IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 25, 2021, accepted March 10, 2021, date of publication March 22, 2021, date of current version April 1, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3068045

A Comprehensive Review of Speech Emotion Recognition Systems

TAIBA MAJID WANI[1], TEDDY SURYA GUNAWAN[1,3], (Senior Member, IEEE), SYED ASIF AHMAD QADRI[1], MIRA KARTIWI[2], (Member, IEEE), AND ELIATHAMBY AMBIKAIRAJAH[3], (Senior Member, IEEE)

[1]Department of Electrical and Computer Engineering, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia
[2]Department of Information Systems, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia
[3]School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia

Corresponding author: Teddy Surya Gunawan (tsgunawan@iium.edu.my)

This work was supported by the Malaysian Ministry of Education through Fundamental Research Grant, FRGS19-076-0684, under Grant FRGS/1/2018/ICT02/UIAM/02/4.

ABSTRACT During the last decade, Speech Emotion Recognition (SER) has emerged as an integral component within Human-computer Interaction (HCI) and other high-end speech processing systems. Generally, an SER system targets the speaker's existence of varied emotions by extracting and classifying the prominent features from a preprocessed speech signal. However, the way humans and machines recognize and correlate emotional aspects of speech signals are quite contrasting quantitatively and qualitatively, which present enormous difficulties in blending knowledge from interdisciplinary fields, particularly speech emotion recognition, applied psychology, and human-computer interface. The paper carefully identifies and synthesizes recent relevant literature related to the SER systems' varied design components/methodologies, thereby providing readers with a state-of-the-art understanding of the hot research topic. Furthermore, while scrutinizing the current state of understanding on SER systems, the research gap's prominence has been sketched out for consideration and analysis by other related researchers, institutions, and regulatory bodies.

INDEX TERMS Speech emotion recognition, database, preprocessing, feature extraction, classifier.

I. INTRODUCTION

We humans have a unique ability to convey ourselves through speech. These days alternative communication methods like text messages and emails are available. Further, instant messages are aided by emojis that have paved the way for visual communication in this digital world. However, speech is still the most significant part of human culture and is data-rich. Both paralinguistic and linguistic information is contained in the speech.

Classical automatic speech recognition systems focused less on some of the essential paralinguistic information passed on by speech like gender, personality, emotion, aim, and state of mind [1]. The human mind utilizes all phonetic and paralinguistic data to comprehend the utterances' hidden importance and has efficacious correspondence [2]. The superiority of communication gets badly affected if there is any meagreness in the cognizance of paralinguistic features. There have been some arguments regarding children who can not comprehend the speaker's emotional conditions evolve substandard social skills. In certain instances, they manifest psychopathological manifestations [3], which accentuates the significance of perceiving speech's emotional conditions leading to ineffective communication. Therefore, creating coherent and human-like communication machines that comprehend paralinguistic data, for example, emotion, is essential [4].

Emotion recognition has been the subject of exploration for quite a long time. The fundamental structure of research in emotion recognition was formed by detecting emotions from facial expressions [5]. Emotion recognition from speech signals has been studied to a great extent during recent times. In human-computer interaction, emotions play an essential role [6]. In recent times, speech emotion recognition (SER), which expects to investigate the emotion states through speech signals, has been drawing increasing consideration. Nevertheless, SER remains a challenging task, with the question of how to extract effective emotional features.

The associate editor coordinating the review of this manuscript and approving it for publication was Joanna Kolodziej.

A classification of methodologies that process and at the same time characterize speech signals to identify emotions embedded in them is an SER system. An SER system needs a classifier, a supervised learning construct, programmed to perceive any emotions in new speech signals. [7]. A supervised system like that introduces the need for labeled data with emotions embedded in it. Before any processing can be done on the data to extract the features, it needs preprocessing. For this reason, the sampling rate across all the databases should be consistent. The classification process essentially requires features. They help reduce raw data into the most critical characteristics only, regardless of whether it suffices to utilize acoustic features for displaying emotions or if it is mandatory to cooperate with different kinds of features like linguistic, facial features, or speech information.

Classifiers' performance can be said to depend mainly on the techniques of feature extraction and those features that are viewed as salient for a particular emotion [8]. If additional features can be consolidated from different modalities, for example, linguistic and visual, it can strengthen the classifiers. However, this relies on the significance and accessibility. These features are then permitted to pass to the classification system with a broad scope of classifiers at its disposal. All have been analyzed to classify emotions according to their acoustic correlation in speech utterances from numerous machine learning algorithms. Linear discriminant classifiers, Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), k-nearest neighborhood (kNN) classifiers, Support Vector Machines (SVM), decision tree, and artificial neural networks (ANN) are a few models that have been generally used to classify emotions dependent on their acoustic features of intrigue [9]. The feature extraction techniques used by these classifiers are the determining factor for the performance of these classifiers. To date, numerous acoustic features and classifiers have been put through experimentation to test their credibility, but the accuracy still needs to be improved. In recent times, deep learning classifiers have become common such as Deep Belief Networks, Deep Neural Network, Deep Boltzmann Machine, Convolution Neural Network, Recurrent Neural Network, and Long Short-Term Memory.

The rest of the paper is organized as follows. Section 2 discusses the SER system with different databases, speech processing, feature extraction techniques, and different classification methods used for SER. Some of the recent works using different databases and different classifiers and their contribution and future direction are shown in Section 3. Different challenges faced by SER are discussed in Section 4, while Section 5 concludes the paper.

II. SPEECH EMOTION RECOGNITION SYSTEM

The development of machines that communicate with humans through interpreting speech leads us down the way for the development of systems designed with human-like intelligence. In Artificial Intelligence, the automatic speech recognition field has been actively involved in generating the machines that communicate with human beings via speech. Human speech is the most common and expedient way of communication, and understanding speech is one of the complex mechanisms that the human brain performs. From a speech signal, numerous amounts of information can be gathered like gender, words, dialect, emotion, and age that could be utilized for various applications. In speech processing, one of the most arduous tasks for the researchers is speech emotion recognition. SER is of most interest while studying human-computer recognition. It implies that the system must understand the user's emotions, which will define the system's actions accordingly. Various tasks such as speech to text conversion, feature extraction, feature selection, and classification of those features to identify the emotions must be performed by a well-developed framework that includes all these modules [10]. The task of classification of features is yet another challenging work, and it involves the training of various emotional models to perform the classification appropriately.

Now comes the second aspect of emotional speech recognition, the database used for training models. It involves selecting only the features that happen to be salient to depict the emotions accurately. Merging all the above modules in the desired way provides us with an application that can recognize a user's emotions and further provide it as an input to the system to respond appropriately. When we take a superior view, it may be isolated into a few fields, as depicted in Figure 1. The enhancement of the classification process can be attributed to a better understanding of emotions.

Numerous approaches are present to display emotions. Still, it is an open issue. In any case, the dimensional and discrete models are ordinarily utilized. Subsequently, emotional models have been reviewed.

A. EMOTIONAL MODELS

Human beings are known for an uncountable set of emotions that can be expressed in multidimensional ways. Some of the commonly used ways are writing, speech, facial expression, body language, and gesture [11]. Now, since these emotions pertain to a human's different aspects, they can likewise be arranged with various emotion models. If we intend to determine and further analyze emotions from any content, it needs to be studied using an appropriate emotion model. For the problem, such an emotion model should determine the applicable set of emotions correctly.

The grouping of human emotions is different from a psychological point of view based on emotion intensity, emotion type, and various parameters boundaries, which could be consolidated and acknowledged into emotion models [10]. When categorized and defined based on some scores, ranks, or dimensions, various human emotions give us the emotion models. These models define various emotions as understood from the name itself, based on duration, behavioral impact, synchronization, rapidity of change, intensity, appraisal elicitation, and event focus.

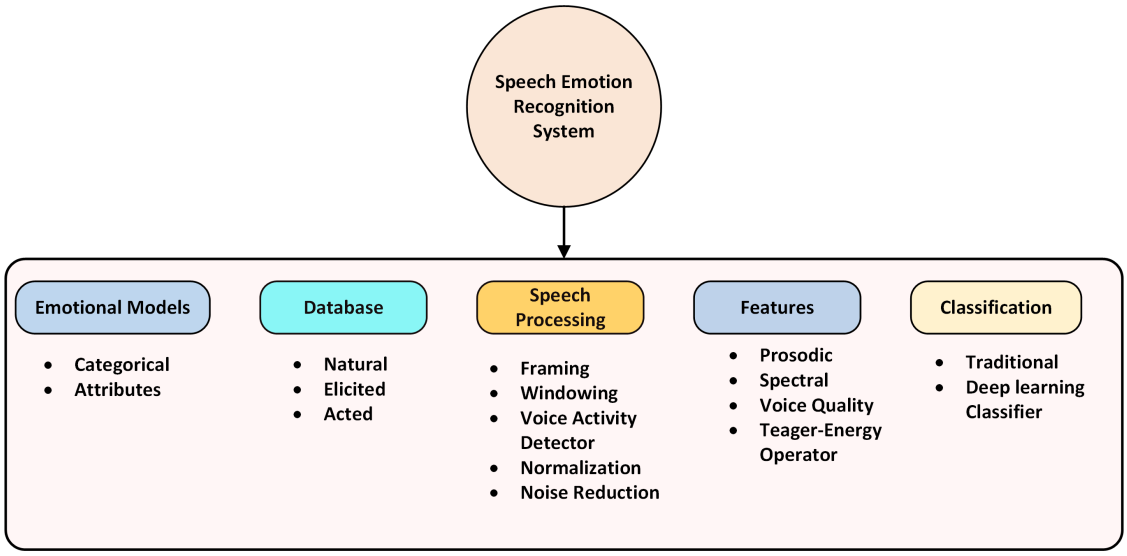

FIGURE 1. General SER system.

We can sum up that in light of various emotion hypotheses, existing emotion models can be separated into two categories: Attributes and Categorical [12]. Attribute emotion models describe some dimensions with individual variables, determine emotions as per given dimensions, and categorical emotion models characterize a rundown of classes of disjunctive emotions from each other.

B. DATABASES

Facial emotions are said to be a universal feature by various researchers. Nevertheless, the other emotions that include vocal signatures show variations in features like the emotion anger. If we consider the fundamental frequency values, it is known that anger typically has a high fundamental frequency compared to the low fundamental frequency of sadness. One simple reason for this variation is the difference between the actual data and the simulated data [13]. However, other causes of variations are differences in culture, language, gender, and situations. Since emotions are a set of complex input variables, any evaluation that they are put through must have a set degree of naturalness so that the assessed performance is closer to reality [14].

If the quality of the database is compromised, the conclusions are deemed to be incorrect. Also, the design of a database is still significant, given the importance of the classification task. Moreover, the database's design is critically important to the classification task being considered to evaluate the frameworks' implementation. The purposes and strategies for gathering speech corpora differ profoundly as per the inspiration driving the improvement of speech frameworks. Speech corpora utilized for creating emotional speech frameworks can be organized into three kinds, specifically.

1) SPONTANEOUS SPEECH

The most claimed of the databases is spontaneous speech, which contains the most natural and authentic emotions. In this case, hidden mechanisms are used to record the speaker's exact emotional conditions, allowing the subject to react naturally, unaware that his reactions are being monitored [15]. Nevertheless, the technology used for the data collection of emotions has never been an easy task. At times, the emotional expressions tend to be spontaneous. They are collected as multimodal samples. Some ways of collecting these are from TV chat shows, interviews, and other environments of such kinds.

2) ACTED SPEECH

Creating the acted speech database is not faced by those difficulties that the natural database faces, making it easier to control. As stated by Cao *et al.*, acted speech is merely conforming the type of emotion [15]. Some collections of acted speech might also be composed of recordings with professional actors or mature artists.

3) ELICITED SPEECH

Elicited speech uses the procedure of evoking a specific emotion by inducing certain emotions [16]. It is quite possible to put a subject into such a situation that evokes a certain kind of emotion. This speech is then recorded. However, the introduced emotions face a general problem of being significantly mild. However, the induction method gives some control over the stimulus. There are various datasets utilized for emotion recognition. A portion of the conspicuous datasets is summarized in Table 1.

TABLE 1. Prominent databases in speech emotion recognition system.

