IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 25, 2019, accepted August 5, 2019, date of publication August 19, 2019, date of current version September 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936124

Speech Emotion Recognition Using Deep Learning Techniques: A Review

RUHUL AMIN KHALIL[1], EDWARD JONES[2], MOHAMMAD INAYATULLAH BABAR[1], TARIQULLAH JAN[1], MOHAMMAD HASEEB ZAFAR[3], AND THAMER ALHUSSAIN[4]

[1]Department of Electrical Engineering, Faculty of Electrical and Computer Engineering, University of Engineering and Technology, Peshawar, Pakistan
[2]Department of Electrical and Electronics Engineering, National University of Ireland, Galway, H91 TK33 Ireland
[3]Department of Information Technology, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia
[4]Department of E-Commerce, Saudi Electronic University (SEU), Riyadh 11637, Saudi Arabia

Corresponding author: Ruhul Amin Khalil (ruhulamin@uetpeshawar.edu.pk)

This article was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah. The authors therefore, acknowledge with thanks DSR for technical and financial support.

ABSTRACT Emotion recognition from speech signals is an important but challenging component of Human-Computer Interaction (HCI). In the literature of speech emotion recognition (SER), many techniques have been utilized to extract emotions from signals, including many well-established speech analysis and classification techniques. Deep Learning techniques have been recently proposed as an alternative to traditional techniques in SER. This paper presents an overview of Deep Learning techniques and discusses some recent literature where these methods are utilized for speech-based emotion recognition. The review covers databases used, emotions extracted, contributions made toward speech emotion recognition and limitations related to it.

INDEX TERMS Speech emotion recognition, deep learning, deep neural network, deep Boltzmann machine, recurrent neural network, deep belief network, convolutional neural network.

I. INTRODUCTION

Emotion recognition from speech has evolved from being a niche to an important component for Human-Computer Interaction (HCI) [1]–[3]. These systems aim to facilitate the natural interaction with machines by direct voice interaction instead of using traditional devices as input to understand verbal content and make it easy for human listeners to react [4]–[6]. Some applications include dialogue systems for spoken languages such as call center conversations, onboard vehicle driving system and utilization of emotion patterns from the speech in medical applications [7]. Nonetheless, there are many problems in HCI systems that still need to be properly addressed, particularly as these systems move from lab testing to real-world application [8]–[10]. Hence, efforts are required to effectively solve such problems and achieve better emotion recognition by machines.

Determining the emotional state of humans is an idiosyncratic task and may be used as a standard for any emotion recognition model [11]. Amongst the numerous models used for categorization of these emotions, a discrete emotional approach is considered as one of the fundamental approaches. It uses various emotions such as anger, boredom, disgust, surprise, fear, joy, happiness, neutral and sadness [12], [13]. Another important model that is used is a three-dimensional continuous space with parameters such as arousal, valence, and potency.

The approach for speech emotion recognition (SER) primarily comprises two phases known as feature extraction and features classification phase [14]. In the field of speech processing, researchers have derived several features such as source-based excitation features, prosodic features, vocal traction factors, and other hybrid features [15]. The second phase includes feature classification using linear and non-linear classifiers. The most commonly used linear classifiers for emotion recognition include Bayesian Networks (BN) or the Maximum Likelihood Principle (MLP) and Support Vector Machine (SVM). Usually, the speech signal is considered to be non-stationary. Hence, it is considered that non-linear classifiers work effectively for SER. There are many non-linear classifiers available for SER, including Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) [16]. These are widely used for classification of information that is derived from basic level features.

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

Energy-based features such as Linear Predictor Coefficients (LPC), Mel Energy-spectrum Dynamic Coefficients (MEDC), Mel-Frequency Cepstrum Coefficients (MFCC) and Perceptual Linear Prediction cepstrum coefficients (PLP) are often used for effective emotion recognition from speech. Other classifiers including K-Nearest Neighbor (KNN), Principal Component Analysis (PCA) and Decision trees are also applied for emotion recognition [17].

Deep Learning has been considered as an emerging research field in machine learning and has gained more attention in recent years [18]. Deep Learning techniques for SER have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning; tendency toward extraction of low-level features from the given raw data, and ability to deal with un-labeled data.

Deep Neural Networks (DNNs) are based on feed-forward structures comprised of one or more underlying hidden layers between inputs and outputs. The feed-forward architectures such as Deep Neural Networks (DNNs) and Convolutional Neural Networks (CNNs) provides efficient results for image and video processing. On the other hand, recurrent architectures such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) are much effective in speech-based classification such as natural language processing (NLP) and SER [18]. Apart from their effective way of classification these models do have some limitations. For instance, the positive aspect of CNNs is to learn features from high-dimensional input data, but on the other hand, it also learns features from small variations and distortion occurrence and hence, requires large storage capability. Similarly, LSTM-based RNNs are able to handle variable input data and model long-range sequential text data.

