures that incorporate rectangle-shaped filters. To ensure raise the importance of HCI, Mao et al.(9) enhanced the efficiency of SER for instantaneous application analysis and feature extraction. To estimate emotion intensities multi type as metatype in a hierarchical manner, Min et al.(10) applied the SER for explaining emotion in movies utilizing content analysis of arousal and violence discriminative elements. Miguel et al.(11) used SER for concerns of privacy reasons to identify the speaker using paralinguistic indications and a privacy-preserving hashing technique.

Numerous researchers have offered a range of approaches in the rapidly developing field of SER research. To accurately identify a speaker's emotion, most researchers are trying to determine which ones are effective, prominent and biased characteristics speech signals for categorization. Deep learning methodologies possess recently been used by researchers to identify the key and discriminative characteristics for SER. Convolutional Neural Networks (CNN) are employed to recognize and comprehend lines., dots, curves, and forms by constructing high-level features on top of low-level data. To obtain greater accuracy than low-level handmade features, algorithms for Deep Learning like CNN, LSTM, DNN, DBN, and others aim at recognizing the conspicuous features at highest level. The model gets more computationally complex when deep neural networks are used.

1. The voice recognition industry has many challenges. One of them is the absence of considerable improvement in precision and cost complexity of current CNN designs in speech signal processing.
2. While Long Short-Term Memory (LSTM) and RNN able to be accustomed to sequential training information, we are harder to work out efficiently and have increased computational aspects complexity.
3. Most of the research have employed concatenation and Encoding at the frame level techniques regarding feature fusion, we are not appropriate for SER at the utterance level.
4. Lack of data makes it difficult to determine the precise word border and causes a huge concatenation feature fusion.

Owing to the problems and difficulties, we suggested a unique CNN architecture with unique steps in place a pooling strategy to extract vital higher-level characteristics after speech spectrograms of data. We recognize vocal signals. hidden patterns within convolutional layers by employing specific steps to bring down sample, the maps with attributes, in order to determine, The significance and efficiency of suggested model relative to other cutting-edge techniques, comprehensive experiments were carried out on two metrics that were measured Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[13] and Interactive Emotional Dyadic Motion Capture (IEMDMC)[12] datasets. The test component of this study includes the comprehensive tests and examination of the suggested approach in comparison with additional baseline techniques.

The key contributions of this article are reframed are outlined as follows:

- Pre-Processing: to guarantee a precise forecast of emotional outcomes, SER feeds in data that has been refined, which constantly requires detailed speech signals. Preprocessing stages, which efficiently improve the data and contribute to improving the final classifier's accuracy, are not given enough attention in the SER literature's methodology. Our preprocessing approach in this paper involves first removing silent sections from audio data using the Novell adaptive thresholding technique. Hence, a major part of the entire SER system is our preprocessing approach.
- CNN (Convolutional Neural Network) Model: we used convolutional neural network technique[14] and suggested a novel CNN architecture, the DSCNN for SER, to teach the convolutional layers salient and discriminative features. Rather than using pooling layers, the convolutional layer's unique strides are used for down sampling the feature maps. Spectrophotograms are specifically used in the DSCNN model to solve the SER problem.
- Computational Complexity: CNN architectures that we have suggested have modest respective fields. employ minimum convolutional layers to obtain discriminative, crucial, and deep characteristics from speech spectrograms. This reduces computational complexity and increases accuracy because the proposed CNN model has a simple structure, as demonstrated by the experiments.

Related work

Investigating digital signal processing (DSP) is broad. Researchers have developed some effective methods for SER currently that use digital audio signals from the voice to ascertain the affective condition regarding the speaker. In this domain, robust feature selection which accurately detects a speaker's emotions is a difficult assignment.[15] The feature selection procedure, which extracts high-level characteristics derived from audio data, and the classifier procedure for selecting, which chooses which classifiers to use to accurately identify emotions from speech, make up a typical SER. While some studies have trained CNN, DNN, and NN models using low-level hand-crafted features to raise the accuracy of SER, several researchers have recently revealed methods for SER employing deep learning approaches for enhancing accuracy of recognition using audio speech data.

Feature-Based Speech Emotion Recognition (SER) Produced by Hand

For researchers, choosing the strong features for SERI is a difficult task. SER's handcrafted features have been utilized by a few researchers. So, instead of using supplementary low-level characteristics like volume, linear productivity code (LPC),[16] formant, etc. Dave et al.[17] analysed several characteristics of speech emotions and shown frequency intended for optimal Mel frequency cepstral coefficient (MFCC)[18] characteristics of SER. By utilizing extra speech features including jitter and shimmer, Liu and colleagues[19] calculated the cepstral coefficient of gammatone frequency. The unweighted accuracy of SER can be increased by up to 3,6 % compared to MFCCs using (GFCC) characteristics. For classification, Liu et al. employed a decision tree based on an extreme learning machine (ELM).[20] suggested SER technique, which employed a Chinese voice dataset and applied correlation to find hidden emotional features (CASIA).[21] The technique in the direction of choose characteristics for DNN based on models for SER training based on glottal and MFCCs was reported by Fahad et al.[22] Wei et al.[2] used the autoencoder and sparse classifier suggested the SER approach to produced positive outcomes on a dataset of Chinese speaking emotions. To identify all type of the emotions in voice signals, the large-scale hidden properties are retrieved via the autoencoder, and sparse network is utilized to trace the attributes with tiny dimensions to a support vector machine (SVM) classification system.