NO.	DATABASE	LANGUAGE	TYPE	SIZE	EMOTIONS
1.	Berlin Emotional Database [EMO-DB] [18]	German	Acted	7 emotions, 10 utterances, 10 speakers (5 male and 5 female)	Neural, anger, sadness, fear, boredom, happiness, disgust, boredom.
2.	Surrey Audio-Visual Expressed Emotion (SAVEE)[19]	English	Acted	7 emotions, 4 speakers (male), 120 utterances	Surprise, anger, fear, disgust, sadness, neutral, happiness.
3.	RECOLA Speech Database [20]	French	Natural	7 hours of speech, 46 speakers (27 females, 19 males)	5 social behaviors (engagement, performance, agreement, rapport, dominance); valence and arousal.
4.	SAMAINE Database [21]	English Greek Hebrew	Natural	959 conversation, 150 speakers.	power, valence, expectation, activation, overall emotional intensity.
5.	eNTERFACE'05 Audio-Visual Emotion Database [22]	English	Elicited	1116 video sequences, 8 females, 34 males, a total of 42 speakers, from 14 different countries.	Surprise, disgust, happiness, fear, anger, sadness.
6.	Interactive Emotional Motion Capture (USC-IEMOCAP)[23]	English	Elicited	Five sessions where each session includes the conversation between two people (one male and one female) and its corresponding labelled speech text	Anger, happiness, sadness, frustration, neutral
7.	FAU Aibo Emotion Corpus [24]	German	Natural	51 children talking to robot dog Aibo, 9 hours of speech	Bored, joyful, helpless, touchy, anger, reprimanding, emphatic, surprised, neutral, motherese, rest.
8.	BAUM-1 Speech Database [25]	Turkish	Acted and Natural	1222 spontaneous video clip, 288 acted, 31 speakers (13female, 18 male)	Anger, surprise, sadness, disgust, contempt, fear, concentration, bothered, being thoughtful, unsure, happiness, boredom, interest.
9.	Oriya Emotion Speech Dataset [26]	Odia/Oriya	Elicited	35 speakers (12 female and 23 male) recoded the text fragments of Oriya drama scripts.	Astonish, sadness, fear, anger, happiness, neutral.
10.	Persian Emotion Speech Dataset [27]	Persian	Simulated	33 native speakers (15 females and 18 males) recorded 748 utterances from Persian Drama Radio Emotional Corpus (PDREC)	Happiness, anger, sadness, surprise, fear, boredom, neutral, disgust.
11.	Assamese Emotion Speech Dataset [28]	Assamese	Simulated	30 students and faculty members (3 males and 3 females per language) recorded 140 utterances of 5 native languages of Assam.	Happiness, surprise, sadness, anger, fear, disgust, neutral.
12.	Chinese Emotion Speech Dataset [29]	Chinese	Simulated	A professional actress of a Reader's Digest Collection recorded 3649 phases and 1500 utterances	Happiness, anger, fear, anger, neutral
13.	Situation Analysis in a Fictional and Emotional corpus (SAFE) [30]	English	Elicited	4724 segments of speech were recorded by students. 400 sequences of audio/visual taken from 30 movies of 7 hours duration.	Positive, negative, neutral
14.	Multilingual Database [31]	Japanese, English, German	Natural and Simulated	Four emotional databases, 1. LEGO emotion database (English) 2. EMO-DB (GERMAN database), 3. UUDB (The Utsunomiya University Spoken Dialogue Database for paralinguistic information studies), and 4. SAVEE (Surrey Audio-Visual Expressed Emotion) corpus in (English).	1. Angry, slightly angry, very angry, neutral, friendly, and nonspeech (critical noisy recordings or just silence) 2. Neutral, anger, fear, joy, sadness, boredom, or disgust 3. Happy-exciting, angry, anxious, sad-bored, relaxed, serene. 4. Anger, disgust, fear, happiness, sadness, surprise, and neutral.
15.	Multilingual Database [32]	Indian English Malayalam and Tamil	Simulated	10 speakers Emotionally biased utterances	Angry, sad, happy.

C. SPEECH PROCESSING

The recoded audio signals contain the target speaker's speech and background noise, non-target speakers' voices, and reverberation. To automatically suppress such interference signals, various speech enhancement technologies are employed, such as speech processing. Speech processing

involves manipulating signals to change the signal's essential characteristics or extract vital information from it. Speech processing consists of the following steps.

1) PREPROCESSING

The first step after collecting the data is preprocessing. The collected data would be utilized to prepare the classifier in an SER system. While few of these preprocessing procedures are utilized for feature extraction, others take care of the normalization of the features so that the variations in the recordings of the speakers do not affect the recognition process [17].

2) FRAMING

The next step is known as signal framing. It is also alluded to as speech segmentation and is the way toward apportioning constant speech signals into fixed length sections to surpass a few SER difficulties. Emotions often tend to vary during a speech as a result of the signals being non-stationary. Despite this fact, the speech remains invariant even though it is for a very short period, such as 20 to 30 milliseconds. Speech signal, when framed, helps to estimate the semi-fixed and local features [33]. We can also retain the connection and data between the frames by intentionally covering 30% to 40% of these segments. The utilization of processing methods, for example, Discrete Fourier Transform (DFT) for feature extraction, SER can be controlled by persistent speech signals. Accordingly, fixed size frames are appropriate for classifiers, for example, ANNs, while holding the emotion data in speech.

3) WINDOWING

Once the framing in a speech signal is conducted, the frame is subject to the window function. During Fast Fourier Transform (FFT) of information, leakages occur due to discontinuities at the edge of the signals, henceforth reduced by the windowing function [34]. Generally, one of the sorts of the windowing function is Hamming window as defined in Eq. (1),

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{M-1}\right) \tag{1}$$

where the frame is $w(n)$, the window size is M, and $0 \leq n \leq M-1$.

4) VOICE ACTIVITY DETECTION

Three sections are included in utterance: unvoiced speech, voiced speech, and silence. If vocal cords play an active role in sound production, voiced speech is produced [10]. On the contrary, the speech is unvoiced if vocal cords are inactive. Voiced speech can be distinguished and extricated because of its periodic behavior. A voice activity detector could be used to detect voiced/unvoiced speech and silence in a speech signal.

5) NORMALIZATION

It is a methodology for adjusting the volume of sound to a standard level [17]. For normalization, the maximum value of the signal is obtained, and then the whole signal sequence is divided by the calculated maximum to estimate that every sentence has a similar level of volume. Z-normalization is generally used for normalization and is calculated as

$$z = \left(\frac{x-\mu}{\sigma}\right) \tag{2}$$

where μ is the mean, and σ is the standard deviation of the given speech signal.

6) NOISE REDUCTION

The environment is full of noises, and these noises are also encapsulated with every speech signal. Critically, the accuracy will be affected by the presence of noise in the speech signal. Therefore, for reducing this noise, several noise reduction algorithms can be utilized, like minimum mean square error (MMSE) and log-spectral amplitude MMSE (LogMMSE) [35].

The crucial phases in emotion recognition are feature selection and dimension reduction. Speech consists of numerous emotions and features, and one cannot state with certainty which set of features must be modeled and thus making a requirement for the utilization of feature selection techniques [36]. It is essential to do as such to preclude that the classifiers are not confronted with the scourge of dimensionality, incremented training time, and over-fitting that profoundly influence the prediction rate.

D. SPEECH FEATURES

The most predominant characteristics of SER are speech features. The recognition rate is enhanced by each precisely made arrangement of features that effectively describe every emotion. Different features have been utilized for SER frameworks. However, there is no commonly acknowledged arrangement of features for exact and particular classification [14]. To date, the existed studies have all been experimental. As we know, speech is practically a continuous signal but of varying length carrying both information and emotion. Thus, depending upon our needs, we may extract both global and local features or both. The comprehensive statistics like maximum and minimum values, standard deviation, and mean are represented by global features, also known as supra-segmental or long-term features.

On the contrary, the temporal dynamics are represented by local features, also called segmental or short-term features, to imprecise a fixed state [37]. The significance of these fixed features originates from the certitude that emotional features are not consistently appropriated over all the points of the speech signal. SER frameworks' global and local features are examined in the four classes, prosodic features, spectral features, voice quality features, and Teager Energy Operator (TEO) based features.

1) PROSODIC FEATURES

Several features like rhythm and intonation that human beings can recognize are known as prosodic features, or para-linguistic features as these features manage the components of speech that are properties of massive units as in sentences, words, syllables, and expressions and sentences [14]. Prosodic features are extricated from massive units and, thus, are long-term features. These features are the ones passing on unique properties of emotional substance for speech emotion recognition. Energy, duration, and fundamental frequency are some characteristics on which broadly utilized prosodic features are based.

2) SPECTRAL FEATURES

The vocal tract filters a sound when produced by an individual. The shape of the vocal tract controls the produced sound. An exact portrayal of the sound delivered and the vocal tract is resulted by precisely simulated shape. The vocal tract features are competently depicted in the frequency domain [38]. Fourier transform is utilized for obtaining the spectral features transforming the time domain signal into the frequency domain signal.

3) MEL FREQUENCY CEPSTRAL COEFFICIENTS

The most widely used spectral feature in automatic speech recognition is Mel Frequency Cepstral Coefficient (MFCC). MFCCs represent the envelope of the short-time power spectrum, which represents the shape of the vocal tract. The utterances are split into various segments before converting into the frequency domain using short-time discrete Fourier transform to obtain MFCC. Mel filter bank is utilized to calculate several sub-band energies. After that, the logarithm of respective sub-bands is computed. Lastly, MFCC is determined by applying the inverse Fourier transform [39].

4) LINEAR PREDICTION CEPSTRAL COEFFICIENTS

Linear prediction cepstral coefficients (LPCC) captures the emotion-specific information expressed through vocal tract characteristics. There are differences between the characteristics and emotions. Linear Prediction Coeficient (LPC) is primarily equivalent to the even envelope of the log spectrum of the speech, and the coefficients of all the pole-filters are used for obtaining the LPCC by a recursive method. The speech signal is flattened before processing to avoid additive noise error as LPCCs are more exposed to noise than MFCCs [40].

5) GAMMATONE FREQUENCY CEPSTRAL COEFFICIENTS

Gammatone frequency cepstral coefficients (GFCC) is computed by a method similar to that of MFCC, except that Gammatone filter-bank is applied in place of Mel filter bank to the power spectrum [3].

6) VOICE QUALITY FEATURES

Irrespective of other spectral features, the voice quality features define the qualities of the glottal source. The impact of the vocal tract is expiated to a large extent by inverse filtering. Some of the automatic changes might deliver a speech signal that may distinguish between various emotions utilizing the features like harmonics to noise ratio (HNR), shimmer, and jitter. The emotional content and voice quality of the speech have a compelling correlation between them [41].

7) TEAGER ENERGY OPERATOR BASED FEATURES

Teager energy operator (TEO) was introduced by Teager [42] and Kaiser [43]. TEO was framed on the confirmation that the hearing process is responsible for energy detection. It has been perceived that under stressful conditions, there is a change in fundamental frequency and critical bands because of the distribution of harmonics. A distressing circumstance affects the speaker's muscle pressure, resulting in modifying the airflow during the sound creation. Kaiser recorded the operator created by Teager to quantify the energy from a speech by this nonlinear process in Eq. (3), where Ψ is TEO and $x\,(n)$ is the sampled speech signal.

$$\Psi\left[x\,(n)\right] = x^2\,(n) - x\,(n-1)\,x\,(n+1) \tag{3}$$

Normalized TEO auto-correlation envelope area, TEO-decomposed frequency modulation variation, and critical band-based TEO auto-correlation envelope area are the three new TEO-based features introduced by [44].

E. CLASSIFIERS

For any utterance, the underlying emotions are classified using speech emotion recognition. Classification of SER can be carried out in two ways: (a) traditional classifiers and (b) deep learning classifiers. Numerous classifiers have been utilized for the SER system, but determining which works best is difficult. Therefore the ongoing researches are widely pragmatic.

SER systems generally utilize several traditional classification algorithms. The learning algorithm predicted a new class input, which requires the labeled data that recognizes the respective classes and samples by approximating the mapping function [45]. After the training process, the remaining data is utilized for testing the classifier performance. Examples of traditional classifiers include Gaussian Mixture Model, Hidden Markov Model, Artificial Neural Network, and Support Vector Machines. Some other traditional classification techniques involve k-Nearest Neighbor, Decision Trees, Naïve Bayes Classifiers [46], and k-means are preferred. Additionally, an ensemble technique is used for emotion recognition, which combines various classifiers to acquire more acceptable results.

