Journal of Harbin Engineering University Vol 44 No. 7
ISSN: 1006-7043 July 2023

Deep learning for speech emotion feature extraction and classification: current trends andfuture directions

Himashri Deka[1*], Vikas Mittal[1]

[1*]Department of Electronics and Communication Engineering,
Kurukshetra 136119, Haryana, India.

*Corresponding author(s).
E-mail(s): himashrideka.nitkkr@gmail.com;
Contributing authors: vikas mittal@nitkkr.ac.in;

Abstract-Speech Emotion Recognition (SER) systems have been very useful in different domains like social media, customer satisfaction, etc. Traditional SER systems use old datasets and feature extraction techniques that render the recognition less reliable and robust. In recent times, with the advancement of deep learningalgorithms and the production of massive amounts of speech-emotion data, the use of unsupervised learning has been widely used to automate feature extrac- tions. Additionally, deep learning algorithms have also been proven very effective for emotion classification and transfer learning of emotions. In this paper, we cover recent trends of deep learning algorithms in speech emotion feature extrac- tion and classification, along with their comparative study. We also present the advantages of deep learning algorithms in building SER over traditional SER. Further, we have also compared the advantages of unsupervised learning over supervised learning due to the variation in speech emotion data. We have also discussed popular datasets along with some recently developed datasets to leverage the development of SER.

Keywords: Speech emotion recognition, Feature extraction, Unsupervised Learning, Deep Learning, Datasets

1 Introduction

Identification of the emotional state of a speaker from his or her speech is defined as SER. Speech has been one of the most promising modalities for the automatic recognition of human emotions, aside from facial expressions [1]. In the continuously evolving area of Human-Computer Interaction (HCI), SER is a crucial topic of interest. It has wide potential applications, such as the interface with robots, banking, detection of lies, health assistance including psychiatric aid, call centers, cardboard systems, video games, computer games, etc.[2]. The multidisciplinary nature of SER research and applications is demonstrated in Fig. 1, with 20.4% of Scopus documents on SERbeing from the engineering domain.

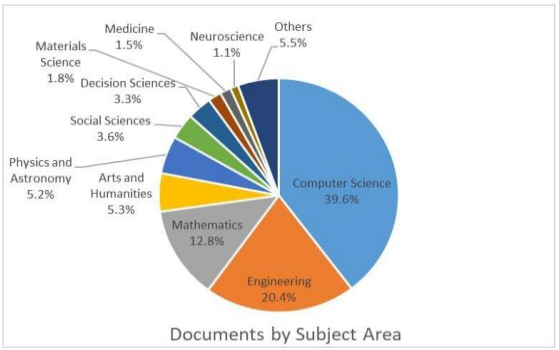


Fig. 1 Distribution of SER in terms of subject area.

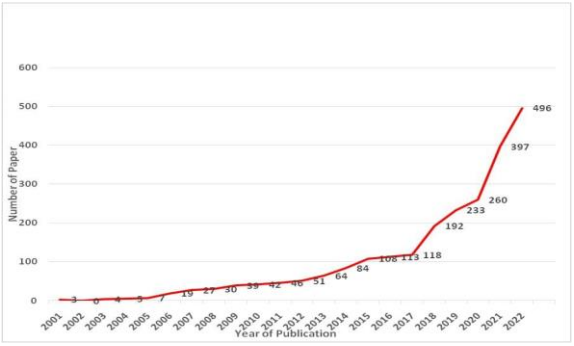
Williamson developed the very first method to interpret a person's emotional state from a speech in the late 1970s [2] by developing a speech analyzer to identify an underlying emotion by examining changes in pitch or frequency.

The first research publication [3] on this subject was published in 1996, which also introduced statistical strategies for recognizing patterns present in voice emotions. The Hidden Markov Model(HMM) [4] was observed to perform best with high accuracy for identifying joy, sorrow, rage, boredom, fear, and neutral emotions using energy and pitch variables. Support Vector Machine (SVM), Decision Trees(DT), and Gaussian Mixture Model(GMM) are a few among the early machine learning models that have consistently served their purpose over time. During the 2000s, studies on voice emotion recognition frequently used Neural Networks(NN) [5]. However, over the last decade, deep learning models have become more popular, yielding encouraging results, and are fast, replacing the conventional machine learning models for SER.

Deep learning algorithms [6] have a multi-layer architecture where each layerconsists of multiple neurons. Multiple layers help in learning multiple levels of abstrac- tion from data. In a hierarchical manner, multiple layers learn higher-level features from lower-level features. Some of the deep learning algorithms highly used in SER are Recurrent Neural Networks(RNN), Long short-term Memory Networks(LSTM), Bidirectional Long short-term Memory Networks(BLSTM), Convolutional Neural Net- works(CNN), and Deep Convolutional Neural Network(DCNN). Multitask learning and attention mechanisms are also currently being exploited for better performance. The transfer learning technique[7] is commonly utilized for cross-lingual and cross- corpus speech emotion recognition. Speech emotion recognition has been a very important topic in research in recent decades.

Due to its inherent multi-disciplinary nature, the field of SER has benefitted fromthe other fields of research and, in turn, has also benefited many of the other research fields. This has resulted in the trend of growing research work in the field of SER in the last 20 years, as the graph below depicts.

Fig. 2 Distribution of the papers in terms of years.



The general architecture of the SER system has three components, as described below. A speech processing system that extracts key parameters or features from the signal, such as energy or pitch. These parameters are concise into reduced sets for efficient computer processing. A classifier that associates the parameters to the emotions is then applied to recognize the emotion.

Section 2 gives a literature review of SER along with a comparison of the recent works. Section 3 discusses the most common databases used for speech-emotion recognition (SER). Section 4 discusses the traditional and machine learning-based feature extraction technique along with their relative drawbacks and advantages. Section 5 discusses the classifier used in SER. Section 6 gives an idea about the open research challenges in this field, followed by a conclusion in section 7.

2 State of the Art

In this section, a comparative review of SER related work is done by focusing on six different criteria, i.e., research area, proposed solution, experimentation result, observation, and possible suggestions.

Authors in [8] addressed three key questions for building an effective SER systemas :

1. What feature set would be best?
2. What auditory elements should be included for the most reliable automatic recognition of a speaker's emotion?
3. Which categorization method is the best one?