Convolutional Neural Network (CNN)-Based on Speech Emotion Recognition (SER)

For SER, researchers have recently utilized CNN features. Zhang et al.(23) created the DCNN-based approach for the SER by training an SVM, a conventional classifier, to identify the speaker's emotional state and utilizing CNN's pre-trained Alex Net model to discover deep learning features. George et al.(24) provided a method for SERs that are spontaneous and grounded in CNN and LSM utilizing the Remote Collaboration feature with the Natural Emotion Database Active RECOLA. The writer trained a CNN model to extract the speech's overall discriminatory element, which was input to an LSTM for sequence learning to ascertain the emotions of the speaker. The random deep belief network was utilized for SER by Wen et al.(25). Using deep belief networks (DBNs) high-level characteristics that discriminate are retrieved from speech signals by feeding the low-level features, which were first extracted using LLD. The SVM classifier, which is linked to each DB, receives these high-level features. It then predicts the speaker's weaker sensations and makes choices. by the vast majority vote. Lian et al.(26) employed a complex model in which an SVM classifier was used to predict emotions and a DBN was used to learn hidden features from speech to use the Chinese CASIA dataset to attain high-level accuracy in SER. According to Hajar et al.(27) a technique for SER that divided the discourse inputs into frames as well also retrieved the characteristics of the MFCCs as transformed them into spectrograms, which serve as a representation of the keyframe when chosen as a full audio the act of speaking. The 3D CNN model is trained to forecast the emotions in speech using the key spectrograms that are selected using the K-mean clustering approach. A method for SER employing advance long short-term memory (A-LSTM) to become familiar with the sequences using a pooling recurrent neural network (RNN) scheme was described by Fei et al.(28) This method outperforms simple LSTM. Saurabh et al.(29) assessed the autoencoder's performance in SER using state-of-the-art GAN.(30) after employing it for voice recognition on the IEMDMC dataset. Numerous techniques have been reported in the literature to extract discriminatory features from voice signals using the CNN model for SER with several kinds of input.(9) Deep learning techniques for SER were used in (26,27,28,29,30,31) to raise the ratio of recognition for real-time spontaneous SER Making use of a range of speech datasets, including IEMDMC, SAVEE, RAVDESS, CAISE TITMIT, and others. The use of massive pre-trained CNN architectures increases accuracy but significantly increases the model's cost computations, using spectrograms as an input, several researchers have built techniques to identify the emotions expressed in speech. Utilizing CNN models to extract features, and separate classifiers for classification will increase the computational difficulty of the entire model.(32) Using a simple network, For SER, we presented a brand-new CNN architecture. in this research. To down sample the input feature map, strides were employed in the pooling layer in place of the convolutional layers. It lowers the total computational expense. of the CNN model and improves SER accuracy by utilizing the publicly accessible benchmark datasets, IEMDMC and RAVDES. The suggested approach is fully explained in following segment, and the challenge part examines the system's efficacy and evaluation.

METHOD

We introduce a Convolutional Neural Network (CNN) based framework for SER in this area. The suggested architecture uses spectrograms and a discriminative Convolutional Neural Network (CNN) for the feature learning to identify the speaker's contentious status. The input layers, convolutional layers, and fully connected the layers of the suggested stride CNN architecture are followed by a SoftMax classifier. When it comes to identifying a speaker's emotions, a speech spectrum analysis signals is two-dimensional depiction More information is provided by the frequencies in relation to time than by the words used in text transcription. When we transform the audio speech signal into phonemes or text., we are unable to extract and apply the rich information contained in spectrograms. This skill makes spectrograms more effective at recognizing emotions expressed verbally.

The fundamental goal is to extract highest level of voice signal discrimination characteristics. To this end, used a Convolutional Neural Network (CNN) architecture, as well as the spectrogram works well for this kind of work. In (8), SER and classification are achieved by combining spectrogram and MFCC data with a Convolutional Neural Network (CNN). The spectrogram characteristics are employed in (32) to attain high SER performance. The following sections outline the main component of the suggested framework.

Pre-Processing

Preparing data for preprocessing is crucial to achieving model efficiency and accuracy. Using an adaptive threshold-based method of preprocessing, we purify the sound signals. in this step for the purpose of get rid of background noise, silent portions, and other unnecessary the speech signal's information.(33) Using a direct relation policy, we could ascertain the connection among amplitude and energy in a voice signal using this way. The magnitude of a wave and its amount of vitality it travels through are related, according to the energy amplitude connection. A high energy wave is indicated by a high amplitude, whereas A low energy is indicated by a low amplitude wave. An element's extreme displacement from its position of rest is indicated by the amplitude of a wave. The energy-amplitude relationship makes sense in the following way to exclude the extraneous and silent particle from voice signals This process includes three steps:

Start by gradually listening to the audio file at 20 000 sample rates. The following stage involves determining the relationship between energy and amplitude between waves, calculating the maximum amplitude using equation (1) in each frame. Passing the data through an appropriate threshold to eliminate noise and highlight the important portions and saving the results in an array. The final stage is creating a new audio file from scratch with identical sampling rates, as well as silent, noiseless transmissions.

In equation (1), stands for the particle displacements, f signifies frequency in connection to time t., and A for the signal peak or amplitude. Figure 1 displays the preprocessing block diagram.

$$D = A0 \times \sin (2 \times \pi \times t \times f)$$

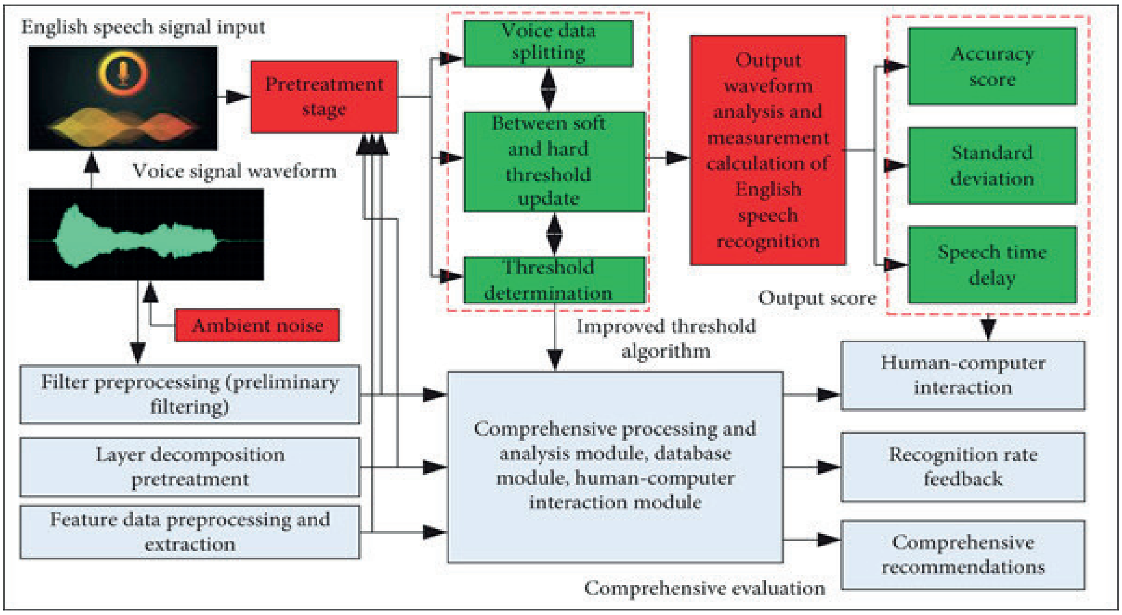
English speech signal input
Pretreatment stage
Voice signal waveform
Ambient noise
Voice data splitting
Between soft and hard threshold update
Threshold determination
Output waveform analysis and measurement calculation of English speech recognition
Accuracy score
Standard deviation
Speech time delay
Output score
Improved threshold algorithm
Filter preprocessing (preliminary filtering)
Layer decomposition pretreatment
Feature data preprocessing and extraction
Comprehensive processing and analysis module, database module, human-computer interaction module
Human-computer interaction
Recognition rate feedback
Comprehensive recommendations
Comprehensive evaluation