1) GAUSSIAN MIXTURE MODEL (GMM)

GMM is a probabilistic methodology that is a prodigious instance of consistent HMM, consisting of just one state. The main aim of using mixture models is to template the data in a mixture of various segments, where every segment has an elementary parametric structure, like a Gaussian. It is presumed that every information guide alludes toward one of

the segments, and it is endeavored to infer the allocation for each portion freely [47].

GMM was contemplated for determining the emotion classification on two different speech databases, English and Swedish [48]. The outcome stipulated that GMM is an expedient method on the frame level. The two MFCC methods show similar performance, and MFCC low features outperformed the pitch features.

A semi-natural database GEU-SNESC (GEU Semi Natural Emotion Speech Corpus), was proposed [49]. Five emotions: happy, sad, anger, surprise, and neutral, were considered for the classification using the GMM classifier. For the characterization of emotions, the linear prediction residual of the speech signal was incorporated. The recognition percentage was discerned to be 50–60%.

2) HIDDEN MARKOV MODEL (HMM)

HMM is a usually utilized technique for recognizing speech and has been effectively expanded to perceive emotions[50]. HMM is a statistical Markov model in which the system is assumed to be a Markov process with an unobserved state. The term ''hidden'' indicates the ineptitude of seeing the procedure that creates the state at an instant of time. It is then possible to use a likelihood to foresee the accompanying state by referencing the current situation's target realities with the framework.

In [51], the authors demonstrated that HMM performs better on log frequency power coefficient features than LPCC and MFCC. The emotion classification was done based on text-independent methods. They attained a recognition rate of 89.2% for emotion classification and human recognition of 65.8%.

Hidden semi-continuous Markov models were utilized to construct a real-time multilingual speaker-independent emotion recognizer [52]. A higher than 70% recognition rate was obtained for the six emotions comprising anger, sadness, fear, joy, happiness, and disgust. INTERFACE emotional speech database was considered for the experiment.

3) SUPPORT VECTOR MACHINE (SVM)

An SVM classifier is supervised and preferential. The classifier is generally described for linearly separable patterns by splitting hyperplane. SVM makes use of the kernel trick to model nonlinear decision boundaries. The SVM classifier aims to detect that hyperplane having a maximum margin between two classes' data points. The original data points are mapped to a new space if the given patterns are not linearly separable by utilizing a kernel function [53].

SVM has been used as a classifier in [54] and was trained over Berlin Emo-DB and Chinese emotional databases (self-built). Only three emotions were considered sad, happy, and neutral. The investigated features included MFCC, energy, LPCC, Mel-energy spectrum dynamic coefficients, and pitch. The Chinese emotional database's overall accuracy rate was 91.3%, and for Berlin emotional database 95.1%.

The SVM classifier was compared with several other classifiers like radial basis function neural network, k-nearest-neighbor, and linear discriminant classifiers to check the accuracy rate for SER [55]. All the four classifiers were trained on emotional speech Chinese corpus. SVM performed best among all the classifiers with an 85% accuracy because of its good discriminating ability.

4) ARTIFICIAL NEURAL NETWORKS (ANN)

ANNs have been typically used for several kinds of issues linked with classification. It essentially consists of an input layer, at least one hidden layer, and an output layer. Since the layers consist of several nodes, the nodes present in an input and output layer depend upon the characterization of labeled class and data, while a similar number of nodes can be present in the hidden layer as per the requirement. The weights are arbitrarily chosen and are related to each layer. The qualities of a picked sample from training data are staked to the information layer and later forwarded to the next layer. The backpropagation algorithm is used for updating the weights at the output layer. The weights are foreseen to be able to classify the new data once the training has finished.

Two models are formulated to recognize emotions from speech based on ANN and SVM in [56], where the effect of feature dimensionality reduction to accuracy was evaluated. The features are extracted from CASIA Chinese Emotional Corpus. Initially, the ANN classifier showed 45.83% accuracy, but after the principal component analysis (PCA) over the features, ANN resulted in 75% improvement while SVM showed slightly better results, i.e., 76.67% of accuracy.

5) K-NEAREST NEIGHBOR (KNN)

k-NN is an uncomplicated supervised algorithm. The implementation of k-NN is easy and is utilized for solving both regression and classification problems. The algorithm is based on proximity, i.e., the data having similar characteristics near each other with a small distance. The calculation of distance depends upon the problem that is to be solved. Typically, Euclidean distance is used, as shown in Eq. (4).

$$d\,(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{4}$$

where x and y are two points in Euclidean space, while x_i and y_i are Euclidean vectors and N is the N-th space.

In the case of classification, the data is classified based on the vote of its neighbor. The data is assigned to the most common class among its k-nearest neighbors. If the value of k is 1, the data is assigned to the class of that single nearest neighbor.

In [57], kNN and ANN were utilized as classifiers, and Hurst parameters and LPCC features were extracted from speech recordings of the Telegu and Tamil languages. The evaluation results showed that the combination of features yielded better results than individual features with an accuracy of 75.27% for ANN and 69.89% for kNN.

6) DECISION TREE

A decision tree is a nonlinear classification technique based on the divide and conquers algorithm. This method can be considered a graphical representation of trees consisting of roots, branches, and leaf nodes. Roots indicate tests for the particular value of a specific attribute, and from where decision alternative branches originate, edges/branches represent the output of the text and connects to the next leaf/ node, and leaf nodes represent the terminal nodes that predict the output and assign class distribution or class labels. Decision Tree helps in solving both regression and classification problems. For regression problems, continuous values, which are generally real numbers, are taken as input. In classification problems, a Decision Tree takes discrete or categorical values based on binary recursive partitioning involving the fragmentation of data into subsets, further fragmented into smaller subsets. This process continues until the subset data is sufficiently homogenous, and after all the criteria have been efficiently met, the algorithm stops the process.

A binary decision tree consisting of SVM classifiers was utilized to classify seven emotions in [58]. Three databases were used, including EmoDB, SAVEE, and Polish Emotion Speech Database. The classification done was based on subjective and objective classes. The highest recognition rate of 82.9% was obtained for EmoDB and least for Polish Emotional Speech Database with 56.25%.

7) NAÏVE BAYES CLASSIFIER

Naïve Bayes Classifier is a decent supervised learning method. The classification is based on Bayes theorem as given in Eq. (5).

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \tag{5}$$

where x represents a class variable and y represents the features/parameters.

Naïve Bayes Classifier is a probabilistic algorithm that utilizes the joint probabilities assuming that the existence of a particular feature in a class is independent of the existence of any other feature. This independence allows the features to be learned separately, simplifying and increasing the computation operations [59].

Naïve Bayes Classifier was trained on EmoDB for emotion recognition in [60]. The authors combined the spectral (MFCC) and prosodic (pitch) features to enhance the SER system's performance. The evaluation result was divided into four classes based on speakers and emotions considered. The highest accuracy of 95.23% was obtained for the class, consisting of one male speaker's speech samples.

Deep learning algorithms usually allude to deep neural networks, and the vast majority are dependent on ANN. Though a traditional neural network consists of two or three few hidden layers, there could be hundreds of hidden layers in neural networks. Thus the expression ''deep'' originates from the hidden layers. Generally, the deep learning algorithms' execution outperforms the traditional machine learning algorithms, thus emphasizing their application to SER [9]. A portion of these algorithms' benefit is that there is no requirement for feature extraction and selection steps. Deep learning algorithms select all the features automatically. Most generally utilized deep learning algorithms in the SER area are Deep Neural Networks, Deep Belief Networks, Deep Boltzmann Machine, Recurrent Neural Networks and Long-Short Term Memory.

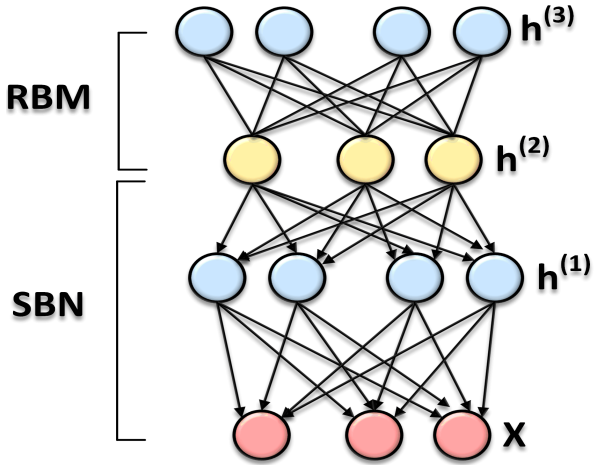

FIGURE 2. Basic Architecture of Deep Belief Network.

8) DEEP NEURAL NETWORKS

Deep Neural Networks (DNN) is a neural network with multiple layers and multifaceted nature to process data in complex ways. It can be described as networks with a data layer, an output layer, and one hidden layer in the center. Each layer performs precise types of organizing and requisites in a method that some suggest as ''feature hierarchy.'' One of the keys implementations of these refined neural networks is overseeing unlabeled or unstructured data.

A custom-made database was proposed in [61]. For the recognition of emotions, DNN was utilized. First, the network was optimized for four emotions, giving the recognition rate of 97.1% and then for three emotions, resulting in a 96.4% recognition rate. Only the MFCC feature was considered for the experiment.

An amalgam of the traditional classification approach – GMM with the neural network was utilized to recognize emotions [62]. A total of four distinct algorithms were used for the classification process: DNN, GMM, and two different variations of Extreme Machine Learning (EML). It was found that the DNN-EML approach outshined the GMM-based algorithms in terms of accuracy.

9) DEEP BELIEF NETWORKS

Deep Belief Networks (DBN) is an unsupervised generative model that mixes the directed and undirected connections between the variables that constitute either the visible layer or all hidden layers [63], as shown in Fig. 2.

DBN is not a feedforward network. It is a specific model where the hidden units are binary stochastic random variables Figure 2 depicts the basic architecture of DBN with three hidden layers and one visible layer. There are undirected interactions between $h^{(3)}$ and $h^{(2)}$ and directed connections between $h^{(2)}$ and $h^{(1)}$, and $h^{(1)}$ and x. The top layers in DBN form a restricted Boltzmann machine, while the other layers form Bayesian Network with directed interactions. The conditional distribution of a layer in Figure 2 is shown in Eq (6).

$$p\left(h_j^{(1)} = 1|h^{(2)}\right) = sigm\left(b^{(1)} + W^{(2)^T} h^{(2)}\right)$$
$$p\left(x_i = 1|h^{(1)}\right) = sigm\left(b^{(0)} + W^{(1)^T} h^{(1)}\right) \quad (6)$$

where $h^{(1)}$ and $h^{(2)}$ are the first and second hidden layer respectively, h_j and x_i are the hidden and inputs, respectively, b is the bias, and $W^{(2)}$ is the connection between $h^{(2)}$ and $h^{(1)}$ and $W^{(1)}$ is the connection between $h^{(1)}$ and x. The probability of either of the units to be equal to 1. The sigmoid is applied on the linear transformation of the layer above it, e.g., for $h^{(1)}$, the linear transformation of $h^{(2)}$ is taken. This type of interaction is referred to as Sigmoid Belief Network (SBN) [64].

In [65], several features were fused to explore the relationship between the emotion recognition performance and feature combination. For the classification process, DBN was utilized and was trained on the Chinese Academy of Science emotional speech database. Along with the DBN, the classification experiment was performed using SVM. The evaluation results revealed that DBN performed better than SVM with an accuracy of 94.6%, while SVM achieved 84.54% accuracy.

A novel classification technique consisting of the combination of DBN and SVM was proposed in [66]. The evaluation process was carried on the Chinese speech emotion dataset. DBN was utilized to extract features and was trained by the conjugate gradient method, while as classification process was carried by SVM. The results showed that the proposed method worked well for small training samples and achieved an accuracy of 95.8%.

10) DEEP BOLTZMANN MACHINE

Deep Boltzmann Machine (DBM) is a probabilistic unsupervised generative model. DBM is a network of symmetrically connected two stochastic binary units, consisting of visible units and multiple hidden layers, as shown in Fig. 3. The network has undirected connections between the neighboring layers. The units within the layers are independent of each other but are dependent on the neighboring layers. The basic architecture of DBM is shown in Figure 3. The probability of the visible/hidden units is achieved by marginalizing the joint probability [63] defined in Eq. (7).

$$p(v, h) = \frac{e^{-E(v,h)}}{\sum_{m,n} e^{-E(m,n)}} \quad (7)$$

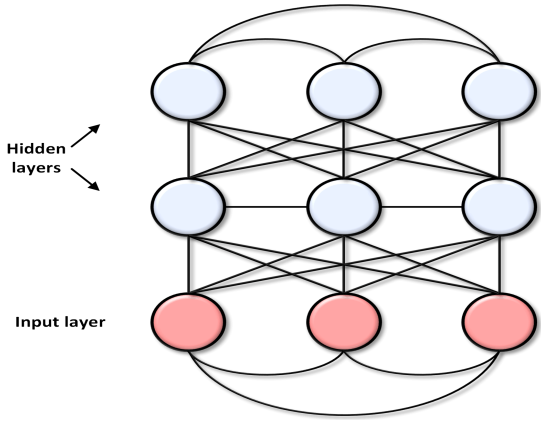

FIGURE 3. Basic Architecture of Deep Boltzmann Machine.