They highlighted the importance of having a sufficient and suitable collection of phrases in the database to effectively train the emotion recognition system and eval- uate its performance. In their experimental work, Multivariate Linear Regression (MLR), Recurrent Neural Networks (RNN), and Support Vector Machine (SVM) classifiers have been utilized for the classification of speech emotion. Authors at [9] investigated the Golden Ratio-based Equilibrium Optimization (GREO) algorithm, which combines the Equilibrium Optimization (EO) and the Golden Ratio Optimiza- tion (GRO) approaches [10]. The proposed hybrid Feature Selection (FS) method, referred to as Golden Ratio-based Equilibrium Optimization (GREO), combines the Equilibrium Optimization (EO) and Golden Ratio Optimization (GRO) algorithms. It aims to reduce the feature vector size and enhance the accurateness of the Speech Emotion Recognition (SER) task. The key objective of the given model is to identify the optimal feature set that increases classification accuracy while minimizing computational and storage costs. The EO and GRO optimization algorithms form the foundation of the proposed hybrid meta-heuristic FS technique. The suggested technique has demonstrated recognition accuracies of 98.46% and 97.31% on two renowned, publicly accessible SER datasets, EmoDB and SAVEE respectively.

Authors at [11] presented a fresh set of harmonic features for speech emotion identification inspired by psychoacoustic perception from music theory. Authors at [12] demonstrated the advantages of utilizing pre-trained networks in complex multi-class picture categorization by initializing network training procedures with pre-trained

network parameters. They explored the possible application of the most powerful pre-trained neural networks, originally trained for image classification, to the problem of SER. The advantages of pre-trained networks in the initialization of network training and the requirement of lower numbers of training data can be exploited. By doing this, the training procedure can be shortened to a quick fine-tuning process that utilizes a modest training data set. The classification of images has been a particular strength of pre-trained networks.

According to recent research, the task of the speech classification challenge can be reformed to an image classification problem[13] and resolved with the aid of an image classification network that has already been trained. By computing speech amplitude spectrograms and converting them into RGB images, the speech-to-image transition was accomplished.

Authors at [14] noted a spoken signal's

spectral energy distribution changes depending on how it is emotionally charged. Emotions are divided into low-arousal and high-arousal categories as a result. High-arousal emotions like happiness or anger exhibit greater energy at higher frequencies compared to low-arousal emotions like melancholy. The research paper proposed utilizing semi-Nonnegative Matrix Factor-ization (semi-NMF) with singular value decomposition (SVD) initialization for feature optimization.

In order to reduce the issue of unreliable information loss and the resulting volatil- ity in the system's performance, authors at[15] proposed an approach that involves transforming the feature vector into a lower rank output vector through linear trans-formation. Spectral characteristics[16] are better able to accurately define emotional components in speech than other speech features.

Authors at [17] proposed a PCA-based dimensionality reduction approach to deal with the higher dimension of the dataset for SER. The best solution for the high dimensionality issue is feature dimension reduction; however, reducing the feature vector numbers results in an unknown loss of information, which in turn produces instability in the system's performance.

Authors at [18] utilized a method to identify emotions in speech signals by break- ing them down into intrinsic mode functions. They measured randomness through entropy calculations. Different classes of emotions had distinct median values for the entropy features, making these features effective discriminators for emotion classifi- cation. However, not a single entropy measure yielded accurate classification for all emotions, and the accuracy relied on the chosen number of Intrinsic Mode Functions (IMFs). The paper identified advances in the method put forward to account for the gender and age bias in speech emotion recognition as the future area of research.

The authors at[19] presented a system that makes use of the Fractional Fourier Transform for adaptive time-frequency feature extraction. The Discrete Fractional Fourier Transform (DFrFT), a technique for adaptive time-frequency feature extrac-tion, is created by expanding the standard Discrete Fourier Transform (DFT) and its eigenvalues and eigenvectors. A Fractional Fourier transform, which is an adaptive fea-ture in the time-frequency domain, is retrieved by adjusting the optimal value of angle. The proposed feature extraction approach takes into account the temporal dimension, enabling the use of these features as inputs for deep networks like CNN and LSTM.

This makes it possible to use these networks to extract advanced characteristics for upcoming applications.

Comparatively, there are more SER experiments carried out in English, German, and French than there are in other languages, as observed by authors at[20]. Studies on the recognition of cross-lingual speech emotions have been conducted. In a cross-lingual investigation, employing the RAVDESS datasets, authors at [21] proposed an approach of blending a BiLSTM network with a deep CNN network. In the cross-lingual trial, transfer learning has produced weighted accuracy of 82.7% on the RAVDESS dataset. Authors at[22] have identified the parameters to increase the accuracy, efficiency, and security of emotion recognition systems such as: Compound emotions (e.g., sur- prisingly happy, happily surprised, fearfully sad, angrily sad); multi-modal approach combining body movements, voice, text or facial expressions to increase the efficiency of the system; cancelable biometrics to secure the database from breaching by applying different transformations on the original data.

By using neutral speech as a reference, authors at [23] present an intriguing approach to the feature extraction challenge for emotion recognition. Between an emo- tional speech and a neutral, non-emotional speech, they capture differences. They take a different technique than that often used in the state-of-the-art[22]. According to their report, their hierarchical binary decision tree-based approach is on par with or superior to the current prosody and spectral features.

SER classification can be conducted using two methods: (a) deep learning clas- sifiers and (b) conventional classifiers. SER systems often make use of a number of conventional classification algorithms. A new class input was foreseen by the learning method, which calls for labeled data that identifies the relevant samples and classes by roughly approximating the mapping function. The leftover data is used to evaluate the performance of the classifier after training. The Hidden Markov Model, Support Vector Machines, the Gaussian Mixture Model, and the Artificial Neural Network are a few examples of classic classifiers. Decision Trees, K-Nearest Neighbour, K-means, and Naive Bayes Classifiers are some further conventional classification techniques.
In Table 1 a comparative analysis of different databases, classification techniques, results, and possible directions for future studies are summarized.

3 Datasets

There are various databases available in the literature, as described in this section, along with their limitations.