Figure 1. A pre-processing block diagram that uses an adaptive threshold setting to improve speech signals

Production of Spectrograms

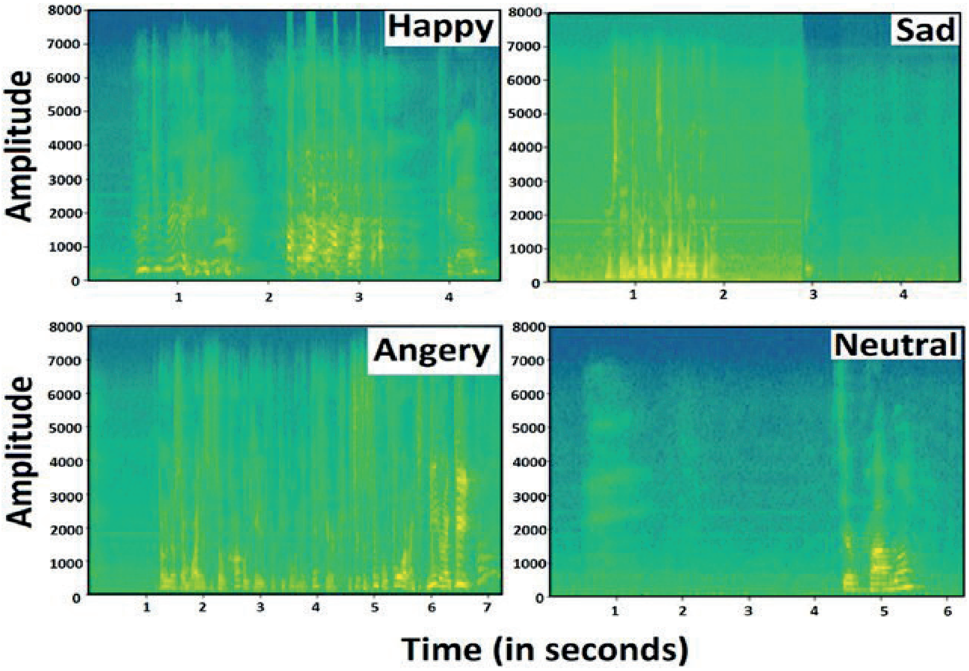

Figure 2. Two-dimension spectrograms of different emotions with visual representations of speech signals

https://doi.org/10.56294/dm2025715

Among the challenging problems in SER using 2D Convolutional Neural Network (CNN) is voice signal's dimension. The main objective of this study is to use the High-level characteristics are extracted from audio signals with a CNN model. Therefore, we must transform the speech signal's one-dimensional representation into a suitable two-dimensional depiction for 2D Convolutional Neural Network (CNN). The most appropriate and optimal two-dimensional representation of audio speech signals that illustrate speech signal strength across a range of frequencies is a spectrogram.[8]

Speech signals are subjected to Short-Term Fourier Transform (STFT) in order to visually display frequencies over various time intervals. Using Quick Fourier Transform (FFT) on a frame to calculate its Fourier spectrum after Converting a Longer Speech Stream using STFT into a shorter, equal length segment or frame. In spectrograms, the frequencies f of each brief interval is symbolized by the y-axis, and the x-axis represents the time t. The comparable the Signal for Speech S (t, f) in Spectrogram S has several type frequencies f across distinct times t. Spectrophotoms with dark colours depict frequencies at low magnitudes whereas those with light hues depict frequencies at greater magnitudes. For several types of speech analysis, including SER, spectrograms are ideal.[34] Figure 2 displays an example of each audio file's extracted spectrograms using STFT.

Convolutional Neural Network -CNN

The most advanced models available today are CNNs, which are used for extracting high-level characteristics from raw pixel data at a low level. Convolutional Neural Network (CNN) extracts high-level characteristics from pictures using several kernels and then uses these features to train a CNN model to carry out important classification tasks.[34] Three parts make up the Convolutional Neural Network (CNN) architecture: convolutional layers, which have several filters that can be applied to input. Every filter generates the number of feature mappings in a single convolutional layer by scanning the input through the submit process utilizing the dot product. Layer pooling, that are accustomed to down sample or reduce the number of dimensions of feature maps, make up the second component. Several strategies, such as maximum and minimum pooling, mean and median pooling that are supplied to etc., are employed to reduce dimensionality. The final element consists of CNN's fully connected layers (FC), which are primarily utilized to extract global characteristics that are fed into a SoftMax classification to determine the likelihood for each class. These layers are arranged in a hierarchical structure by a Convolutional Neural Network (CNN): FC first, followed by the SoftMax classifier, Convolutional Layers (CL), and Pooling Layers (PL). The upcoming section provides an explanation of the suggested architectures.

Proposed Deep Stride Convolutional Neural Network Architecture (DSCNN)

Figure 3 displays the suggested Deep Stride Convolutional Neural Network (DSCNN) model for SER. The concept of plain nets,[14] that is specifically created for computer problems with vision including image categorization, identification, tracking, and localization to ensure high-level accuracy[35] largely inspires our deep stride Convolutional Neural Network (CNN) model. From image classification to speak emotion identification and categorization, we investigate the basics network, we present the stride deep Convolutional Neural Network (CNN) architecture, which learns deep features using a little same field in convolutional layers by utilizing a similar and small filter size (3 × 3). It is based on straightforward principles: if the size of the feature maps is cut in half, the number of filters needs to be doubled to preserve the complexity of time for each layer. The same number of kernels produces the same output feature maps.