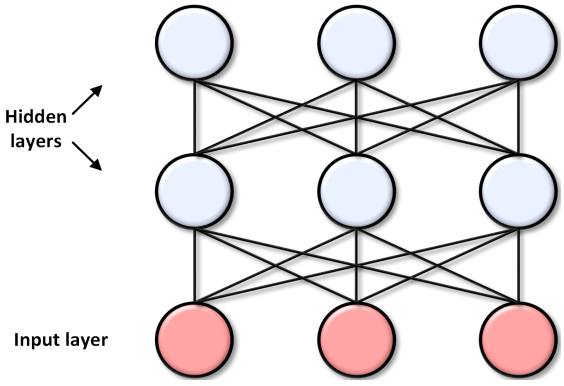

FIGURE 4. Basic Architecture of Restricted Boltzmann Machine.

where v and h are the visible and hidden units respectively and $E(v, h)$ is the Boltzmann Machine's energy function defined in Eq. (8).

$$E(v, h) = \sum_i v_i b_i - \sum_{n=1}^{N-1} \sum_k h_{n,k} b_{n,k} - \sum_{i,k} v_i w_{i,k} h_k - \sum_{n=1}^{N-1} \sum_{k,l} h_{n,k} w_{n,k,l} h_{n,l} \quad (8)$$

where b is the bias and w are the connections between visible and hidden layer, and N is the number of hidden layers.

11) RESTRICTED BOLTZMANN MACHINE

Restricted Boltzmann Machine (RBM) is a particular case of Boltzmann Machine, having the restrictions in terms of connections either between hidden units or between visible units as depicted in Fig. 4. The model becomes a bipartite graph by removing the connections between hidden and visible units [63]. The energy function $E(v, h)$ of RBM is defined in Eq. (9).

$$E(v, h) = \sum_i v_i b_i - \sum_k h_k b_k - \sum_{i,k} v_i h_k w_{i,k} \quad (9)$$

where v and h are the visible and hidden units, respectively, b is the bias, and w is the connections between visible and hidden layers.

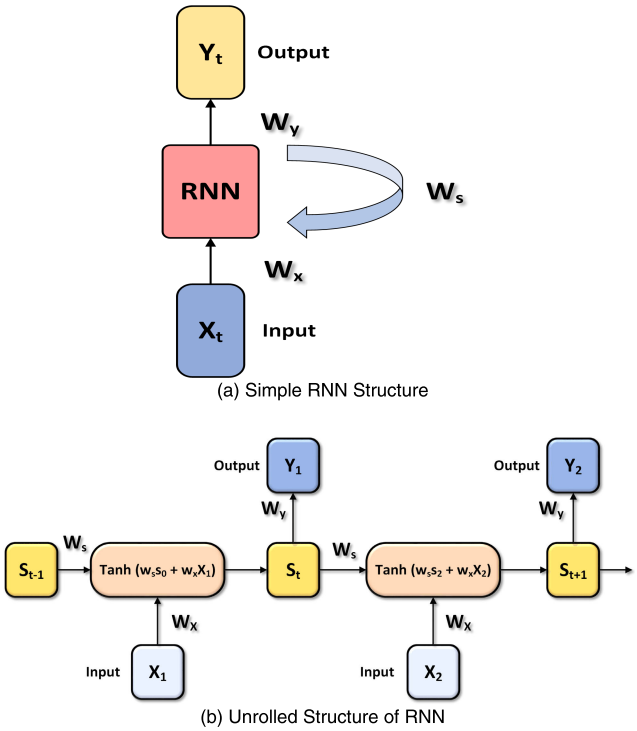

FIGURE 5. Recurrent Neural Networks Architectures.

12) RECURRENT NEURAL NETWORKS

Recurrent Neural Networks (RNN) is designed for capturing information from sequence/time series data and are generally utilized for temporal problems like natural language processing, image capturing, and speech recognition. They are eminent by the "memory" as they take data from previous inputs to influence the current input and output. RNNs work on the recursive formula given in Eq. (10).

$$S_t = F_W\left(S_{t-1}, X_t\right) \tag{10}$$

where X_t is the input at time t, S_t(new state), and S_{t-1} (previous state) is the state at time t and t-1, respectively, and F_W is the recursive function. The recursive function is a *tanh* function. The equation is simplified as given in Eq. (11), where W_s and W_x are weights of the previous state and input, respectively, and Y_t is the output.

$$\begin{aligned} S_t &= \tanh\left(W_s S_{t-1} + W_x X_t\right) \\ Y_t &= W_y S_t \end{aligned} \tag{11}$$

Figure 5(a) shows the simple RNN structure, while Figure 5(b) depicts the unrolled structure of RNN. Unfortunately, the gradient in deep neural networks is unstable as they tend to either increase or decrease exponentially, which is known as the vanishing/exploding gradient problem [67]. This problem is much worse in RNNs. When we train RNNs, we calculate the gradient through all the different layers and through time, which leads to many more layers, and thus the vanishing gradient problem becomes much worse. This problem is solved by Long Short-Term Memory architecture.

An efficient approach based on RNN for emotion recognition was presented in [68]. The evaluation was done in terms of weighted accuracy and unweighted accuracy and contemplated the long-range contextual effects. The proposed framework resulted in 62.85% and 63.89% of weighted and unweighted accuracies measures, respectively.

In [69], it was depicted that RNN can learn temporal aggregation and frame-level characterization over a long period. Besides, a different procedure for feature pooling using RNN with local attention showed robust accuracy in classification compared to the traditional SVM approach.

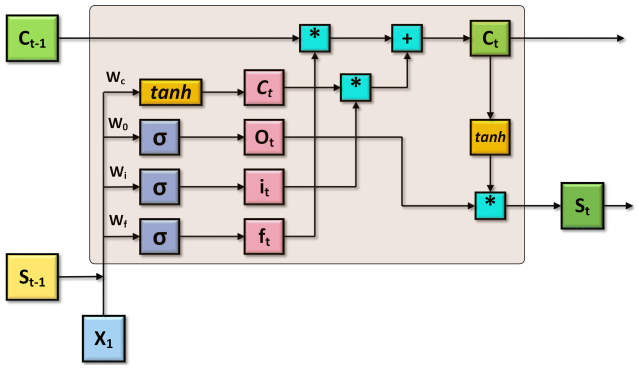

FIGURE 6. Basic Architecture of LSTM.

13) LONG SHORT-TERM MEMORY

Long Short-Term Memory (LSTM) is precisely designed to solve vanishing gradient by adding extra network interactions. LSTM consists of three gates (forget, input and output) and one cell state. The forget gate decides what information from previous inputs to forget, the input gate decides what new information to remember, and the output gate decides which part of the cell state to output. Therefore, LSTM, as shown in Fig. 6, can forget and remember the information in the cell state using gates and retain the long-term dependencies by connecting the past information to the present [70].

The governing equations of forget gate, input gate, output gate, and cell state are presented in Eq. (12).

$$\begin{aligned} f_t &= \sigma\left(W_f S_{t-1} + W_f X_t\right) \\ i_t &= \sigma\left(W_i S_{t-1} + W_i X_t\right) \\ o_t &= \sigma\left(W_o S_t + W_o X_t\right) \\ c_t &= \left(i_t * \tilde{C}_t\right) + \left(f_t * c_{t-1}\right) \end{aligned} \tag{12}$$

where f_t, i_t, o_t and c_t are the forget gate, input gate, output gate, and cell gate, respectively, σ is the sigmoid activation function, S_{t-1} is the previous states, X_t is the input at time t, W_f, W_i and W_o are a respective set of weights of the forget gate, input gate, and output gate. $\tilde{C}_t$ is the intermediate cell state defined in Eq. (13), and the new next state is obtained using Eq. (14).

$$\tilde{C}_t = \tanh\left(W_c S_{t-1} + W_c X_t\right) \tag{13}$$

$$S_t = o_t * \tanh\left(c_t\right) \tag{14}$$

where W_c is the weight of the cell state and S_t is the new state. All the multiplications are element-wise multiplication.

14) CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNN) is the systematic neural network that consists of various layers sequentially. Generally, the CNN model consists of various convolution layers, pooling layers, fully connected layers, and a SoftMax unit. This sequential network forms a feature extraction pipeline modeling the input in the form of an abstract. CNN is mainly used for analyzing image/data classification problems. The basis of CNN is convolutional layers, which constitute filters. These layers perform a convolutional operation and pass the output to the pooling layer. The pooling layer's main aim is to reduce the resolution of the output of convolutional layers, therefore reducing the computational load. The resulting outcome is fed to a fully connected layer, where the data is flattened and is finally classified by the SoftMax unit, which extends the idea of a multiclass world.

In [71], deep CNN is utilized for emotion classification. The input of the deep CNN were spectrograms generated from the speech signals. The model consisted of three convolution layers, three fully connected layers, and a SoftMax unit for the classification process. The proposed framework achieved an overall accuracy of 84.3% and showed that a freshly trained model gives better results than a fine-tuned model.

III. REVIEW OF SPEECH EMOTION RECOGNITION

For the last two decades, the field of SER has been studied extensively. Numerous speech features and various techniques for their extraction have been explored and implemented. Different supervised and unsupervised classification algorithms have been evaluated for effective emotion recognition. However, there is no common consensus on how to measure or categorize emotions as they are subjective. The SER system's crucial aspect is selecting the speech emotion corpora (database), recognizing various features inherited in speech, and a flexible model for the classification of those features.

In recent times, deep learning techniques have shown some promising results for emotion recognition. They are considered best suited for the SER system over traditional techniques because of their advantages like scalability, all-purpose parameter fitting, and infinitely flexible function. On the other hand, traditional classifiers are fast as they need less data for training. Table 2 gives a brief comparison of traditional classifiers and deep learning classifiers. A survey of various databases and different classifiers that have been implemented in recent years has been presented in Table 3. Moreover, their contributions in terms of novelty and recognition rates are provided, and some future work recommendations.

GMM and KNN are implemented in [72] and used three features: wavelet, pitch, and MFCC. GMM classifier performed better than the KNN classifier in terms of precision and F-score. A traditional classifier requires a minimum dataset for the training process. Taking this advantage

TABLE 2. Comparison between traditional and deep learning classifiers.

Deep-Learning Classifiers	Traditional Classifiers
Unsupervised algorithms	Supervised Algorithms
Scalable	Non-scalable
No need for feature engineering	Feature engineering is a crucial
Larger datasets are required for training	Fewer data required for training
Expensive	Financially cheap
Hard to interpret	Easier to interpret
Libraries: TensorFlow, Pytorch, Keras, Theans, Caffe	Libraries: Scikit-learn

in [73], the authors used the Naïve Bayes classifier to classify four emotions. Four different statistical features have been extracted, including pitch, zero-crossing rate (ZCR), MFCC, and energy from the created dataset of 200 utterances. The highest recognition accuracy of 81% is obtained for anger emotion and least for sad with 76%. The results revealed that the presented system is capable of real-time emotion recognition.

For the representation of speech signals, random forest and decision tree are used for classification [81]. The feature vector is selected with a reasonable length of 14 for less time usage. A two-stage classification technique using the decision tree and SVM is used in [87]. Firstly, a rough classification is performed, and then a fine classification for eliminating redundant parameters and improvement in performance.

Various features like MFCC, log power, delta MFCC, double delta MFCC, LPCC, and LFPC have been used with HMM and SVM to classify seven different emotions [76]. MFCC obtained the best accuracy of 82.14% for SVM and performed consistently with less computational complexity. The model could be used in call center applications for recognizing emotions over the telephone. Binary classification is implemented in [74] instead of a single multiclass classifier for emotion recognition.

A hierarchical binary tree framework is built where the easy task of classification is performed at the top level, and more challenging tasks are solved at the end level. Bayesian Logistic Regression and SVM are used for preventing the overfitting and finding enough separation between the two classes, respectively. The proposed method achieves unweighted and weighted accuracy of 58.46% and 56.38%, respectively, for the USE-IEMOCAP database.