EmoDB Database

The EmoDB database[35], a freely accessible emotional database, originates from the Institute of Communication Science at the Technical University in Berlin, Germany. The database includes recordings of 10 professional speakers (5 females and 5 males) and consists of 535 utterances. It covers seven emotions: sadness, happiness, anger, anxiety, boredom, disgust, and neutral. Initially, the recordings were sampled at the rate of 48 kHz and subsequently down sampled to 16 kHz.

Table 1 Literature survey on different databases, classification techniques, results, and possible areas of future studies

Publication	Dataset	Emotions considered	Solutions/Experiments	Results/Observations	Possible areas of future studies
Leila Kerkeni et al.2018[24]	Berlin EmoDB and Spanish Database.	Anger, Disgust, Fear, Joy, Neutral, Surprise, Sadness.	Use of Recurrent Neural Network (RNN), Multivariate linear regression (MLR), and Support vector machine (SVM) classifiers.	They explored different algorithms such as SVM, RNN using (MFCC), and modulation spectral features (MSFs) and observed that combined features achieved an accuracy of 90.5% using RNN.	Alternative interpretation of frequency and an alternative view of non-linear and non-stationary phenomena in SER.
Surekha Reddy Bandela et al.2019[25]	IEMOCAP, EmoDB.	Boredom, Disgust, Anger Anxiety/Fear, Happiness, Sadness ,and Neutral.	Semi-nonnegative matrix factorization (semi-NMF) with singular value decomposition (SVD) initialization for feature optimization.	The SER system's performance is influenced by both the database type employed and the classification model under consideration. The accuracy of classifying the SER system improves as more features are added, up to the optimal rank of semi-NMF. However, after reaching this point, the accuracy begins to decline, suggesting the presence of the curse of dimensionality.	Further enhancements can be made to the proposed SER system to achieve language independence through cross-corpus analysis.
Margaret	EMO-DB	Anger, Bore-	Application of the most pow-	Baseline approach achieved an average accu-	When the sampling frequency

Lech, et al.2020[26]		dom, Disgust Fear, Joy, Neutral, Sadness	erful pre-trained neural networks trained for image classification to the problem of SER.	racy of 82% when trained on the Berlin Emotional Speech (EMO-DB) data.	was lessened from the baseline of 168 kHz, it resulted in a decline in SER accuracy of approximately 3.3%. Privacy issues arising from transfer learning should be addressed by building a more robust privacy-preserving machine learning model.
ArijitDey et al.2020[27]	SAVEE and EmoDB.	Fear, Anger Disgust, Surprise, Anxiety, Happiness, Sadness, and Normal.	To increase the accuracy of the SER system, a hybrid feature selection (FS) method called Golden Ratio based Equilibrium Optimization (GREO) was proposed. This method combined the	The achieved accuracies on SAVEE and EmoDB were 97.31% and 98.46% respectively.	The approach involves utilizing clustering algorithms to form populations based on specific dataset properties, thereby offering a hybrid approach that combines various optimization algorithms.

			principles of Equilibrium Optimization (EO) and the Golden Ratio Optimization (GRO) algorithms to reduce the size of the feature vector.		
PalaniThanaraj Krishnan ·et al.2021 [28]	TESS (Toronto Emotional Speech Set)	Pleasant surprise, Anger Disgust, Sadness, Fear, Happiness, and Neutral.	The median values of entropy features vary across different emotion classes, indicating their potential as effective discriminators for emotion classification.	In order to accurately detect emotions from speech signals, it is essential to leverage the complementary information provided by all entropy measures. No individual entropy measure alone can achieve satisfactory classification accuracy for all emotions present in speech signals. The selection of the number of IMFs plays a vital role in determining the accuracy of classification.	To address gender and age bias in speech emotion recognition, further advancements can be made to the proposed method.

Publication	Dataset	Emotions con- sidered	Solutions/Experiments	Results/Observations	Possible areas of future stud- ies
Mohammad Mehedi et al.2019[29]	DEAP(A dataset for psychological signals-based emotional analysis)	Relaxed, Sad, Happy, Dis- gust and Neutral.	Deep Belief Network (DBN)	They introduced a novel feature fusion tech- nique combining PPG, Zemg, and EDA. The DBN and FGSVM(Fine Gaussian SupportVector Machine) architecture achieved an accuracy of 89.53%.	Additionally, the study sug- gests that employing an ensemble classifier technique along with the proposed technique can increase the generalizability and robust- ness of SER. In addition to text data and audio signals, further explo- ration of multiple modalities can be conducted by incorpo- rating video inputs.
Seunghyun Yoon et al.2019 [30]	IEMOCAP	Happy, angry Neutral and Sad	Dual Recurrent Neural Net- work (RNN)	Introducing a novel architecture called MDRE, which incorporates both text data and audio signals. The accuracy obtained by the proposed method ranges from 68.8% to 71.8%. There are possibilities for development to raise the method's accuracy.	
TM Wani et al.2020 [31]	SAVEE	Happy, Sad, Neutral and Angry	CNN and DSCNN(Deep Stride Convolutional Neural Network) The DSCNN frame- work was utilized, resulting in an accuracy of 87.8% and		

Z.Huijuan et al.2020 [32]	IEMOCAP	Angry, excited, Frustrated and excited ,Happy.	79.4% for DSCNN and CNN, respectively. CNN block with RNN atten-tion module.	A newly introduced hierarchical classification system operates in a top-down manner. Uti- lizing 3D-HMTL, an F1 score of 0.4673 is attained.	Performance enhancements can be made by incorporating fine-grained emotion classifi-cation.
Mingke Xu et.al2020[33]	IMOCAP	Happy, Sad, Excited and Neutral.	ACNN(Attention-based Con- volutional Neural Network)	They employed ACNN, a multi-head self- attention model, for Speech Emotion Recog- nition (SER)	To generate additional emo- tional training data, a Gen- erative Adversarial Network (GAN) can be employed.
D.Bharti et al.2020[34]	RAVDEES	Happiness, Joy, Angry and Sad.	ALO(Ant Lion Optimiza-tion), MVSM(Multi-class Support Vector Machine), and GFCC(Gammatone Fre-quency Cepstral Coefficient).	Impressively, their approach achieved a weighted accuracy (WA) of 76.18% and 76.36%, as well as an unweighted accuracy (UA) of 76%. For the classification process, a novel MVSM is utilized, and ALO+MSVM achieves an accuracy rate of 97%.	Efficient feature fusion can boost the performance of the multimodal system.