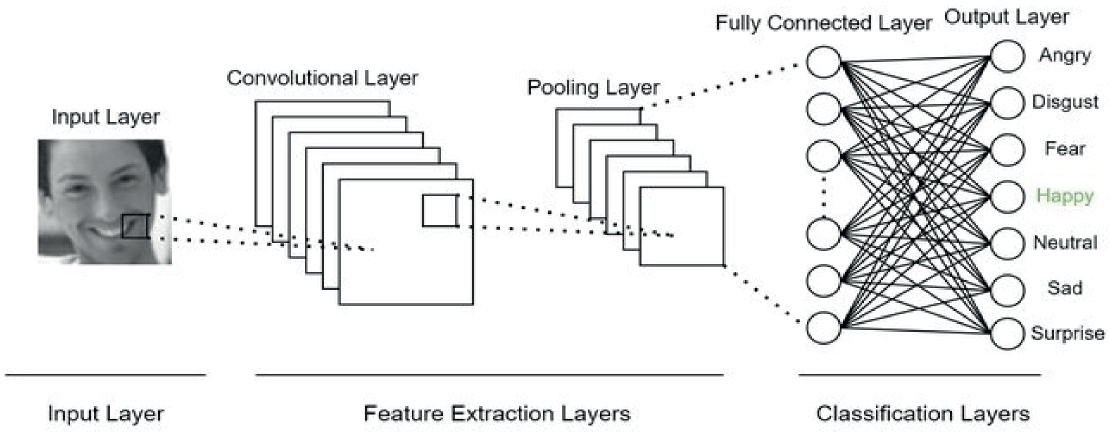

Figure 3. The general Block Diagram of the proposed Deep Stride Convolutional Neural Network (DSCNN) for voice feeling identification is displayed

https://doi.org/10.56294/dm2025715

To implement this approach, we created the DSCNN model for SER, which down-samples feature map sizes straight into convolutional layers instead of the pooling layer using the stride (2 x 2) scheme. (9) layers make up the DSCNN; (2) completely connected layers and (7) convolutional layers are sent into SoftMax to produce the speech emotion probabilities. Convolutional filters are applied to the resulting spectrograms as input to derive maps of features from a certain voice spectrogram.

We set up CL, in the suggested design in a sequential manner. 16 squire-shaped kernels with a size of 7 × 7 are applied to the first convolutional layer's input spectrogram (C1)., the same (2 × 2) pixel padding and stride setting. Likewise, Features of the second convolutional layer (C2) 32 filters with a stride the size of 5 × 5 and the setting of 2 × 2, the same amount is used by the C3 layer of padding, striding and filtering the C2 layer; however, the filters are 3 × 3. The 64 (3 × 3) kernels in the C4 and C5 layers have a stride setting of 2 × 2 pixels. Similarly, the C6 and C7 layers feature 128 (3 × 3) kernels with the same padding and stride. The final convolutional layer is followed by a flattening layer, which feeds the features to FC after transforming the data shape into vector form. There are 512 neurons in the first FC layer, and the final FC layer has as many neurons as there are classes. The suggested DSCNN uses batch normalization and the rectified linear unit activation function to regularize the model after each convolutional layer. Following the initial FC, a 25 % dropout ratio is used to mitigate the model's overfitting.[35] To determine the likelihood of each class, the final FC layer was sent to the SoftMax classifier. Spectrograms are used in the design of the DSCNN for SER. Convolutional and FC layers are present in the DSCNN, which also uses a small filter size and unique steps and removes the pooling layer from the entire network. To extract discriminative features from spectrograms, CL first uses a large kernel size to learn local characteristics. It then gradually increases the number of kernels while maintaining the same filter size and shape. Employing the identical filter size and shape across the network to discover deep features, restricting the quantity of FC levels to prevent duplication, and down sampling while removing the pooling layer are the essential elements of this architecture, the robust prominent features using spectrograms are captured by the proposed DSCNN method because of the qualities.

Model Organization and Computational Setup

Python is used to implement the suggested DSCNN model layout, with additional resources including the machine learning software scikit-learn. The 128 × 128 spectrograms are produced from every file. To train and test, the entire set of produced spectrograms is split into two halves, 20 % and 80 %, respectively. Regarding the suggested DSCNN prototype for SER, the model training procedure was evaluated on a single 12-GB on-board graphics card NVIDIA GeForce GTX 1070. With decaying every 15 epochs utilizing a 0,001 learning rate, the model was trained over 50 epochs. The maximum accuracy was attained after 50 epochs, with 0,3225 lost during training and 0,5472 lost during validation. The batch size for the whole training phase is 128, using 34,10 MB model size that is smaller, the model trains quickly, demonstrating its simplicity of computation.

RESULTS

Within this section, we are used spectrograms that assess our SER model on the RAVDESS and IEMDMC datasets. Using spectrograms, the suggested Convolutional Neural Network (CNN)models' performance is compared to modern CNN architectures for SER. Several experiments were carried out for SER; the specifics and findings are covered in the section that follows.

Data sets

Interactive Emotional and Dyadic Motion Capture (IEMDMC)

Ten actors are used in the acted English and multilingual language speech emotion dataset known as IEMDMC[12] to record a range various feelings, such as fear, rage, happiness, disgust, sadness, neutrality, and excitement. Twelve hours of audiovisual data total from five sessions make up the IEMDMC dataset. Two performers record a screenplay in a range of emotions during each session. Only 4 emotions angry, joyful, neutral, and sad were used in our work to compare experimental evaluations with cutting-edge methods.

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

One English and Multilingual language emotional dataset that is frequently utilized for emotional songs and voices detection is the Ryerson audio-visual library of emotional speech and song.[13] Twelve male and twelve female performers make up the dataset, which is used to record 8 different feelings, encompassing angry, calm, joyful, sad, surprised, disgusted, neutral, afraid etc. There are 1450 wav files overall with a 50 000 Hz sample frequency of utterances.