The Fisher criteria are used for feature selection, and an average recognition rate of 83.75% is obtained. The preprocessing technique involving sampling, normalization, segmentation, and feature extraction are implemented [101]. Commonly used features are extracted by fixing an initial and endpoint detection algorithm to infer the speech signal's initial and last sample point. The training process is tested several times to get desirable results using ANN. The overall accuracy of 86.87% is obtained for neutral, happy, angry, and sad.

Feature extraction is one of the most critical tasks in speech emotion recognition. Because of the presence of numerous features in the speech signal, there is no prominent set of

TABLE 3. Literature survey on different databases and classification techniques for speech emotion recognition.

S. No	Paper	Database	Emotions Considered	Classifier	Contribution	Future Recommendation
1.	C. Chun Lie et al. (2011) [74]	AIBO and USC IEMOCAP	Angry, rest, positive, negative, and empathetic.	SVM and BLR	A hierarchical computational structure is proposed to identity emotions. The method achieved an improvement of 3.37% for AIBO and 7.44% for USC IEMOCAP.	An automatic hierarchical structure can be generated that would decrease the number of iterations.
2.	L. Li et al. (2013) [75]	eNTERFACE and EmoDB	Happy, sad, angry, disgust, fear, and surprise.	DNN-HMM, RBM	DNN-HMM is implemented with RBM. Based on unsupervised pre-training, an accuracy of 74.28% is obtained and 77/92% for discriminative pre-training.	The method can be evaluated for audio-video and facial action emotion recognition.
3	M.Swain et al. (2015) [76]	Multilingual database consisting of two native language of Odisha (Sambalpuri and Cuttacki).	Happiness, disgust, sadness, fear, neutral and surprise	HMM and SVM	MFCC with SVM is computationally more efficient. Better results are obtained for speaker-independent using HMM at 78.81% and SVM, with 82.14% accuracy for the Sambalpuri language.	Hierarchical classification of emotion can be utilized for enhancing the performance of the system.
4.	Rahul B. et al. (2015) [72]	Berlin Emotional Database (Emo-DB)	Angry, happy, sad, surprised, neutral, and fearful	GMM and K-NN	Enhancement of computational speed using K-NN. GMM technique recognizes the 'angry' emotion with 92% accuracy, while the K-NN technique detects the 'happy' emotion with a 90% rate.	The two techniques can be fused for effective emotion recognition.
5.	Sagar K et al. (2016) [73]	Self-created database, consisting of male utterances. Age group 18-30.	Angry, happy, sad, and neutral	Naïve Bayes Classifier	MFCC, pitch, and energy features are used to train the classifier. The accuracy is 76% for sad, 77% for neutral, 78% for happy, and 81% for angry.	Voice quality features and prosodic features can be used to enhance the performance of the system.
6.	Akash Shaw et al. (2016) [77]	A batch of self-recorded speech	Happy, angry, sad, and neutral	ANN	The selected features, formant, pitch, energy, and MFCC, prove effective for speech emotion recognition with an 85% classification rate.	The system can be designed to recognize the images also.
7	Wootaek Lim et al. (2017) [78]	Berlin Emotional Database (Emo-DB)	Neutral, anger, fear, disgust, sadness, boredom, and happy	CNN, time distributed CNN and LSTM	Verified that time-distributed CNNs show better results. The F1 scores (%) for CNN, LSTM, and time-distributed CNNs are 86.06, 78.31, and 86.65, respectively.	Better results expected by utilizing concatenated CNNs.
8.	Parvol H. et al. (2017) [79]	Berlin Emotional Database (Emo-DB)	Angry, sad, and neutral	DNN with 6 convolutional layers, pooling and 3 fully connected layers and Voice Activity Detection (VAD)	The approached method achieves 96.67% of accuracy on testing data and 69.55% for prediction data.	The results can be improved by implementing RNNs
9.	Guihua Wen et al. (2017) [80]	EmoDB, SAVEE, CASIA, and FAU-AIBO	Happy, anger, sadness, fear, disgust, neutral, and surprise	RDBN, RBF-SVM, L-SVM, SOFTMAX, KNN	An ensemble of RDBN's technique is presented. RBF-SVM obtained the highest accuracy for 3 databases with 48.50%, 53.60%, and 82.32% for CASIA, SAVEE, and EmoDB.	RDBN can be trained on a more extensive database for high accuracy and enhance the network's performance.

TABLE 3. ***(Continued)*** **Literature survey on different databases and classification techniques for speech emotion recognition.**

10.	F. Noroozi et al. (2017) [81]	SAVEE	Happiness, sadness, surprise, fear, disgust, and fear.	Random forest and Decision tree	Random forest algorithm is adopted Decision tree approach for the voice-based emotion recognition. 78% of the recognition rate is obtained using the proposed method.	Different features like MFCC and FBE can be utilized for improving the performance of the system.
11.	Po-W.Hsiao et al. (2018) [82]	FAU-Aibo	Anger, neutral, emphatic, rest and positive	RNN and LSTM	The performance measured by unweighted accuracy recall rate enhances from 37.0% of RNN model to 46.3% of LSTM attention model.	Better performance can be achieved by utilizing deeper network architecture.
12.	L. Kerkeni et al. (2018) [83]	Berlin Emo-DB and Spanish Database	Anger, disgust, joy, fear, surprise, sadness, and neutral	RNN and MLR	The best recognition rate is achieved for the combination of MFCC and MS features with 90.05% for the Spanish dataset using RNN and 82.41% using MLR for the Berlin dataset.	The system can be improved for real-time speech emotion recognition.
13.	Zhen-Tao Liu et al. (2018) [84]	CASIA, SAVEE, FAU-Aibo and INTERSPEECH 2009	Happy, sad, angry, fear, surprise and neutral	GA-BEL	Genetic Algorithm (GA) is used for updating the weights of brain emotional learning (BEL). The highest average accuracy obtained is 70.28%, 71.05%, and 76.40% for CASIA, FAU-Aibo, and SAVEE.	The proposed method can be used in the Artificial Intelligence area and can be applied to multi-model-based human-robot interaction.
14.	Gang Liu et al. (2018) [85]	CASIA	Normal, happy, sad, fear and surprise	CNN and DNN	A fusion of spectral-based and pitch-based hyper-prosodic features is proposed. An average precision of 86.58% is obtained.	Time-domain signal and the frequency-domain signal can be implemented for a robust SER system.
15.	Alghifari et al. (2018) [61]	EmoDB	Happy, sad, angry, and neutral	DFNN	An accuracy of 96.3% is obtained for 3 emotions and 97.1% for 4 emotions.	Future work includes the enhancement of the SER system by utilizing different databases and real-life applications.
16.	M. Neumann et al. (2019) [86]	IEMOCAP and MSP-IMPROV	Angry, sad, happy, and neutral.	ACNN	The visualizations of t-SNE are presented, which reveals the discriminative strength. The model achieves a recognition rate of 59.54%.	More experiments with various variants of autoencoders
17.	L.Sun (2019) [87]	CASIA and Berlin Emo-DB	Angry, happy, boring, neutral, sad, fear, and disgust	Decision tree SVM with Fisher feature selection	The proposed method is verified to reduce emotional confusion effectively. 83.75% or recognition rate is achieved for CASIA and 86.86% for Emo-DB.	More effective feature selection methods and feature parameters for feature selection to be found
18.	Mohammad Mehedi et al. (2019) [88]	DEAP (A Dataset for Emotional Analysis using Phycological signals)	Sad, relaxed, happy, neutral, and disgust	Deep Belief Network (DBN) and FGSVM	A new feature fusion of PPG, EDA, and zEMG is put forward. DBM and FGSVM architecture achieve 89.53% of accuracy.	Along with the proposed method, an ensemble classifier technique can be employed to increase the generalizability and robustness of SER.
19.	Linhui Sun et al. (2019) [89]	Chinese Academy Sciences Emotional Corpus	Anger, happy, sad, fear, neutral, and surprise	SVM, DNN-SVM, DNN-decision tree SVM	A multi-level classifier DNN-decision tree SVM is introduced and obtains 75.83% accuracy outperforming SVM and DNN-SVM.	Different features and models can be explored to improve the working of the SER system.
20.	Seunghyun Yoon et al. (2019) [90]	IEMOCAP	Happy, sad, angry, and neutral	Dual Recurrent neural network (RNN)	A novel architecture MDRE that utilized audio signals and text data is introduced. The proposed method obtains accuracy ranging from 68.8% to 71.8%.	Apart from audio signals and text data, video inputs can be investigated using multiple modalities.

TABLE 3. *(Continued)* **Literature survey on different databases and classification techniques for speech emotion recognition.**

21.	Siddique Latif et al. (2019) [91]	IEMOCAP and MSP-IMPROV	Angry, happy, neutral, and sad	CNN, LSTM, and DNN	For feature extraction, a parallel convolutional joined with LSTM is proposed. A combination of CC-LSTM-DNN achieved an accuracy of 60.23% for IEMOCAP and 52.43% for MSP-IMPROV	Raw speech can be used to examine the proposed system. Also, raw speech and hand-engineered feature-based methods can be compared.
22.	H.S. Kumbhar et al. (2019) [92]	RAVDEES	Happy, sad, calm, angry, surprise, fearful, disgust	LSTM	Multidimensional complex data is learned using MFCC and LSTM and obtained an accuracy of 80.81%	For a true positive rate of ROC, different features can be used.
23.	D. Bharti et al. (2020) [93]	RAVDEES	Happiness, sad, angry and joy	ALO, GFCC, and MSVM	A new MVSM is used for the classification process, and ALO+MSVM achieved 97% of accuracy.	An efficient feature fusion may be implemented to enhance the multimodal system performance.
24.	Leila Kerkerni et. al (2020) [94]	EmoDB and Spanish Database.	Angry, joy, sadness, disgust, neutral, fear, and sadness	RNN, MLR, and SVM	83% of accuracy is obtained for EmoDB and 94% for Spanish Database.	A pedagogical interaction system is aimed to be developed by using more feature extraction methods.
25.	Z. Yao et al. (2020) [95]	IEMOCAP	Happy, angry, sad, and neutral	HSF-DNN. MS-CNN and LLD-RNN	A new confidence-based fusion method is proposed. The WA of 57.1% and UA of 58.3% are achieved.	A fusion of more deep learning methods can be carried out.
26.	Mustaqeem et al. (2020) [96]	IEMOCAP, EmoDB, and RAVDEES	Happy, sad, angry, and neutral	RBFN, CNN, and BiLSTM	A novel architecture of the SER system has been developed using RBFN. An accuracy of 77.02% has been obtained for RAVDEES, 72.25%, and 85.57% accuracy for IEMOCAP and EmoDB, respectively.	DBN and GRU can be implemented for better accuracies.
27.	TM Wani et al. (2020) [97]	SAVEE	Angry, sad, happy, and neutral	CNN and DSCNN	A different framework DSCNN has been used. 87.8% and 79.4% of accuracy are obtained for DSCNN and CNN, respectively.	The methods can be improved for the enhancement of accuracy.
28.	Z. Huijuan et al. (2020) [98]	IEMOCAP	Happy, angry, frustrated, and excited	CNN blocks with RNN attention module.	A novel top-down hierarchical classification system has been introduced. F1 score of 0.4673 is achieved using 3D-HMTL	Improvement can be made in performance by using a fine-grained classification of emotions.
29.	Md Shah et al. (2020) [99]	IEMOCAP	Happy, sad, angry, and neutral	DNN-HMM	Epoch-based features have been used. A recognition rate of 60.86% is obtained for MFCC and 65.93% for MFCC + Epoch based features.	For the epoch-based features, LSTM can be used for better recognition rates.
30.	Mingke Xu et. al (2020) [100]	IEMOCAP	Happy, sad, exited and neutral.	ACNN	A multi-head self-attention model for SER has been utilized. 76.18% and 76.36% of weighted accuracy (WA) and 76 unweighted accuracy (UA) are obtained.	A generative adversarial network can be used for developing more emotional data for training.

features that could be efficiently utilized in the SER system to recognize emotions. However, recent researches have used the features fusion, which has enhanced the SER system in terms of recognition accuracy [85], [107]–[109]. The fusion is not limited to the features but has been implemented in classification techniques as well. Many traditional classifiers have been fused with each other to enhance the recognition rate of models. Likewise, many deep learning classifiers with other deep learning classifiers and many traditional classifiers have been assimilated with deep learning methods, showing some good results.