SAVEE Database

The recording of the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset[36] was undertaken as a necessary step toward the formation of an automated emotion recognition system. The dataset comprises recordings where four male actors express a variety of seven distinct emotions. In total, it includes 480 British English utter- ances. Sentences from the widely utilized TIMIT corpus were chosen and phonetically balanced to correspond to each specific emotion in a meticulous manner.

Crowd-sourced Emotional Multimodal Actors Dataset

CREMA-D is an emotional multimodal actor dataset[37] of 7,442 unique clips from 91 actors. The dataset includes clips from 43 female and 48 male actors, ages ranging 20 to 74, representing diverse races and ethnicities such as Asian, Hispanic, African American, Caucasian, and Unspecified. These clips capture the actors speaking 12sentences, each expressing one of the 6 emotions (Happy, Anger, Fear, Neutral, Disgust and Sad) and 4 emotion levels (High, Medium, Low, and Unspecified). Participants assessed both the emotion and emotion levels by evaluating the combined audiovisual presentation, as well as the video and audio independently.

Ryerson Audio-Visual Database of Emotional Speech andSong (Ravdess)

The Ryerson Audio-Visual Database of Emotional Speech[38] and Song (RAVDESS) contains 7536 recordings. There are three different modalities present inside these files, i.e., full audio and video, audio only, and video only, along with 2 vocal channels, i.e., song and speech. A single instance of this dataset, i.e., one audio file, contains the sound of one actor. This sound may represent the following eight categories: disgusted, surprised, fearful, happy, sad, angry, calm, and neutral. These emotional expressions are produced at two levels, i,e., regular and strong. It indicates that there can be the same emotions at two different scales, and these kinds of audio samples can be very useful to build a robust classifier that represents the emotions expressed in a real scenario.

The Interactive Emotional Dyadic Motion Capture (IEMOCAP)

The collection of this dataset took place at the University of Southern California, where 10 different professional people were used to generate speech of different emotions[39]. This dataset contains the speech of both males and females in each session. This dataset contains several different emotions, i.e., anger, sadness, happiness, excitement, fear, frustration, neutral and surprise. 5531 different speech samples are contained init.

Chinese Emotional Corpus(CASIA)

This dataset is the most popular dataset developed in the Chinese language that contains six different emotions[40]. These emotions include fear, happiness, anger,sadness, neutral, and surprise. Each emotion contains around 400 sentences spoken by four different professionals. The dataset was created by performing a data acquisition with a signal-to-noise(SNR) ratio of 35 fb.

Geneva Whispered Emotion Corpus (GeWEC))

It is an emotional database developed in the French language that contains four differ- ent emotions, i.e., fear, happiness, anger, and neutral[41]. The database offers emotions expressed in both normal and whispered states. It contains a total of 1280 recordings and was recorded using high-quality recording tools.

King Saud University Emotions (KSU Emotions) Corpus

It is an Arabic language dataset collected using the recording of 13 females and 10 males from Saudi Arabia, Yemen, and Syria[42]. The total recording is of 5 hr and 16 min. It contains emotions such as sadness, happiness, neutral, surprise, and anger.

4 Feature Extraction Techniques

There are several traditional and machine learning-based feature extraction techniques as described in this section.

Traditional Feature Extraction Techniques

Mel-frequency Cepstral Coefficient (MFCC)

MFCC has been found to be very effective in recent times for feature extraction [42]. MFCC works by assuming the human ear as the listener of audio since the human ear can identify the lower and higher frequencies; such frequency bandwidth can be used to extract different phonetic features.

During the computation of the MFCC feature, the speech signal is first split into different windows resulting in various numbers of frames. After framing the audio signal, Fast Fourier transformation (FFT) is applied to find the power spectrum of each frame. The Mel-scale is then used to apply filter bank processing on the power spectrum. The power spectrum is then transferred into the log domain resulting in the speech signal; after that, Dual Clutch Transmission (DCT) is applied to the speech signal. The calculation of MFCC is represented in equation 1.

$$C_n^{\wedge} = \sum_{n=1}^{k} (log S_k^{\wedge}) \cos n \left[k - \frac{k}{2} \right] \quad (1)$$

In equation 1, the variables k, Cn and Sk denote the number of mel coefficients, the final MFCC coefficients and the output obtained from the filter bank, respectively.

Zero-Crossing Rate

Zero-Crossing Rate (ZCR)[43] is the total number by which the signal changes value from positive to negative and vice-versa. It can also be described as the total amount of noise present in the signal. ZCR is calculated with the help of equation 2.

$$(i) = \frac{1}{2W_L} \sum_{n=1}^{W_L} |sgn[x(n)] - sgn[x(n-1)]|$$

Where, the function sgn(.) corresponds to the sign function, i.e.

$\text{sgn}[x_i(n)] = 1 x_i(n) <= 0 \quad 1 x_i(n) < 0$

From (2), it is observed that the value of *n*, i.e., noise is directly proportional tothe ZCR value.

Chroma stft

Chroma is an effective feature that represents sounds of speech where there are dif- ferent semitones that are divided and displayed into 12 different bins. These bins are *C, C#, D, D#, E, ..., B* , which are present in an audio signal. The chroma feature contains chroma values and a pitch height that represents the octave present inside it. These chroma features can extract the harmonic and melodic properties contained in a music. Chroma features represent the harmonic content for a very short period of an audio frame. In chromastft features, a feature vector is built by extracting the magni- tude spectrum with the help of Short Time Fourier Transformation (STFT), Chroma Energy Normalized (CENS), Constant-Q trans-formation (CQT), and other methods as discussed.

Root Mean Square (RMS)

To perceive the loudness of a sound, the audio signal is divided into different window frames, and the maximum amplitude in each window frame is found. The maximum amplitude can then be used to plot a graph. The plot with maximum amplitude can then be used for event detection. Additionally, it is much more robust to outliers than Amplitude Envelope (AE) discussed in [44]. This robustness leads to the detection of events such as a person speaking or crying more precisely.