Experimental Evaluations

In this part, we tested our suggested method using the IEMDMC and RAVDESS benchmark datasets. Two types of spectrograms were used in our experiments: raw spectrograms and spectrograms created directly from audio recordings. Second, the audio signals are treated to eliminate the silence and noise component before being turned into spectrograms on the clean spectrogram. Our research on SER involved utterance-based tests using a five-fold cross-validation method. 85 percentage of the information were utilized for training and 20 percent were used for testing the model. We ran 2 sets of experiments. In the first, we used raw spectrograms to train a DSCNN model and evaluated the model's accuracy and efficiency in making predictions. We assessed the model's performance in the second set after training it on clean spectrograms. Evaluate the model utilizing the subsequent metrics: precision, recall, f1 score, weighted (class-specific accurately forecasted samples of emotion divided by class-specific motion samples) and unweighted accuracy (total samples correctly predicted divided by total samples in the dataset). The suggested model's training results are presented in Tables 1-3, and the discussion sections include the in-depth analysis and comparisons.

Table 1 made it evident how the two data sets' raw and clean spectrograms differed. For this reason, rather than using pooling layers to reduce the feature maps' sample size, the suggested The DSCNN model made use of convolutional layers.[14] When utilizing spectrograms to identify emotions from speech data, the performance is superior to the pooling approach.

Table 1. A contrasting it suggested model with both clean and raw spectrograms

Input	Data-set	Weighted- acc %	Unweighted- acc%	F1 -score%
RAW- SPEC	IEMDMC	77	73	80
CLEAN- SPEC	IEMDMC	85	78	85
RAWS- PEC	RAVDESS	69	83	71
CLEAN -SPEC	RAVDESS	82	63	85

Results and Performance of Deep Stride Convolutional Neural Network (DSCNN) Model

Utilizing the Convolutional Neural Network (CNN) architecture of simple nets,[14] we created the DSCNN method for SER and conducted studies on speech spectrograms obtained from utterances. Applying a dividing approach of 80 %/20 %, the DSCNN method was taught using these produced spectrograms. The two common benchmark datasets utilized for instruction and testing the algorithms prediction outcomes were IEMDMC and RAVDESS, respectively. Using the IEMDMC dataset, table 2 displays the performance of the system training on clean and unprocessed spectrograms, comprising weighted and unweighted accuracy, f1 score, recall, and precision at the class level. The total performance utilizing the RAVDESS dataset is shown in table 3, which includes weighted and unweighted accuracy, recall, precision and accuracy at the class level, and f1 score for both raw and clean spectrograms.

Table 2. Shows how well the suggested DSCNN method trained on clean and raw spectrograms made with IEMDMC

Nature emotion	Result on raw spectrograms			Result on clean spectrograms		
	Precision	Recall	F1score	Precision	Recall	F1score
ANGER	0,97	0,88	0,90	0,88	0,97	0,92
HAPPY	0,59	0,86	0,70	0,98	0,68	0,79
NEUTRAL	1,10	0,78	0,88	0,78	0,89	0,85
SAD	0,70	0,90	0,80	0,84	0,93	0,85
WEIGHTED AVG	0,79	0,76	0,77	0,87	0,86	0,86
UNWEIGHTED AVG	0,76	0,74	0,74	0,87	0,85	0,84
ACCURACY	---	---	0,79	---	---	0,86

Table 3. Shows suggested DSCNN method trained in raw cleaning and raw utilizing RAVDESS to create spectrograms

Nature emotion	Result on raw spectrograms			Result on clean spectrograms		
	Precision	Recall	F1score	Precision	Recall	F1score
Anger	0,45	1,10	0,58	0,80	0,92	0,85
Happy	0,93	0,30	0,45	0,80	0,91	0,85
Neutral	0,92	0,43	0,58	0,72	1,10	0,84
Sad	0,99	0,99	0,99	0,99	0,97	0,94

Clam	0,83	0,76	0,79	0,73	0,95	0,82
Fearful	0,00	0,00	0,00	1,10	0,55	0,68
Surprised	0,92	0,47	0,60	0,90	0,88	0,89
Disgust	0,93	0,87	0,90	1,10	0,39	0,56
Weighted avg	0,80	0,72	0,69	0,86	0,82	0,82
Unweighted avg	0,74	0,60	0,62	0,86	0,82	0,80
Accuracy	---	---	0,72	---	---	0,83

Prediction Performance of Proposed Deep Stride Convolutional Neural Network (DSCNN)

To demonstrate an effectiveness of system, the prediction performance of suggested DSCNN method is assessed using the RAVCDESS and IEMDMC datasets. Table 3 displays the methods prediction performance IEMDMC dataset in terms of confusion matrix. The method's confusion matrix for the Table 4 displays the RAVDESS dataset. The suggested Convolutional Neural Network (CNN) model's prediction performance increases overall prediction accuracy, demonstrating the model's relevance and resilience.

Table 4 shows confusion matrix for forecasting emotions of IEMOCAM. Then average recall value for this matrix is 82,75 %, and each row shows how each emotion is confused between the predictions and the ground truth.

Table 4. Confusion matrix for forecasting emotions of IEMOCAM

Emotion class	Anger	Happy	Neutral	Sad
Anger	0,93	0,064	0,014	0,033
Happy	0,093	0,745	0,084	0,093
Neutral	0,00	0,073	0,905	0,024
Sad	0,00	0,025	0,253	0,733
Overall accuracy 82,75 %				

An IEMDMC dataset's total prediction performance for each of the four emotion files is displayed in Table 4. Even Nevertheless, the model's overall performance (82,75 %) is good for the IEMDMC dataset, with happy and sorrow receiving significantly less prediction accuracy than angry and neutral. In a similar vein, table 5 shows the RAVDESS dataset's prediction performance for eight classes. Although fear and disgust are partly intermingled with clam and rage emotions, the model obtains superior prediction for all classes in the RAVDESS dataset, and its overall prediction accuracy of 80,5 percent is good. Table 6 lists the training and testing accuracy of the suggested DSCNN model for each of the two datasets.

Table 5 confusion matrix for RAVDESS emotion prediction; each row shows how each emotion is confused between the ground truth and the forecasts, with an average recall value of 80,5 %.