Many speech variations are mainly due to different speakers, their speaking styles, and speaking rate. The other reason being the environment and culture in which the speaker expresses certain emotions. The multiple levels of speech signals are easily discovered by Deep Belief Networks (DBN).

This significance is well exploited in [80] by proposing an assemble of random deep belief networks (RDBN) algorithm for extracting the high-level features from the input speech signal. Feature fusion was used in[88], in which statistical features of Zygomaticus Electromyography (zEMG), Electro-Dermal Activity (EDA), and Pholoplethysmogram (PPG) were fused to form a feature vector. This feature vector is combined with DBN features for classification. For the nonlinear classification of emotions, a Fine Gaussian Support Vector Machine (FGSVM) is used. The model successfully implemented and archives an accuracy of 89.53%.

Deep learning classifiers have been tremendously utilized in recent times. A feedforward neural network with multiple hidden layers and many hidden variables is utilized to recognize emotions [61]. DNN is trained on Emo-DB. Only four emotions have been considered, including happy, sad, angry, and neutral. An optimum configuration of 13 MFCC, 12 neurons, and two hidden layers have been used for three emotions, and an accuracy of 96.3% is achieved. For four emotions, 97.1% is obtained with the optimum configuration of 25 MFCC, 21 neurons, and four layers. DNN is widely preferred for the feature extraction of deep bottleneck features used to classify emotions.

In [89], DNN decision trees SVM is presented where initially decision tree SVM framework based on the confusion degree of emotions is built, and then DNN extracts the bottleneck features used to train SVM in the decision tree. The evaluated results revealed that the proposed method of DNN-decision tree outperforms the SVM and DNN-SVM in terms of recognition rates. DNN with convolutional, pooling and fully connected layers are trained on Emo-DB [79], where Stochastic Gradient Descent is used to optimize DNN. Silent segments are removed using voice activity detection (VAD), and an accuracy of 96.67% is achieved. A multimodal text and audio system using a dual RNN is proposed in [90]. Different auto-encoder, including Audio Recurrent Encoder, to predict the class of given audio signal.

Text Recurrent Encoder to process textual data for predicting emotions is used to evaluate the system's efficiency and performance. Two novel architectures are put forward, Multi Dual Recurrent Encoder and Multimodal Dual Recurrent Encoder with Attention (MDREA), to focus on the particular piece of transcripts to encode information from audio and textual inputs independently. The MDREA method solved the problem of inaccuracies in predictions. Ant lion optimization algorithm and multiclass SVM are utilized to select the feature set and train the modal, respectively [93]. The modal efficiently imposes the signal-to-noise ratio (SNR), and a recognition rate of 97% is achieved.

MFCC is widely used to analyze any speech signal and had performed well for speech-based emotion recognition systems compared to other features. In [92], MFCC feature extraction is used, and 39 coefficients are extracted. Long Short-Term Memory (LSTM) is implemented for emotion recognition. The ROC curve determines the recognition accuracy. The training data achieves less accuracy than the test data and the average test data accuracy obtained is 84.81%. Gammatone frequency cepstral coefficients have been used for feature extraction. A concatenated CNN and RNN architecture is employed without utilizing the traditional handcrafted features [78], and a time-distributed CNN network is proposed. Besides, a framework is combined with LSTM network layers for emotion recognition. The highest accuracy of 86.65% is obtained for time-distributed CNN. A coarse fine classification model is presented using the CNN block module RNN, attention module, and feature fusion module to classify fine classes in specific course types [98].

SER system provides an efficient mechanism for systematic communication between humans and machines by extracting the silent and other discriminative features. CNN has been used for extracting the high-level features using spectrograms [71], [97], [102]. A different framework of CNN, referred to as Deep Stride Convolutional Neural Network (DSCNN), using strides in place of pooling layers, has been implemented in [97], [103] for emotion recognition. The proposed model in [91] uses parallel convolutional layers of CNN to control the different temporal resolutions in the feature extraction block and is trained with LSTM based classification network to recognize emotions. The presented model captures the short-term and long-term interactions and thus enhances the performance of the SER system.

An essential sequence segment selection based on a radial basis function network (RBFN) is presented in [96], where the selected sequence is converted to spectrograms and passed to the CNN model for the extraction of silent and discriminative features. The CNN features are normalized and fed to deep bi-directional long short-term memory (BiLSTM) for learning temporal features to recognize emotions. An accuracy of 72.25%, 77.02%, and 85.57% is obtained for IEMOCAP, RAVDEES, and EmoDB, respectively.

An integrated framework of DNN, CNN, and RNN is developed in [95]. The utterance level outputs of high-level statistical functions (HSF), segment-level Mel-spectrograms (MS), and frame-level low-level descriptors (LLDs) are passed to DNN, CNN, and RNN, respectively, and three separate models HSF-DNN, MS-CNN, and LLD-RNN are obtained. A multi-task learning strategy is implemented in three models for acquiring the generalized features by operating regression of emotional attributes and classification of discrete categories simultaneously. The fusion model obtained a weighted accuracy of 57.1% and an unweighted accuracy of 58.3% higher than individual classifier accuracy and validated the proposed model's significance.

A neural network attention mechanism is inspired by the biological visual attention mechanism found in nature. The attention mechanism increases the computational load of the model, but it enhances the accuracy rates as well as the model's performance. In [86], the authors utilized unlabeled speech data for emotion recognition. Autoencoder is incorporated with CNN to improve the performance of the SER system. Attention-based Convolutional Neural Network (ACNN) is used as the baseline structure and

trained on USE-IEMOCAP and tested on MSP-IMPROV and ACNN-AE on Tedium. The ACNN-AE approach achieves better results than the baseline ACNN. Additionally, the t-distributed stochastic neighbor embeddings (t-SNE) have been utilized for 2D projections of speech data, and autoencoder significantly separates the respective high and low arousal speech signals.

In [94], RNN, SVM, and MLR (multivariable linear regression) are compared. RNN performed better than SVM and MLR with 94.1% accuracy for Spanish databases and 83.42% for Emo-DB. The research concluded that SVM and MLR perform better on fewer data than RNN, which needs a more significant amount of training data. In [82], RNN obtained an unweighted accuracy of 37%, and LSTM-attention achieves 46.3% on the dynamic modeling structure of FAU Aibo tasks. A multi-head self-attention method proposed in [100] implemented five convolutional layers and an attention layer using MFCC for emotion recognition. The model is trained on IEMOCAP, and only four emotions, happy, sad, angry, and neutral, are considered. Around 76.18% of weighted accuracy and 76.36% of unweighted accuracy is obtained.

The brain emotional learning model (BEL) [107] is a simple structure algorithm in modeling the brain's limbic system and is contemplated as a significant part of the emotional process. BEL model has lower complexity than other conventional neural networks and hence is utilized in the prediction [108] and classification process [109]. The genetic algorithm (GA) is an adaptive algorithm having superior parallel processing ability and is used with the BEL model (GA-BEL) for the optimization of weights of the BEL model in the SER system [84]. The Low-Level Descriptors (LDDs) of MFCC related features are extracted. PCA, LDA, and PCA+LDA have been utilized for dimension reduction of the feature set. Additionally, the BEL model and GA-BEL model's performance is compared, and the BEL model acquired lower accuracy than the GA-BEL model suggesting the latter is feasible for the SER system.

IV. CHALLENGES

As we might have thought lately, SER is no longer a peripheral issue. In the last decade, the research in SER had become a significant endeavor in HCI and speech processing. The demand for this technology can be reflected by the enormous research being carried out in SER. Human and machine speech recognition have had large differences since, which presents tremendous difficulty in this subject, primarily the blend of knowledge from interdisciplinary fields, especially in SER, applied psychology, and human-computer interface.

One of the main issues is the difficulty of defining the meaning of emotions precisely. Emotions are usually blended and less comprehendible. The collection of databases is a clear reflection of the lack of agreement on the definition of emotions. However, if we consider the everyday interaction between humans and computers, we may see that emotions are voluntary. Those variations are significantly intense as these might be concealed, blended, or feeble and barely recognizable instead of being more prototypical features. Discussing the above facts, we may conclude that additional acoustic features need to be scrutinized to simplify emotion recognition.

One more challenge is handling the regularly co-occurring additive noise involving convolute distortion (emerging from a more affordable receiver or other information obtaining devices) and meddling speakers (emerging from background). The various methodology utilized to record elicited emotional speech, enacted emotional speech, and authentic, spontaneous emotional speech must be unique to each other. Recording certified emotion raises a moral issue, just as challenges control recording circumstance and emotional labeling. A broadly acknowledged recording convention is a deficit for the recording of elicited emotion.

Another challenge is in applying a reduction in dimensionality and feature selection. Feature selection is costlier and unfeasible because of the enhancement's intricacy that focuses on an appropriate feature subset between the large set of features, particularly when utilizing the wrapper techniques. There is an elective strategy that can be utilized, known as filter-based component determination techniques. They are not founded on classification decision however consider different qualities like entropy and correlation. The filter has been recently proved to be more helpful for high-resolution data. It comes with a setback; however, these are not appropriate for a wide range of classifiers. Likewise, the feature selection cut-off points may prompt ignoring some ''significant'' data involved in un-selected features like in CNN.

The problems arise at various stages, including at the time of labeling the utterances. After the utterances are recorded, the speech data is labeled by human annotators. However, there is no doubt that the speaker's actual emotion might vary from the one perceived by the human annotator. Even for human annotators, the recognition rates stay lower than 90%. It is believed that it also depends on both context and content of speech, what the human annotators can infer. SER is affected by culture and language also. Various works have been put forward on cross-language SER that show the ongoing systems and features' insufficiency.

Classification is one of the crucial processes in the SER system as it depends on the classifier's ability to interpret the results accurately generated by the respective algorithm. There are various challenges related to the classifiers, like the deep learning classifier CNN is significantly slower due to max-pooling and thus takes a lot of time for the training process. Traditional classifiers such as kNN, Decision Tree, and SVM take a larger amount of time to process the larger datasets.

Additionally, during the neural network training, there are chances that neurons become co-dependent on each other, and as a result of which their weights affect the organization process of other neurons. It causes these neurons to get specialized in training data, but the performance gets degraded when test data is provided. Hence, resulting in an overfitting

problem. Classifiers like CNN, RNN, and DNN are very notorious for overfitting problems.

We have already discussed various challenges, but not the most ignored challenge, of multi-speech signals. The SER system itself must choose the signal on which the focus should be done. Despite that, this could be controlled by another algorithm, which is the speech separation algorithm at the preprocessing stage itself. The ongoing frameworks nevertheless fail to recognize this issue.

V. CONCLUSION

The capability to drive speech communication using programmable devices is currently in research progress, even if human beings could systematically achieve this errand. The focus of SER research is to design proficient and robust methods to recognize emotions. In this paper, we have offered a precise analysis of SER systems. It makes use of speech databases that provide the data for the training process. Feature extraction is done after the speech signal has undergone preprocessing. The SER system commonly utilizes prosodic and spectral acoustic features such as formant frequencies, spectral energy of speech, speech rate and fundamental frequencies, and some feature extraction techniques like MFCC, LPCC, and TEO features. Two classification algorithms are used to recognize emotions, traditional classifiers, and deep learning classifiers, after the extraction of features. Even if there is much work done using traditional techniques, the turning point in SER is deep learning techniques. Although SER has come far ahead than it was a decade ago, there are still several challenges to work on. Some of them are highlighted in this paper. The system needs more robust algorithms to improve the performance so that the accuracy rates increase and thrive on finding an appropriate set of features and efficient classification techniques to enhance the HCI to a greater extend.

REFERENCES

[1] J. Li, L. Deng, R. Haeb-Umbach, and Y. Gong, "Fundamentals of speech recognition," in *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Waltham, MA, USA: Academic, 2016.

[2] J. Hook, F. Noroozi, O. Toygar, and G. Anbarjafari, "Automatic speech based emotion recognition using paralinguistics features," *Bull. Polish Acad. Sci. Tech. Sci.*, vol. 67, no. 3, pp. 1–10, 2019, doi: 10.24425/bpasts.2019.129647.

[3] M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, A. M. Kulkarni, and J. A. Christensen, "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers," *Hearing Res.*, vol. 322, pp. 151–162, Apr. 2015, doi: 10.1016/j.heares.2014.10.003.

[4] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, Apr. 2018, doi: 10.1145/3129340.