RMS value is calculated with the help of equation 4.

$$RMS_t = \frac{\sum_{k=t.K}^{k} S(K)^2}{(t+1)(K-1)}$$

Proceeding through the window of an audio signal, the amplitudes within the windows are squared and then summed up. In equation 3, K defines the frame length. The summed-up value is divided by the frame length, and the square root is taken, resulting in an RMS value.

Mel Spectrogram

In Mel Spectrogram, each spectrum frame contains a short Fourier transformation [45]. This Fourier transformation ranges from the linear frequency scale to the loga- rithmic scale. If logarithmic transformation is used, then the Mel scale is a logarithmic transformation of these signal's frequency. The transformation equation is presented in equation 5, where the variable f represents the frequency, while m represents the corresponding mel scale.

$$m = 1127 * log(1 + \frac{f}{700}) \quad (4)$$

A filter bank is then applied to the transformed value, and an eigenvector and its eigenvalues are obtained. These eigenvalues represent a distribution of signal energy on Mel-scale frequency. Mel-Spectrogram uses 8 kHZ sampling to process each audio signal and is very effective in audio processing.

Machine Learning-based Feature Extraction Techniques

Autoencoder

Autoencoder is a type of neural network [46] that can be used in noise reduction, image compression, and effective feature extraction. Autoencoder consists of three layers, as shown in figure 3. The first layer is known as the input layer, the second layer is the hidden layer, and the output layer is the last layer. Autoencoder works by training the network such that the result generated from the output layer is equal to the input layer. Since the number of output nodes and input nodes in an autoencoder is identical, such a situation is possible.At every training round, the weight between the output and input nodes is adjusted by minimizing the reconstruction error. Since reconstruction error is minimized, autoencoders learn very robust features of input data such that from which output data can be generated.

Let us have a training dataset $\{x_i, y_i\}$, where $x_i \in R^N$ $y_i \in \{1, 2, 3, \quad , N\}$ represent where y_i different classes and x_i is the n-dimensional feature vector. The encoding is done using equation 6, where, in equation 6, $h(x)_i$ is the hidden represen- tation, f represents the activation function, w and b is the weight and bias associated with it. Hidden representation contains the feature associated with the input data. It tries to construct the best features with the reduced dimension by minimizing the loss function explained in equation 8. In equation 8, we optimize the parameter θ with the help of stochastic gradient descent(SGD) algorithm.

$$h(x)_i = f(w(x) + b) \quad (4)$$

Autoencoder tries to reconstruct the data back with the help of generated feature vectors, i.e., $h(x)_i$. Equation 6 explains this reconstruction process, which is also known as decoding.

$$x_i = g(w_i h(x)_i + b_i)$$

$$L(x_i, x, \theta) = ||x_i - x|^2$$

There are several different kinds of autoencoders available for feature extraction, which are discussed in further sections.

Denoising Autoencoder(DAE)

In this type of autoencoder, we add noise to the training data, such as Gaussian noise and Laplace noise [47]. The addition of noise makes the input data corrupted.

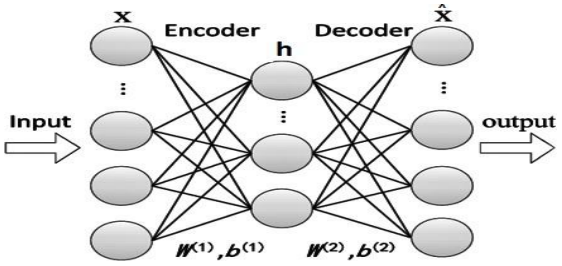


Fig. 3 Autoencoder Network.

The corrupted data is then used to construct hidden features that are more robust in nature. The corrupted input data don't let

autoencoders replicate the original data, and hence more robust features with lower dimensional features can be extracted.

Sparse Autoencoder(SAE)

In this type of autoencoder[48], we have a sparsity enforcer that helps in reducing error while reconstructing the input. Here sparsity enforcer helps in activating fewer neurons in hidden layers so that during training, a hidden representation of a feature can be created that is invariant to changes and linearly separable.

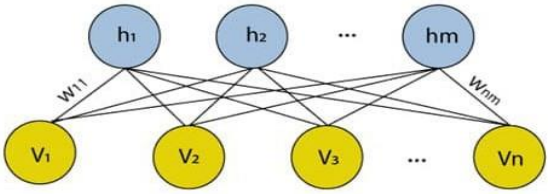
Contractive Autoencoder(CAE)

In this type of autoencoder, a penalty term[49] is added along with the cost function. This penalty term is like the addition of noise in a denoising autoencoder that brings regularization while reconstructing the input data. This regularization helps in learning robust features.

Restricted Boltzmann Machine

Restricted Boltzmann Machine(RBM) is a two-layer generative network that consists of a hidden and visible layer such that bipartite graphs can be created for the develop- ment of algorithms. For effective feature extraction, several boltzmann machines are stacked together from visible to hidden layers, which is shown in figure 4. The top layer is then embedded into a machine learning classifier such that generated features can be fed into the classifier. Two types of boltzmann machines have been for extracting features for speech emotion recognition which are described in sections 4.2.6 and 4.2.7.

Fig. 4 Restricted Boltzmann Machine.



Deep Boltzman Machine

Deep Boltzmann Machine (DBM) is created by stacking numerous hidden layers where parameters are randomly selected along with stochastic units. The learning of features is done in a hierarchical manner. To accomplish this, DBM utilizes a layer-wise pre- training technique based on Markov random field, allowing for the pre-training of unlabeled data. And providing feedback through a bottom-up approach.

Deep Belief Network

Deep belief networks (DBNs) are a type of artificial neural network that can be used to extract features from data. These networks are designed to learn hierarchical rep- resentations of the input data, allowing them to extract increasingly complex and abstract features as they move through the layers of the network. DBNs are particu- larly suitable for tasks such as image recognition and speech processing, as they can identify features such as corners, edges and contours in visual data or phonemes and syllables in audio data.