Table 5. Confusion matrix for RAVDESS emotion prediction

Emo class	Anger	Clam	Disgust	Fear	Happy	Neutral	Sad	Surprised
Anger	0,84	0,00	0,00	0,00	0,16	0,00	0,00	0,04
Clam	0,00	0,84	0,00	0,00	0,00	0,16	0,00	0,00
Disgust	0,19	0,17	0,53	0,00	0,00	0,06	0,00	0,10
Fear	0,22	0,15	0,00	0,44	0,09	0,12	0,00	0,05
Happy	0,05	0,00	0,00	0,00	0,88	0,03	0,03	0,05
Neutral	0,00	0,04	0,00	0,00	0,00	0,96	0,00	0,02
Sad	0,00	0,00	0,00	0,00	0,03	0,05	0,95	0,00
Surprised	0,00	0,02	0,00	0,00	0,00	0,00	0,00	0,99
Overall accuracy 80,5 %								

Table 6. shows the suggested DSCNN model's accuracy in testing and training

Data set	Training Accuracy	Testing Accuracy
IEMDMC	84 %	82,75 %
RAVDESS	81 %	80,50 %

Experiment using Cross Datasets

These days, many researchers have created sufficient method for SER utilizing one corpus has produced good recognition accuracy in tests. Many obstacles, such as diverse linguistic and cultural backgrounds and drastically changing speaker populations, affect the model's performance when it is used in a natural setting for SER. We assess the efficacy of our model by comparing its performance across four different emotions in the RAVDESS and IEMDMC datasets.(36) We use an IEMDMC dataset for model training, and the RAVDESS dataset is used for model testing. Table 7 displays the cross-dataset confusion matrix.

Table 7. Confusion matrix for predicting emotions on a cross-corpus between IEMDMC and RAVDESS

Emotion class	Anger	Happy	Neutral	Sad
Anger	0,78	0,09	0,06	0,10
Happy	0,13	0,45	0,02	0,45
Neutral	0,02	0,14	0,58	0,34
Sad	0,00	0,53	0,00	0,50
Overall accuracy 57,5 %				

Table 7 displays proposed model's outcomes across multiple datasets, demonstrating its efficacy in identifying the 4 emotions with an average identification rate of 57,5 %. The recognition rates of the emotions of anger (78 %), sadness (50 %), neutral (57 %), and happiness (45 %), in a cross-corpus analysis, are as follows. The cross-dataset experiments demonstrate the model's importance and resilience.

DISCUSSION

The primary contributions of this work are the architecture of the DSCNN method and utilizing threshold-based speech signal preprocessing to eliminate noisy and unimportant parts of speech. The concept of plain nets, which are specifically made Our deep stride CNN (model) acknowledgement is greatly motivated to achieve a high degree of precision for computer vision problems such as image classification, localization, tracking, and other related tasks.(37) We define the stride deep Convolutional Neural Network-(DCCNN) architecture and investigate the plain network for SER. It has usually followed straightforward guidelines and employed a small (3 × 3) filter size since there are now more kernels in CL. The resulting feature maps are identical due to the same number of kernels. To keep the time complexity per layer constant, the number of filters needs to be doubled if the size of the feature maps is cut in half. Instead of using the pooling layer, we have directly down sampled the size of feature maps in convolutional layers using the stride (2 × 2). Nine (9) layers make up the DSCNN; the SoftMax layer receives input from seven (7) CL and two (2) FC layers to forecast the likelihood of speech emotions. Our model's value lies in its smaller kernel size and fewer convolutional layers, which enable it to learn deep salient features with a small corresponding field. lower complexity compared to other cutting-edge deep learning techniques and produces excellent SER accuracy. The pooling approach is not more robust in signal processing, but it performs well in computer vision tasks including image recognition, tracking, and retrieval. As a result, the stride Convolutional Neural Network (CNN) model performed better than other cutting-edge techniques.

Table 8. Using the IEMDMC dataset, compare the suggested strategy with the baseline approach

Method	Input	Weighted accuracy	Unweighted accuracy	Accuracy
Fayek et al.(38)	Spectrograms	65,78 %	60,99 %	---
Luo et al.(39)	Spectrograms	61,35 %	62,98 %	---
Tripathi et al.(40)	Spectrograms	72,3 %	61,7 %	---
Yenigalla et al.(41)	Spectrograms	73,00 %	68,05 %	---
Chen et al.(42)	Spectrograms	---	---	64,74 %
Proposed method	Raw_spectrograms	76 %	72 %	74,8 %
Proposed method	Clean_spectrograms	85 %	83 %	82,75 %

The performance of suggested approach is shown in table 8 in contrast to other cutting-edge methods. It performs better than the current findings when employing utterance level speech spectrograms on the IEMDMC dataset. An approach based on RNNs, and neural networks of SER was described by Haytham et al.(38) The method uses using frame spectrograms to train DNN, which is computationally expensive and does not produce very accurate results. Transcript and phoneme were used by Tripathi et al.[43] to define the deep learning methods for SER. The accuracy was increased to 71 % by training many models with different features, but we all employed the same architecture utilized for computer vision related tasks. To improve SER accuracy, Chen et al.[42] created a system for SER utilizing a trained model and 3D CNN architecture, but he also employed the pooling strategy to build the network. Because of this restriction, we looked at the basic Convolutional Neural Network (CNN) architecture and came up with a novel model for SER that performs better than state-of-the-art findings. On the RAVDESS speech emotion dataset, table 9 compares the suggested framework with the baseline technique. It illustrates the importance and effectiveness of the suggested DSCNN model on the RAVDESS dataset, outperforming the SER findings. A spectrogram-based Convolutional Neural Network (CNN) model was published by Yuni et al.[43] for multi-class audio classification. By combining two models, the model achieved an accuracy of 64,48 % in multitasking. SER. Using the log spectrogram and spectral feature, Jalal et al.[44] and Anjali et al.[45] were able to identify the emotion in voice data with 68 % and 75 % accuracy, respectively. Using the IEMDMC dataset for SER, table 10 compares the computational simplicity of the proposed DSCNN model with various baseline Convolutional Neural Network (CNN) models.