[5] F. W. Smith and S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision," *PLoS ONE*, vol. 13, no. 5, May 2018, Art. no. e0197160, doi: 10.1371/journal.pone.0197160.

[6] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2017, doi: 10.1109/ACCESS.2016.2641480.

[7] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Commun.*, vol. 116, pp. 56–76, Jan. 2020, doi: 10.1016/j.specom.2019.12.001.

[8] N. Yala, B. Fergani, and A. Fleury, "Towards improving feature extraction and classification for activity recognition on streaming data," *J. Ambient Intell. Humanized Comput.*, vol. 8, no. 2, pp. 177–189, Apr. 2017, doi: 10.1007/s12652-016-0412-1.

[9] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[10] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, Mar. 2011, doi: 10.1016/j.patcog.2010.09.020.

[11] R. Munot and A. Nenkova, "Emotion impacts speech recognition performance," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Student Res. Workshop*, 2019, pp. 16–21, doi: 10.18653/v1/n19-3003.

[12] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image Vis. Comput.*, vol. 31, no. 2, pp. 120–136, Feb. 2013, doi: 10.1016/j.imavis.2012.06.016.

[13] S. A. A. Qadri, T. S. Gunawan, M. F. Alghifari, H. Mansor, M. Kartiwi, and Z. Janin, "A critical insight into multi-languages speech emotion databases," *Bull. Elect. Eng. Inform.*, vol. 8, no. 4, pp. 1312–1323, Dec. 2019, doi: 10.11591/eei.v8i4.1645.

[14] M. Swain, A. Routray, and P. Kabisatpathy, "Databases, features and classifiers for speech emotion recognition: A review," *Int. J. Speech Technol.*, vol. 21, no. 1, pp. 93–120, Mar. 2018, doi: 10.1007/s10772-018-9491-z.

[15] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 186–202, Jan. 2015, doi: 10.1016/j.csl.2014.01.003.

[16] S. Basu, J. Chakraborty, A. Bag, and M. Aftabuddin, "A review on emotion recognition using speech," in *Proc. Int. Conf. Inventive Commun. Comput. Technol.*, 2017, pp. 109–114, doi: 10.1109/ICICCT.2017.7975169.

[17] B. M. Nema and A. A. Abdul-Kareem, "Preprocessing signal for speech emotion recognition," *Al-Mustansiriyah J. Sci.*, vol. 28, no. 3, p. 157, Jul. 2018, doi: 10.23851/mjs.v28i3.48.

[18] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.

[19] P. Jackson and S. Haq, "Surrey audio-visual expressed emotion (savee) database," Univ. Surrey, Guildford, U.K., 2014. [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. 10th IEEE Int. Conf. Workshops Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–8, doi: 10.1109/FG.2013.6553805.

[21] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 5–17, Jan. 2012, doi: 10.1109/T-AFFC.2011.20.

[22] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proc. 22nd Int. Conf. Data Eng. Workshops*, Apr. 2006, p. 8, doi: 10.1109/ICDEW.2006.145.

[23] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/s10579-008-9076-6.

[24] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: The FAU Aibo emotion corpus," in *Proc. Workshop Corpora Res. Emotion Affect LREC*, 2008, pp. 1–4.

[25] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affect. Comput.*, vol. 8, no. 3, pp. 300–313, Jul. 2017, doi: 10.1109/TAFFC.2016.2553038.

[26] K. S. Rao, S. G. Koolagudi, and R. R. Vempada, "Emotion recognition from speech using global and local prosodic features," *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, Jun. 2013, doi: 10.1007/s10772-012-9172-2.

[27] Z. Esmaileyan, "A database for automatic persian speech emotion recognition: Collection, processing and evaluation," *Int. J. Eng.*, vol. 27, no. 1, pp. 79–90, Jan. 2014, doi: 10.5829/idosi.ije.2014.27.01a.11.

[28] A. B. Kandali, A. Routray, and T. K. Basu, "Emotion recognition from Assamese speeches using MFCC features and GMM classifier," in *Proc. IEEE Region 10 Conf.*, Nov. 2008, pp, 1–5, doi: 10.1109/TENCON.2008.4766487.

[29] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1145–1154, Jul. 2006, doi: 10.1109/TASL.2006.876113.

[30] C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, and G. Richard, "Fear-type emotions of the SAFE corpus: Annotation issues," in *Proc. 5th Int. Conf. Lang. Resour. Eval. (LREC)*, Genoa, Italy, 2006, pp. 1099–1104.

[31] C. Brester, E. Semenkin, and M. Sidorov, "Multi-objective heuristic feature selection for speech-based multilingual emotion recognition," *J. Artif. Intell. Soft Comput. Res.*, vol. 6, no. 4, pp. 243–253, Oct. 2016, doi: 10.1515/jaiscr-2016-0018.

[32] D. Pravena and D. Govind, "Development of simulated emotion speech database for excitation source analysis," *Int. J. Speech Technol.*, vol. 20, no. 2, pp. 327–338, Jun. 2017, doi: 10.1007/s10772-017-9407-3.

[33] T. Ozseven, "Evaluation of the effect of frame size on speech emotion recognition," in *Proc. 2nd Int. Symp. Multidisciplinary Stud. Innov. Technol.*, Oct. 2018, pp. 1–4, doi: 10.1109/ISMSIT.2018.8567303.

[34] H. Beigi, *Fundamentals of Speaker Recognition*. New York, NY, USA: Springer, 2011.

[35] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 670–674, doi: 10.1145/2964284.2967306.

[36] N. Banda, A. Engelbrecht, and P. Robinson, "Feature reduction for dimensional emotion recognition in human-robot interaction," in *Proc. IEEE Symp. Ser. Comput. Intell.*, Dec. 2015, pp. 803–810, doi: 10.1109/SSCI.2015.119.

[37] Y. Gao, B. Li, N. Wang, and T. Zhu, "Speech emotion recognition using local and global features," in *Proc. Int. Conf. Brain Inform.*, 2017, pp. 3–13, doi: 10.1007/978-3-319-70772-3_1.

[38] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, and D. Mürbe, "Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall," *Biomech. Model. Mechanobiol.*, vol. 14, no. 4, pp. 719–733, Aug. 2015, doi: 10.1007/s10237-014-0632-2.

[39] D. Gupta, P. Bansal, and K. Choudhary, "The state of the art of feature extraction techniques in speech recognition," in *Speech and Language Processing for Human-Machine Communications*. Singapore: Springer, 2018, doi: 10.1007/978-981-10-6626-9_22.

[40] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system," in *Proc. 6th Int. Conf.-Cloud Syst. Big Data Eng. (Confluence)*, Jan. 2016, pp. 493–497, doi: 10.1109/CONFLUENCE.2016.7508170.

[41] A. Guidi, C. Gentili, E. P. Scilingo, and N. Vanello, "Analysis of speech features and personality traits," *Biomed. Signal Process. Control*, vol. 51, pp. 1–7, May 2019, doi: 10.1016/j.bspc.2019.01.027.

[42] H. M. Teager and S. M. Teager, "Evidence for nonlinear sound production mechanisms in the vocal tract," in *Speech Production and Speech Modelling*. Dordrecht, The Netherlands: Springer, 1990.

[43] J. F. Kaiser, "Some useful properties of Teager's energy operators," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Apr. 1993, pp. 149–152, doi: 10.1109/icassp.1993.319457.

[44] G. Zhou, J. H. L. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001, doi: 10.1109/89.905995.

[45] S. Asiri, "Machine learning classifiers," *Towards Data Sci.*, Jun. 2018. [Online]. Available: https://towardsdatascience.com/machine-learningclassifiers-a5cc4e1b0623

[46] J. Ababneh, "Application of Naïve Bayes, decision tree, and K-nearest neighbors for automated text classification," *Modern Appl. Sci.*, vol. 13, no. 11, p. 31, Oct. 2019, doi: 10.5539/mas.v13n11p31.

[47] P. Patel, A. Chaudhari, R. Kale, and M. A. Pund, "Emotion recognition from speech with Gaussian mixture models & via boosted GMM," *Int. J. Res. Sci. Eng.*, vol. 3, no. 2, pp. 47–53, Mar./Apr. 2017, doi: 10.1109/ICME.2009.5202493.

[48] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMMs," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 1–4.

[49] S. G. Koolagudi, S. Devliyal, B. Chawla, A. Barthwaf, and K. S. Rao, "Recognition of emotions from speech using excitation source features," *Procedia Eng.*, vol. 38, pp. 3409–3417, Sep. 2012, doi: 10.1016/j.proeng.2012.06.394.

[50] S. Mao, D. Tao, G. Zhang, P. C. Ching, and T. Lee, "Revisiting hidden Markov models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 6715–6719, doi: 10.1109/ICASSP.2019.8683172.

[51] T. L. New, S. W. Foo, and L. C. De Silva, "Detection of stress and emotion in speech using traditional and FFT based log energy features," in *Proc. 4th Int. Conf. Inf.*, Dec. 2003, pp. 1619–1623, doi: 10.1109/ICICS.2003.1292741.

[52] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," *Speech Commun.*, vol. 41, no. 4, pp. 603–623, Nov. 2003, doi: 10.1016/S0167-6393(03)00099-2.

[53] Q. Jin, C. Li, S. Chen, and H. Wu, "Speech emotion recognition with acoustic and lexical features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2015, pp. 4749–4753, doi: 10.1109/ICASSP.2015.7178872.

[54] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *J. Smart Home*, vol. 6, no. 2, pp. 101–108, 2012, doi: 10.1109/kst.2013.6512793.

[55] C. Yu, Q. Tian, F. Cheng, and S. Zhang, "Speech emotion recognition using support vector machines," in *Proc. Int. Conf. Comput. Sci. Inf. Eng.*, 2011, pp. 215–220, doi: 10.1007/978-3-642-21402-8_35.

[56] X. Ke, Y. Zhu, L. Wen, and W. Zhang, "Speech emotion recognition based on SVM and ANN," *Int. J. Mach. Learn. Comput.*, vol. 8, no. 3, pp. 198–202, Jun. 2018, doi: 10.18178/ijmlc.2018.8.3.687.

[57] S. Renjith and K. G. Manju, "Speech based emotion recognition in Tamil and Telugu using LPCC and hurst parameters—A comparitive study using KNN and ANN classifiers," in *Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT)*, Apr. 2017, pp. 1–6.

[58] E. Yüncü, H. Hacihabiboglu, and C. Bozsahin, "Automatic speech emotion recognition using auditory models with binary decision tree and SVM," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 773–778, doi: 10.1109/ICPR.2014.143.

[59] S. Misra and H. Li, "Noninvasive fracture characterization based on the classification of sonic wave travel times," in *Machine Learning for Subsurface Characterization*. Cambridge, MA, USA: Gulf Professional Publishing, 2020.

[60] A. Khan and U. K. Roy, "Emotion recognition using prosodic and spectral features of speech and Naïve Bayes classifier," in *Proc. Int. Conf. Wireless Commun., Signal Process., Netw. (WiSPNET)*, 2018, pp. 1017–1021, doi: 10.1109/WiSPNET.2017.8299916.

[61] M. F. Alghifari, T. S. Gunawan, and M. Kartiwi, "Speech emotion recognition using deep feedforward neural network," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 10, no. 2, p. 554, May 2018, doi: 10.11591/ijeecs.v10.i2.pp554-561.

[62] I. J. Tashev, Z. Q. Wang, and K. Godin, "Speech emotion recognition based on Gaussian Mixture Models and Deep Neural Networks," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2017, pp. 1–4, doi: 10.1109/ITA.2017.8023477.

[63] H. Wang and B. Raj, "On the origin of deep learning," 2017, *arXiv:1702.07800*. [Online]. Available: https://arxiv.org/abs/1702.07800

[64] Z. Gan, R. Henao, D. Carlson, and L. Carin, "Learning deep sigmoid belief networks with data augmentation," in *Proc. Artif. Intell. Statist.*, 2015, pp. 268–276.

[65] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, and S. Yang, "Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services," *Softw., Pract. Exper.*, vol. 47, no. 8, pp. 1127–1138, Aug. 2017, doi: 10.1007/978-3-319-39601-9_5.

[66] L. Zhu, L. Chen, D. Zhao, J. Zhou, and W. Zhang, "Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN," *Sensors*, vol. 17, no. 7, p. 1694, Jul. 2017, doi: 10.3390/s17071694.