Table 2 Handcrafted Features Vs Machine-generated Features

Characteristics	Handcrafted Features	Machine-generated Features
Feature Extraction Process	Feature extraction process is tedious and cannot be guaranteed that the most reliable features can be obtained, which can further decrease the accuracy of a classifier.	Feature extraction process is done by learning numerous patterns, and hence more tedious and more robust fea- tures can be obtained.
Data Preprocessing		
Feature Learning	Noise in the features is automatically handled and requires less processing, like normalization and outlier removal.	Human-crafted features bring expensive noise and
Model Compilation		

	Feature learning becomes more effective and less tedious as they are learned from unlabelled data. It needs a massive amount of data and a longer execution time.	require massive processing. Lots of labeled data are needed, which requires dimen- sionality reduction of features that later on are hard to gen- eralize by classifiers. The execution time is short- ened, and less training data is required.

Comaprison of Handcrafted Features andMachine-generated Features

In table 2, we have done a comparison between handcrafted features and machine generated features on the basis of four key aspects.

5 Machine Learning-based Classifier

Several types of classifiers are found in the literature. The following section describes some commonly used classifiers for speech emotion recognition.

Support Vector Machine (SVM)

Support Vector Machines (SVM) is a widely recognized method for emotion classifica- tion. SVM is commonly employed for tasks involving classification and regression [50]. An N-dimensional hyperplane is created during the categorization process in order to efficiently divide the data into several groups. The categorization in the input feature space of the dataset is achieved by employing a separation surface that can be either linear or nonlinear in nature. The underlying principle of SVM is to achieve optimal classification in a novel feature space by initially converting the original input set into a higher-dimensional feature space through the utilization of a kernel function.

Multilayer Perceptron(MLP)

Speech emotion identification may be done using Multilayer Perceptron (MLP), a form of artificial neural network [50]. It has numerous layers of nodes and is a feedforward neural network. Every node within the network applies a non-linear transformation to the input it receives from the preceding layer and the output of the final layer conveys the inferred emotion.

First, characteristics must be extracted from the speech signal before we can apply MLP for speech emotion identification. Pitch, energy, and formant frequencies are examples of characteristics that are often employed. The MLP then receives these characteristics as input.

The MLP is trained on a dataset of speech samples labeled with their correspond- ing emotions. The MLP modifies its weights during training in order to lessen the difference between the predicted emotion and the actual emotion label. Once trained,the MLP can be used to predict the emotion of new speech samples.

Convolutional Neural Network (CNN)

Shift invariant artificial neural networks (SIANNs) or convolutional neural networks (CNNs) are particular varieties of neural networks with several hidden filters or regions, each responding to different characteristics [51] contained in the input signal. The study of Hubel and Wiesel from 1968, which defined the visual cerebral cortex as a spatially specialized structure where each area responds to certain characteristics of the input signal, served as an inspiration for the construction of these

networks. The quantity of storage required at the time of development rises as a result of CNNs' ability to learn features from tiny oscillations and distortions as well as from high- dimensional input data. As a result, a layer of convolution is typically present in CNNs, followed by a down-sampling method. The weights of the individual filter banks in the convolution layer are modified via frequent back-propagation.

Recurrent Neural Network (RNN)

Recurrent Neural Networks[52] are able to learn from and respond to temporal events without changing the slowly shaped weights. For applications where timing is a crucial component, such as speech processing, music composition, and video explanation, this functionality can be useful. The magnitude of the weights will determine whether the error signals flow backward in time or disappear when they are trained through backpropagation through time. As a result, the network will either produce oscillating weights or train and converge slowly.

6 CHALLENGES IN THE FIELD OF SER

Evaluating the speech preprocessing tasks and handlingthe class imbalance:

Comprehensive experimental evaluation is necessary to assess the impact of various speech pre-processing techniques (such as standardization and normalization) on the speed and accuracy of deep learning (DL) methods [54]. Additionally, the presence of imbalanced classes in speech databases has a negative effect on how well DL algorithms for SER perform when classifying data. Consequently, it is essential to focus on balanc- ing the database to address this issue. Consequently, there is a pronounced necessity for extensive research to devise algorithms that can successfully address the issue of class imbalance in speech datasets. By doing so, the effectiveness of DL approaches for SER can be substantially elevated.

Transfer learning for SER using deep learning techniques :

Pre-trained deep neural networks with the ability to retrain on fresh datasets are included in transfer learning. GoogleNet, AlexNet, LeNet, and Petri Net are a few ofthe well-known transfer learning designs [55]. The pre-trained architectures have been trained on diverse datasets, encompassing images of objects and animals. Although transfer learning enables rapid adaptation, it might not produce better classification outcomes on recent datasets. The classification accuracy in SER datasets may be low as a result of these transfer learning models Therefore, there is a need for research to develop novel transfer learning architectures specifically tailored for speech and speech emotion datasets. These architectures should aim to provide robust SER models, facil- itate the reuse of trained neural networks in different contexts, and reduce the training time required for DL methods.

Constructing multilingual speech emotion databases fordeep learning techniques:

To effectively evaluate DL techniques, it is necessary to have access to large speech-emotion databases [56]. However, many studies rely on single-language databases that have a limited number of emotional utterances. Therefore, considerable research efforts can be dedicated to the creation of extensive multilingual speech-emotion databases.

Improving the classification accuracy of deep learningmethods through feature fusion:

In studies, deep neural networks were trained to produce good results using the feature vector of various auditory representations. On acoustic characteristics, deep neural networks performed well, but the combination of temporal and frequency domain information may further enhance deep neural networks.

Active deep learning (ADL) for SER :

Deep neural network-based Active Deep

Learning (ADL) offers two significant advantages: scalability and robustness, owing to its ability to handle large volumes of data. This makes it particularly suitable and advantageous for a vast number of training instances. In the context of Speech Emotion Recognition (SER), a substantial amount of speaker utterances is required. However, labeling all these utterances, typically with the speaker ID, is a time-consuming task. In these circumstances, using ADL to train deep neural networks for SER can significantly minimize the amount of manual effort.

Fusion of automatic speaker recognition and SER systems:

In the future, it will be possible to recognize speakers as well as the speakers' emo- tions, such as happiness, sadness, and rage. Such a categorization system, built on contemporary deep neural network-based techniques, will be an enhancement of SER with possible applications like real-time classification of YouTube videos based on the speaker's voice tone.