Table 9. Using the RAVDESS dataset, compare the suggested strategy with the baseline approach

Method	Input	Weighted accuracy	Unweighted accuracy	Accuracy
Zengetal.[43]	Spectrograms	--	--	64,50 %
Jalaletal.[44]	Log-spectrogram	--	70,4 %	68,15 %
Bhavanetal.[45]	Spectral features	--	--	75,70 %
Proposed model	Raw_ Spectrograms	68 %	62 %	70,10 %
Proposed model	Clean_ Spectrograms	80 %	80 %	80,5 %

Table 10. A computational comparison between the proposed DSCNN model and existing baseline Convolutional Neural Network (CNN) algorithms presented

Method	Training Time	Model Size	Accuracy
Alex-Net (transfer Learning)[46]	38min	201MB	70,54 %
Vgg16(transfer Learning)[37]	55min	420MB	73,00 %
ResNet50(transfer Learning)[14]	30min	75MB	75,50 %
Proposed DSCNN model	14min	34,5MB	82,75 %

On the Interactive Emotional Dyadic Motion Capture (IEMDMC) and RAVDESS datasets for SER, the suggested DSCNN model was also contrasted with cutting-edge CNN models in terms of training time, model size, and accuracy. Table 10 presents the efficacy and importance of the suggested model in comparison to baseline techniques. It illustrates how the suggested model is computationally simple in terms of training time, model size, and accuracy on speech signal spectrograms that are generated. utilizing the IEMDMC dataset, we train the Convolutional Neural Network (CNN) models Alex-Net, Vgg-16,[47] and Resnet-50[48] utilizing transfer learning approaches. The effectiveness and relevance of the proposed Convolutional Neural Network (CNN) model for SER were demonstrated by the proposed DSCNN model, which outperformed findings.

CONCLUSIONS

There are too many obstacles in the way of the SER literature's attempts to reduce the overall model's computing complexity and increase recognition accuracy. In response to these difficulties, we suggested a Convolutional Neural Network (CNN) design that includes a few key features extraction mechanisms to reduce computational complexity and increase accuracy of the SER model. In this research, we extracted speech signals from noise and silence data using a dynamic adaptive threshold approach. Then, to improve the accuracy and reduce the computational complexity of the suggested model, the improved speech signals are transformed into spectrograms. As an alternative to pooling layers, we employed stride Convolutional Neural Network (CNN) architectures for SER using spectrograms to learn most salient and discriminative features in a convolutional layer using some unique strides set to down-sample feature maps. Two common benchmark datasets, IEMDMC and RAVDESS, are used to assess the efficacy of the suggested model. The findings hold up well and can identify different moods in voice patterns. Our approach produces a computationally-friendly output system with accuracy up to 80,5 % when used to Ryerson Audio Visual Database of Emotional Speech and Song (RAVDESS) and 82,75 % when applied to an Interactive Emotional Dyadic Motion Capture (IEMDMC) dataset with less parameters in the model. This shows how important and successful the suggested system for SER using voice signal spectrograms.

BIBLIOGRAPHIC REFERENCES

1. Grewe, L.; Hu, C. ULearn: Understanding and reacting to student frustration using deep learning, mobile vision and NLP. In Proceedings of the Signal Processing, Sensor/Information Fusion, and Target Recognition XXVIII, Baltimore, MD, USA, 7 May 2019; p. 110180W.

2. Wei, B.; Hu, W.; Yang, M.; Chou, C.T. From real to complex: Enhancing radio-based activity recognition using complex-valued CSI. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 35.

3. Zhao, W.; Ye, J.; Yang, M.; Lei, Z.; Zhang, S.; Zhao, Z. Investigating capsule networks with dynamic routing for text classification. arXiv 2018, arXiv:1804.00538.

4. Sabour, S.; Frosst, N.; Hinton, G.E. Dynamic routing between capsules. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017; pp. 3856-3866.

5. Bae, J.; Kim, D.-S. End-to-End Speech Command Recognition with Capsule Network. In Proceedings of the Interspeech, Hyderabad, India, 2-6 September 2018; pp. 776-780.

6. Fiore, U.; Florea, A.; Pérez Lechuga, G. An Interdisciplinary Review of Smart Vehicular Traffic and Its Applications and Challenges. J. Sens. Actuator Netw. 2019, 8, 13.

7. Kim, S.; Guy, S.J.; Hillesland, K.; Zafar, B.; Gutub, A.A.-A.; Manocha, D. Velocity-based modeling of physical interactions in dense crowds. Vis. Comput. 2015, 31, 541-555.

8. Badshah, A.M.; Rahim, N.; Ullah, N.; Ahmad, J.; Muhammad, K.; Lee, M.Y.; Kwon, S.; Baik, S.W. Deep features-based speech emotion recognition for smart affective services. Multimed. Tools Appl. 2019, 78, 5571-5589.

9. Mao, Q.; Dong, M.; Huang, Z.; Zhan, Y. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Trans. Multimed. 2014, 16, 2203-2213.

10. Kang, S.; Kim, D.; Kim, Y. A visual-physiology multimodal system for detecting outlier behavior of participants in a reality TV show. Int. J. Distrib. Sens. Netw. 2019, 15.

11. Dias, M.; Abad, A.; Trancoso, I. Exploring hashing and cryptonet based approaches for privacy-preserving speech emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15-20 April 2018; pp. 2057-2061.

12. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMDMC: Interactive emotional dyadic motion capture database. Lang. Resour. Eval. 2008, 42, 335.

13. Livingstone, S.R.; Russo, F.A. 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 2018, 13.

14. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June-1 July 2016;pp. 770-778.

15. Jiang, S.; Li, Z.; Zhou, P.; Li, M. Memento: An Emotion-driven Lifelogging System with Wearables. ACM Trans. Sens. Netw. (TOSN) 2019, 15, 8.

16. Erol, B.; Seyfioglu, M.S.; Gurbuz, S.Z.; Amin, M. Data-driven cepstral and neural learning of features for robust micro-Doppler classification. In Proceedings of the Radar Sensor Technology XXII, Orlando, FL, USA, 16-18 April 2018; p. 106330J..