[67] J. Brownlee, "A gentle introduction to exploding gradients in neural networks," *Mach. Learn. Mastery*, Aug. 2017. [Online]. Available: https://machinelearningmastery.com/transfer-learning-for-deep-learning/

[68] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1–4.
[69] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2017, pp. 2227–2231, doi: 10.1109/ICASSP.2017.7952552.
[70] S. Pranjal, "Essentials of deep learning?: Introduction to long short term memory," *Anal. Vidhya*, Dec. 2017. [Online]. Available: http://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/
[71] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. Int. Conf. Platform Technol. Service*, 2017, pp. 1–5, doi: 10.1109/PlatCon.2017.7883728.
[72] R. B. Lanjewar, S. Mathurkar, and N. Patel, "Implementation and comparison of speech emotion recognition system using Gaussian mixture model (GMM) and K-nearest neighbor (K-NN) techniques," *Procedia Comput. Sci.*, vol. 49, pp. 50–57, 2015, doi: 10.1016/j.procs.2015.04.226.
[73] S. K. Bhakre and A. Bang, "Emotion recognition on the basis of audio signal using Naïve Bayes classifier," in *Proc. Int. Conf. Adv. Comput., Commun. Inform.*, Sep. 2016, pp. 2363–2367, doi: 10.1109/ICACCI.2016.7732408.
[74] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1162–1171, Nov. 2011, doi: 10.1016/j.specom.2011.06.004.
[75] L. Li, Y. Zhao, D. Jiang, Y. Zhang, F. Wang, I. Gonzalez, E. Valentin, and H. Sahli, "Hybrid deep neural network–hidden Markov model (DNN-HMM) based speech emotion recognition," in *Proc. Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 312–317, doi: 10.1109/ACII.2013.58.
[76] M. Swain, S. Sahoo, A. Routray, P. Kabisatpathy, and J. N. Kundu, "Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition," *Int. J. Speech Technol.*, vol. 18, no. 3, pp. 387–393, Sep. 2015, doi: 10.1007/s10772-015-9275-7.
[77] A. Shaw, R. Kumar, and S. Saxena, "Emotion recognition and classification in speech using artificial neural networks," *Int. J. Comput. Appl.*, vol. 145, no. 8, pp. 5–9, Jul. 2016, doi: 10.5120/ijca2016910710.
[78] W. Lim, D. Jang, and T. Lee, "Speech emotion recognition using convolutional and Recurrent Neural Networks," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Dec. 2017, pp. 1–4, doi: 10.1109/APSIPA.2016.7820699.
[79] P. Harár, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *Proc. 4th Int. Conf. Signal Process. Integr. Netw.*, Feb. 2017, pp. 137–140, doi: 10.1109/SPIN.2017.8049931.
[80] G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random deep belief networks for recognizing emotions from speech signals," *Comput. Intell. Neurosci.*, vol. 2017, Mar. 2017, Art. no. 1945630, doi: 10.1155/2017/1945630.
[81] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *Int. J. Speech Technol.*, vol. 20, no. 2, pp. 239–246, Jun. 2017, doi: 10.1007/s10772-017-9396-2.
[82] P.-W. Hsiao and C.-P. Chen, "Effective attention mechanism in dynamic models for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2526–2530, doi: 10.1109/ICASSP.2018.8461431.
[83] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "Speech emotion recognition: Methods and cases study," in *Proc. ICAART*, 2018, pp. 175–182, doi: 10.5220/0006611601750182.
[84] Z.-T. Liu, Q. Xie, M. Wu, W.-H. Cao, Y. Mei, and J.-W. Mao, "Speech emotion recognition based on an improved brain emotion learning model," *Neurocomputing*, vol. 309, pp. 145–156, Oct. 2018, doi: 10.1016/j.neucom.2018.05.005.
[85] G. Liu, W. He, and B. Jin, "Feature fusion of speech emotion recognition based on deep learning," in *Proc. Int. Conf. Netw. Infrastruct. Digit. Content*, Aug. 2018, pp. 193–197, doi: 10.1109/ICNIDC.2018.8525706.
[86] M. Neumann and N. T. Vu, "Improving speech emotion recognition with unsupervised representation learning on unlabeled speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2019, pp. 7390–7394, doi: 10.1109/ICASSP.2019.8682541.
[87] L. Sun, S. Fu, and F. Wang, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," *EURASIP J. Audio, Speech, Music Process.*, vol. 2019, no. 1, Dec. 2019, pp. 1–14, doi: 10.1186/s13636-018-0145-5.
[88] M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, "Human emotion recognition using deep belief network architecture," *Inf. Fusion*, vol. 51, pp. 10–18, Nov. 2019, doi: 10.1016/j.inffus.2018.10.009.
[89] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on DNN-decision tree SVM model," *Speech Commun.*, vol. 115, pp. 29–37, Dec. 2019, doi: 10.1016/j.specom.2019.10.004.
[90] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *Proc. IEEE Spoken Lang. Technol. Workshop*, Dec. 2019, pp. 112–118, doi: 10.1109/SLT.2018.8639583.
[91] S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, "Direct modelling of speech emotion from raw speech," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, 2019, pp. 3920–3924, doi: 10.21437/Interspeech.2019-3252.
[92] H. S. Kumbhar and S. U. Bhandari, "Speech emotion recognition using MFCC features and LSTM network," in *Proc. 5th Int. Conf. Comput., Commun., Control Automat.*, Sep. 2019, pp. 1–3, doi: 10.1109/ICCUBEA47591.2019.9129067.
[93] D. Bharti and P. Kukana, "A hybrid machine learning model for emotion recognition from speech signals," in *Proc. Int. Conf. Smart Electron. Commun.*, Sep. 2020, pp. 491–496, doi: 10.1109/ICOSEC49089.2020.9215376.
[94] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cleder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning*. Rijeka, Croatia: InTech, 2020.
[95] Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun.*, vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005.
[96] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM," *IEEE Access*, 2020, doi: 10.1109/ACCESS.2020.2990405.
[97] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, "Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks," in *Proc. 6th Int. Conf. Wireless Telematics*, Sep. 2020, pp. 1–6, doi: 10.1109/icwt50448.2020.9243622.
[98] Z. Huijuan, Y. Ning, and W. Ruchuan, "Coarse-to-fine speech emotion recognition based on multi-task learning," *J. Signal Process. Syst.*, vol. 93, nos. 2–3, pp. 299–308, Mar. 2021, doi: 10.1007/s11265-020-01538-x.
[99] M. S. Fahad, A. Deepak, G. Pradhan, and J. Yadav, "DNN-HMM-based speaker-adaptive emotion recognition using MFCC and epoch-based features," *Circuits, Syst., Signal Process.*, vol. 40, no. 1, pp. 466–489, Jan. 2021, doi: 10.1007/s00034-020-01486-8.
[100] M. Xu, F. Zhang, and S. U. Khan, "Improve accuracy of speech emotion recognition with attention head fusion," in *Proc. 10th Annu. Comput. Commun. Workshop Conf.*, Jan. 2020, pp. 1058–1064, doi: 10.1109/CCWC47524.2020.9031207.
[101] S. A. Firoz, S. A. Raji, and A. P. Babu, "Automatic emotion recognition from speech using artificial neural networks with gender-dependent databases," in *Proc. Int. Conf. Adv. Comput., Control, Telecommun. Technol.*, Dec. 2009, pp. 162–164, doi: 10.1109/ACT.2009.49.
[102] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech*, 2018, pp. 3688–3692, doi: 10.21437/Interspeech.2018-1811.
[103] Mustaqeem and S. Kwon, "A CNN-assisted enhanced audio signal processing for speech emotion recognition," *Sensors*, 2020, doi: 10.3390/s20010183.
[104] C. Balkenius and J. MorEn, "Emotional learning: A computational model of the amygdala," *Cybern. Syst.*, vol. 32, no. 6, pp. 611–636, Sep. 2001, doi: 10.1080/01969720118947.
[105] E. Lotfi, A. Khosravi, M. R. Akbarzadeh, and S. Nahavandi, "Wind power forecasting using emotional neural networks," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2014, pp. 311–316, doi: 10.1109/SMC.2014.6973926.
[106] Y. Mei, G. Tan, and Z. Liu, "An improved brain-inspired emotional learning algorithm for fast classification," *Algorithms*, vol. 10, no. 2, p. 70, Jun. 2017, doi: 10.3390/a10020070.
[107] C. Balkenius and J. Morén, "Emotional learning: A computational model of the amygdala," *Cybern. Syst.*, vol. 32, no. 6, pp. 611–636, 2001, doi: 10.1080/01969720118947.

[108] L. Guo, L. Wang, J. Dang, L. Zhang, and H. Guan, "A feature fusion method based on extreme learning machine for speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2018, pp. 2666–2670, doi: 10.1109/ICASSP.2018.8462219.
[109] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in *Proc. 8th Int. Conf. Comput., Commun. Netw. Technol.*, Jul. 2017, pp. 1–5, doi: 10.1109/ICCCNT.2017.8204149.

TAIBA MAJID WANI received the bachelor's degree in electronics and communication engineering from the Islamic University of Science and Technology, Kashmir. She is currently pursuing the M.Sc. degree in communication engineering with International Islamic University Malaysia (IIUM). Her research interests include speech processing, speech emotion recognition, and deep learning. She is also an IEEE IIUM Student Branch Member.

TEDDY SURYA GUNAWAN (Senior Member, IEEE) received the B.Eng. degree *(cum laude)* in electrical engineering from the Institut Teknologi Bandung (ITB), Indonesia, in 1998, the M.Eng. degree from the School of Computer Engineering, Nanyang Technological University, Singapore, in 2001, and the Ph.D. degree from the School of Electrical Engineering and Telecommunications, The University of New South Wales, Australia, in 2007. His research interests include speech and audio processing, biomedical signal processing and instrumentation, image and video processing, and parallel computing. He was awarded the Best Researcher Award from IIUM, in 2018. He was a Chairman of the IEEE Instrumentation and Measurement Society–Malaysia Section (2013, 2014, and 2020), a Professor (since 2019), the Head of Department (from 2015 to 2016) with the Department of Electrical and Computer Engineering, and the Head of Programme Accreditation, and the Quality Assurance for Faculty of Engineering (from 2017 to 2018), International Islamic University Malaysia. He has been a Chartered Engineer (IET, U.K.) and Insinyur Profesional Madya (PII, Indonesia) since 2016, a registered ASEAN Engineer since 2018, and an ASEAN Chartered Professional Engineer since 2020.

SYED ASIF AHMAD QADRI received the B.Tech. degree in computer science and engineering from Baba Ghulam Shah Badshah University, Kashmir, and the M.Sc. degree in computer and information engineering from International Islamic University Malaysia (IIUM). He worked as a Graduate Research Assistant with the Department of Electrical and Computer Engineering, IIUM. His research interests include speech processing, machine learning, artificial intelligence, and the Internet of Things.

MIRA KARTIWI (Member, IEEE) is currently an Associate Professor with the Department of Information Systems, Kulliyyah of Information and Communication Technology, and the Deputy Director of E-Learning with the Centre for Professional Development, International Islamic University Malaysia (IIUM). She is also an Experienced Consultant, specializing in health, financial, and manufacturing sectors. Her areas of expertise include health informatics, e-commerce, data mining, information systems strategy, business process improvement, product development, marketing, delivery strategy, workshop facilitation, training, and communications. She was one of a recipients of the Australia Postgraduate Award (APA) in 2004. For her achievement in research, she was awarded the Higher Degree Research Award for Excellence in 2007. She has also been appointed as an Editorial Board Member in local and international journals to acknowledge her expertise.

ELIATHAMBY AMBIKAIRAJAH (Senior Member, IEEE) received the B.Sc. degree (Hons.) in engineering from the University of Sri Lanka and the Ph.D. degree in signal processing from Keele University, U.K. His key publications led to his repeated appointment as a short-term Invited Research Fellow with the British Telecom Laboratories, U.K., for ten years from 1989 to 1999. After previously serving as the Head of the School of Electrical Engineering and Telecommunications, he is currently serving as the Acting Deputy Vice-Chancellor Enterprise with the University of New South Wales (UNSW), Australia, from 2009 to 2019. He has authored and coauthored approximately 300 journal and conference papers. His research interests include speaker and language recognition, emotion detection, and biomedical signal processing. He is a Fellow and a Chartered Engineer of the IET U.K., and an Engineers Australia (EA). He is also a Life Member of APSIPA. He was a recipient of many competitive research grants. He was an APSIPA Distinguished Lecturer from 2013 to 2014.