Multilingual SER using deep learning techniques:

The majority of researchers used single speech databases to evaluate the proposed strategies in their initial research [57]. A few research compared the effectiveness of the suggested technique on various speech databases using different speech databases. There is therefore room for study into categorizing utterances of various languages using deep neural network-based methods. Additionally, DL approaches learn various ways to express the same feeling in various languages. More study is necessary in order to train DL systems using multilingual voice databases as a result.

Contextual information in speech emotion recognitionsystems:

Exploring the spontaneous behavior of humans has given rise to numerous research challenges, ranging from techniques for collecting databases to the incorporation of various features, such as lexical information alongside prosodic features. Though it is rarely acknowledged, contextual information plays an important part in speech emotion detection. Given that it is obvious that the environment has a significant impact on how emotions are perceived, investigating the significance of contextual information may be a useful study addition.

7 CONCLUSION

Deep Learning has proven to be a highly effective technique for speech emotion feature extraction and classification. Through the use of deep neural networks, researchers have been able to accurately identify and classify emotions in speech data. This has numerous applications, including in areas such as healthcare, education, and enter- tainment. One of the main advantages of Deep Learning for speech emotion feature extraction is its ability to learn complex patterns and relationships within the data. Unlike traditional machine learning techniques, which rely on hand-crafted features, Deep Learning can automatically extract features from raw speech data. This not only simplifies the process of feature extraction, but also results in more accurate classifi- cation. Additionally, Deep Learning models can be trained on large datasets, allowing for improved performance and generalization. This makes it possible to develop models that can accurately classify emotions across different speakers, languages, and cul- tural backgrounds. In the future, due to the massive growth in speech data, the need for deep learning algorithms like unsupervised learning can be of great potential to automate feature extraction and data labeling.

Statements and Declarations Ethics approval Not applicable **Consent to participate** Not applicable

Competing interests The authors declare that they have no competing interests.

Availability of supporting data Not applicable

Funding Not applicable

Authors' contributions Himashri Deka-Conceptualization, Writing (Original18 Draft), Vikas Mittal-Writing (reviewing), Supervision, Validation.

Acknowledgments Not applicable

References

[1] Lieberman, Philip. "The evolution of human speech: Its anatomical and neural bases." Current anthropology 48.1 (2007): 39-66.

[2] Singh, Youddha Beer, and Shivani Goel. "A systematic literature review of speech emotion recognition approaches." Neurocomputing (2022).

[3] Schuller, Björn W. "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends." Communications of the ACM 61.5 (2018): 90-99.

[4] Nogueiras, Albino, et al. "Speech emotion recognition using hidden Markov mod- els." Seventh European conference on speech communication and technology.2001.

[5] Han, Kun, Dong Yu, and Ivan Tashev. "Speech emotion recognition using deep neural network and extreme learning machine." Interspeech 2014. 2014.

[6] LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436-444.

[7] Jain, Saachi, et al. "A data-based perspective on transfer learning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[8] Atmaja, Bagus Tris, Akira Sasou, and Masato Akagi. "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion." Speech Communication (2022).

[9] Singh, Amarjeet. "Speech emotion recognition using enhanced cat swarm opti- mization algorithm." International Journal of Information Technology (IJIT) 6.5 (2020).

[10] Flower, T. Mary Little, and T. Jaya. "Speech emotion recognition using Ramanu- jan Fourier Transform." Applied Acoustics 201 (2022): 109133.

[11] Globerson, Eitan, et al. "Psychoacoustic abilities as predictors of vocal emotion recognition." Attention, Perception, Psychophysics 75 (2013): 1799-1810.

[12] Zhang, Hua, et al. "Pre-trained deep convolution neural network model with attention for speech emotion recognition." Frontiers in Physiology 12 (2021): 643202.

[13] Mustaqeem, and Soonil Kwon. "A CNN-assisted enhanced audio signal processing for speech emotion recognition." Sensors 20.1 (2019): 183.

[14] Scheidwasser-Clow, Neil, et al. "SERAB: A multi-lingual benchmark for speech emotion recognition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[15] Aggarwal, Apeksha, et al. "Two-way feature extraction for speech emotion recognition using deep learning." Sensors22.6 (2022): 2378.

[16] Morais, Edmilson, et al. "Speech emotion recognition using self-supervised fea- tures." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[17] Pandey, Sandeep Kumar, Hanumant Singh Shekhawat, and S. R. M. Prasanna. "Attention gated tensor neural network architectures for speech emotion recogni- tion." Biomedical Signal Processing and Control 71 (2022): 103173.

[18] Latif, Siddique, et al. "Self supervised adversarial domain adaptation for cross- corpus and cross-language speech emotion recognition." IEEE Transactions on Affective Computing (2022).

[19] Maji, Bubai, and Monorama Swain. "Advanced fusion-based speech emotion recognition system using a dual-attention mechanism with conv-caps and bi-gru features." Electronics11.9 (2022): 1328.

[20] Gat, Itai, et al. "Speaker normalization for self-supervised speech emotion recog- nition." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[21] Rudd, David Hason, Huan Huo, and Guandong Xu. "Leveraged mel spectro-grams using harmonic and percussive components in speech emotion recognition." Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II. Cham: Springer International Publishing, 2022.

[22] Abdelhamid, Abdelaziz A., et al. "Robust speech emotion recognition using CNN+ LSTM based on stochastic fractal search optimization algorithm." IEEE Access 10 (2022): 49265-49284.

[23] Bhangale, Kishor, and K. Mohanaprasad. "Speech emotion recognition using mel frequency log spectrogram and deep convolutional neural network." Futuristic Communication and Network Technologies: Select Proceedings of VICFCNT 2020.Springer Singapore, 2022.

[24] Kerkeni, Leila, et al. "Speech Emotion Recognition: Methods and Cases Study."ICAART (2) 20 (2018).

[25] Bandela, Surekha Reddy, and K. U. M. A. R. T KISHORE. "Speech emotion recognition using semi-NMF feature optimization." Turkish Journal of Electrical Engineering Computer Sciences 27.5 (2019): 3741-3757.

[26] Lech, Margaret, et al. "Real-time speech emotion recognition using a pre-trained image classification network: Effects of bandwidth reduction and companding." Frontiers in Computer Science 2 (2020): 14.

[27] Dey, Arijit, et al. "A hybrid meta-heuristic feature selection method using golden ratio and equilibrium optimization algorithms for speech emotion recognition." IEEE Access 8 (2020): 200953-200970.