17. Dave, N. Feature extraction methods LPC, PLP and MFCC in speech recognition. Int. J. Adv. Res. Eng. Technol. 2013, 1, 1-4

18. Luque Sendra, A.; Gómez-Bellido, J.; Carrasco Muñoz, A.; Barbancho Concejero, J. Optimal Representation of Anuran Call Spectrum in Environmental Monitoring Systems Using Wireless Sensor Networks. Sensors 2018, 18, 1803.

19. Liu, G.K. Evaluating Gammatone Frequency Cepstral Coefficients with Neural Networks for Emotion Recognition from Speech. arXiv 2018, arXiv:1806.09010.

20. Liu, Z.-T.; Wu, M.; Cao, W.-H.; Mao, J.-W.; Xu, J.-P.; Tan, G.-Z. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing 2018, 273, 271–280.

21. Liu, C.-L.; Yin, F.; Wang, D.-H.; Wang, Q.-F. CASIA online and offline Chinese handwriting databases. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 37–41.

22. Fahad, M.; Yadav, J.; Pradhan, G.; Deepak, A. DNN-HMM based Speaker Adaptive Emotion Recognition using Proposed Epoch and MFCC Features. arXiv 2018, arXiv:1806.00984.

23. Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. IEEE Trans. Multimed. 2017, 20, 1576–1590.

24. Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schuller, B.; Zafeiriou, S. Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 5200–5204.

25. Wen, G.; Li, H.; Huang, J.; Li, D.; Xun, E. Random deep belief networks for recognizing emotions from speech signals. Comput. Intell. Neurosci. 2017, 2017.

26. Zhu, L.; Chen, L.; Zhao, D.; Zhou, J.; Zhang, W. Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. Sensors 2017, 17, 1694.

27. Hajarolasvadi, N.; Demirel, H. 3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms. Entropy 2019, 21, 479.

28. Tao, F.; Liu, G. Advanced LSTM: A study about better time dependency modeling in emotion recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2906–2910.

29. Sahu, S.; Gupta, R.; Sivaraman, G.; Abd Alma geed, W.; Espy-Wilson, C. Adversarial auto-encoders for speech-based emotion recognition. arXiv 2018, arXiv:1806.02146.

30. Bao, F.; Neumann, M.; Vu, N.T. Cycle GAN-based emotion style transfer as data augmentation for speech emotion recognition. Manuscripts. Submit. Publ. 2019, 35–37.

31. Yu, D.; Seltzer, M.L.; Li, J.; Huang, J.-T.; Seide, F. Feature learning in deep neural networks-studies on speech recognition tasks. arXiv 2013, arXiv:1301..

32. 3605 Liu, P.; Choo, K.-K.R.; Wang, L.; Huang, F. SVM or deep learning? A comparative study on remote sensing image classification. Soft Comput. 2017, 21, 7053–7065

33. Alkaya, A.; Eker, I˙. Variance sensitive adaptive threshold-based PCA method for fault detection with experimental application. ISA Trans. 2011, 50, 287–302. [PubMed]

34. Abdel-Hamid, O.; Mohamed, A.-r.; Jiang, H.; Deng, L.; Penn, G.; Yu, D. Convolutional neural networks for speech recognition. IEEE/ACM Trans. Audio Speech Lang. Process. 2014, 22, 1533–1545.

35. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. 2014, 15, 1929–1958.

36. Latif, S.; Qayyum, A.; Usman, M.; Qadir, J. Cross Lingual Speech Emotion Recognition: Urdu vs. Western Languages. In Proceedings of the 2018 International Conference on Frontiers of Information Technology (FIT), Islamabad, Pakistan, 17–19 December 2018; pp. 88–93.

37. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv2014, arXiv:1409.1556.

38. Fayek, H.M.; Lech, M.; Cavedon, L. Evaluating deep learning architectures for Speech Emotion Recognition. Neural Netw. 2017, 92, 60-68.

39. Luo, D.; Zou, Y.; Huang, D. Investigation on Joint Representation Learning for Robust Feature Extraction in Speech Emotion Recognition. In Proceedings of the Interspeech, Graz, Austria, 19 September 2019; pp. 152-156.

40. Tripathi, S.; Kumar, A.; Ramesh, A.; Singh, C.; Yenigalla, P. Deep Learning based Emotion Recognition System Using Speech Features and Transcriptions. arXiv 2019, arXiv:1906.05681.

41. Yenigalla, P.; Kumar, A.; Tripathi, S.; Singh, C.; Kar, S.; Vepa, J. Speech Emotion Recognition Using Spectrogram & Phoneme Embedding. In Proceedings of the Interspeech, Hyderabad, India, 2-6 September 2018; pp. 3688-3692.

42. Chen, M.; He, X.; Yang, J.; Zhang, H. 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. IEEE Signal Process. Lett. 2018, 25, 1440-1444.

43. Zeng, Y.; Mao, H.; Peng, D.; Yi, Z. Spectrogram based multi-task audio classification. Multimed. Tools Appl. 2019, 78, 3705-3722.

44. Jalal, M.A.; Loweimi, E.; Moore, R.K.; Hain, T. Learning Temporal Clusters Using Capsule Routing for Speech Emotion Recognition. Proc. Interspeech 2019, 2019, 1701-1705.

45. Bhavan, A.; Chauhan, P.; Shah, R.R. Bagged support vector machines for emotion recognition from speech. Knowl.-Based Syst. 2019, 184, 104886.

46. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4-9 December 2017; pp. 1097-1105.

47. Molchanov, P.; Tyree, S.; Karras, T.; Aila, T.; Kautz, J. Pruning convolutional neural networks for resource efficient transfer learning. arXiv 2016, arXiv:1611.06440.

48. George, D.; Shen, H.; Huerta, E. Deep Transfer Learning: A new deep learning glitch classification method for advanced LIGO. arXiv 2017, arXiv:1706.07446.

FINANCING

None.

CONFLICT OF INTEREST

Authors declare that there is no conflict of interest.

AUTHORSHIP CONTRIBUTION

Conceptualization: Chandupatla Deepika, Swarna Kuchibhotla.
Data curation: Chandupatla Deepika, Swarna Kuchibhotla.
Formal analysis: Chandupatla Deepika, Swarna Kuchibhotla.
Drafting - original draft: Chandupatla Deepika, Swarna Kuchibhotla.
Writing - proofreading and editing: Chandupatla Deepika, Swarna Kuchibhotla.