[28] Krishnan, PalaniThanaraj, Alex Noel Joseph Raj, and VijayarajanRajangam. "Emotion classification from speech signal based on empirical mode decomposition and non-linear features." Complex Intelligent Systems 7.4 (2021): 1919-1934.

[29] Hassan, Mohammad Mehedi, Md Golam Rabiul Alam, Md Zia Uddin, Shamsul Huda, Ahmad Almogren, and Giancarlo Fortino. "Human emotion recognition using deep belief network architecture." Information Fusion 51 (2019): 10-18.

[30] Yoon, Seunghyun, Seokhyun Byun, and Kyomin Jung. "Multimodal speech emo-tion recognition using audio and text." In 2018 IEEE Spoken Language Technology Workshop (SLT), pp. 112-118. IEEE, 2018.

[31] Xu, Haiyang, et al. "Learning alignment for multimodal emotion recognition from speech." arXiv preprint arXiv:1909.05645 (2019).

[32] Huijuan, Zhao, Ye Ning, and Wang Ruchuan. "Coarse-to-fine speech emotion recognition based on multi-task learning." Journal of Signal Processing Systems 93 (2021): 299-308.

[33] Xu, Mingke, Fan Zhang, and Samee U. Khan. "Improve accuracy of speech emo-tion recognition with attention head fusion." In 2020 10th annual computing and communication workshop and conference (CCWC), pp. 1058-1064. IEEE, 2020.

[34] Bharti, Deepak, and Poonam Kukana. "A hybrid machine learning model for emo-tion recognition from speech signals." In 2020

international conference on smart electronics and communication (ICOSEC), pp. 491-496. IEEE, 2020.

[35] Burkhardt, Felix, et al. "A database of German emotional speech." Interspeech. Vol. 5. 2005.

[36] Jackson, Philip, and SJUoSG Haq. "Surrey audio-visual expressed emotion (savee) database." University of Surrey: Guildford, UK (2014).

[37] Cao, Houwei, et al. "Crema-d: Crowd-sourced emotional multimodal actors dataset." IEEE transactions on affective computing 5.4 (2014): 377-390.

[38] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal setof facial and vocal expressions in North American English." PloS one 13.5 (2018):e0196391.

[39] Tripathi, Samarth, Sarthak Tripathi, and Homayoon Beigi. "Multi-modal emo- tion recognition on iemocap dataset using deep learning." arXiv preprint arXiv:1804.05788 (2018).

[40] Liu, Zhen-Tao, et al. "Speech emotion recognition based on an improved brain emotion learning model." Neurocomputing 309 (2018): 145-156.

[41] Deng, Jun, et al. "Fisher kernels on phase-based features for speech emotion recognition." Dialogues with social robots: Enablements, analyses, and evaluation (2017): 195-203.

[42] Mohamed, Omar, and Salah A. Aly. "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset." arXiv preprint arXiv:2110.04425 (2021).

[43] Patnaik, Suprava. "Speech emotion recognition by using complex MFCC and deep sequential model." Multimedia Tools and Applications 82.8 (2023): 11897-11922

[44] Roppel, Anian, Nils Unger, and Marliese Uhrig-Homburg. "Zero Crossing." (2023).

[45] Hodson, Timothy O. "Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not." Geoscientific Model Development 15.14 (2022): 5481-5487.

[46] Kaneko, Takuhiro, et al. "iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[47] Chen, Xiaokang, et al. "Context autoencoder for self-supervised representation learning." arXiv preprint arXiv:2202.03026 (2022).

[48] Kaur, Navdeep, and Parminder Singh. "Modelling of Speech Parameters of Pun- jabi by Pre-trained Deep Neural Network Using Stacked Denoising Autoencoders." ACM Transactions on Asian and Low-Resource Language Information Processing 22.3 (2023): 1-17.

[49] Sivaraman, Vishal Balaji, et al. "Sparse Autoencoder-Based Speech Emotion Recognition." Communication and Intelligent Systems: Proceedings of ICCIS 2021. Singapore: Springer Nature Singapore, 2022. 533-544.

[50] Oyedotun, Oyebade K., and Djamila Aouada. "A Closer Look at Autoencoders for Unsupervised Anomaly Detection." ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[51] Khattak, Muhammad Irfan, et al. "ERBM-SE: Extended Restricted Boltzmann Machine for Multi-Objective Single-Channel Speech Enhancement." IJIMAI 7.4 (2022):

185-195.

[52] Chen, Xiajin. ”Design of Political Online Teaching Based on Artificial Speech Recognition and Deep Learning.” Computational Intelligence and Neuroscience 2022 (2022).

[53] Senthilkumar, N., et al. ”Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks.” Materials Today: Proceedings 57 (2022): 2180-2184.

[54] Jain, Kabir, et al. ”Investigation Using MLP-SVM-PCA Classifiers on Speech Emotion Recognition.” 2022 IEEE 9th Uttar Pradesh Section International Con- ference on Electrical, Electronics and Computer Engineering (UPCON). IEEE, 2022.

[55] Femi, D., and S. Thylashri. ”Human Voice Emotion Recognition Using Multilayer Perceptron.” 2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). IEEE, 2022.

[56] Falahzadeh, Mohammad Reza, et al. ”Deep convolutional neural network and gray wolf optimization algorithm for speech emotion recognition.” Circuits, Systems, and Signal Processing 42.1 (2023): 449-492.

[57] Goel, Dev Priya, et al. ”Towards an efficient backbone for preserving features in speech emotion recognition: deep-shallow convolution with recurrent neural network.” Neural Computing and Applications 35.3 (2023): 2457-2469.

[58] Chan, Jireh Yi-Le, et al. ”State of the art: a review of sentiment analysis based on sequential transfer learning.” Artificial Intelligence Review 56.1 (2023): 749-780.

[59] Wang, Chunyi, et al. ”Speech emotion recognition based on multi-feature and multi-lingual fusion.” Multimedia Tools and Applications 81.4 (2022): 4897-4907.

[60] Al-onazi, Badriyya B., et al. ”Transformer-based multilingual speech emotion recognition using data augmentation and feature fusion.” Applied Sciences 12.18 (2022): 9188.