The organization of this paper is as follows. A review of background for speech-based emotion detection and recognition using traditional classification techniques is given in Section II. Section III reviews the need for deep learning techniques utilized in a different context for SER. In Section IV, different deep learning techniques are discussed on the basis of their layer-wise architecture for SER. Further, Section V provides a summary of the papers based on these deep learning techniques for SER along with detailed discussion and future directions. Finally, concluding remarks are presented in Section VI.

A list of nomenclature used throughout this review paper is provided in Table 1 as follow.

TABLE 1. List of nomenclature used in this review paper.

Nomenclature	Referred to
ABC	Airplane Behavior Corpus
AE	Auto Encoders
ANN	Artificial Neural Network
AVB	Adversarial Variational Bayes
AVEC	Audio/Visual Emotion Challenge
BN	Bayesian Networks
CAM3D	Cohn-Kanade dataset
CAS	Chinese Academy of Science database
CNN	Convolutional Neural Network
ComParE	Computational Paralinguistic challenge
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DCNN	Deep Convolutional Neural Network
DES	Danish Emotional Speech Database
DNN	Deep Neural Networks
eGeMAPS	extended Geneva Minimalistic Acoustic Parameter Set
ELM	Extreme Learning Machine
Emo-DB	Berlin Emotional database
FAU-AEC	FAU Aibo Emotion Corpus
GMM	Gaussian Mixture Model
HCI	Human-Computer Interaction
HMM	Hidden Markov Model
HRI	Human-Robot Interaction
IEMOCAP	Interactive Emotional Dyadic Motion Capture database
KNN	K-Nearest Neighbor
LIF	Localized Invariant Features
LPC	Linear Predictor Coefficients
LSTM	Long-Short Term Memory
MEDC	Mel Energy-spectrum Dynamic Coefficient
MFCC	Mel-Frequency Cepstrum Coefficient
MLP	Maximum Likelihood Principle
MT-SHL-DNN	Multi-Tasking Shared Hidden Layers Deep Neural Network
PCA	Principle Component Analysis
PLP	Perceptual Liner Prediction cepstrum coefficient
RBM	Restricted Boltzmann Machine
RE	Reconstruction-Error-based (RE)
RvNN	Recursive Neural Network
RECOLA	Remote Collaborative and Affective Interactions database
RNN/RCNN	Recurrent Neural Network
SAE	Stacked Auto Encoder
SPAE	Sparse-Auto Encoders
SAVEE	Surrey Audio-Visual Expressed Emotion
SDFA	Salient Discriminative Feature Analysis
SER	Speech Emotion Recognition
SVM	Support Vector Machine
VAE	Variational Auto Encoder

II. TRADITIONAL TECHNIQUES FOR SER

Emotion recognition systems based on digitized speech is comprised of three fundamental components: signal preprocessing, feature extraction, and classification [19]. Acoustic preprocessing such as denoising, as well as segmentation, is carried out to determine meaningful units of the signal [20]. Feature extraction is utilized to identify the relevant features available in the signal. Lastly, the mapping of extracted feature vectors to relevant emotions is carried out by classifiers. In this section, a detailed discussion of speech signal processing, feature extraction, and classification is provided [21]. Also, the differences between spontaneous and acted speech are discussed due to their relevance to the topic [22], [23]. Figure 1 depicts a simplified system utilized for speech-based emotion recognition. In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, and feature selection. The required features are extracted from the preprocessed speech signal and the selection is made from

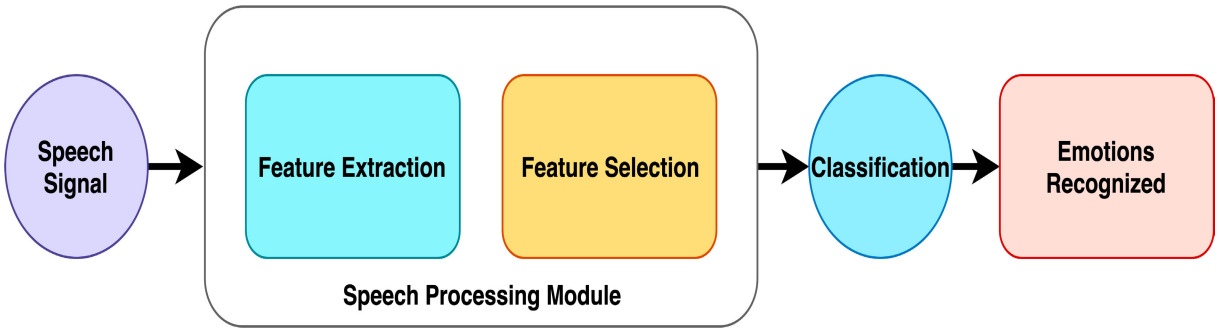

FIGURE 1. Traditional Speech Emotion Recognition System.

the extracted features. Such feature extraction and selection is usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers such as GMM and HMM, etc. are utilized for classification of these features. Lastly, based on feature classification different emotions are recognized.

A. ENHANCEMENT OF INPUT SPEECH DATA IN SER

The input data collected for emotion recognition is often corrupted by noise during the capturing phase [24]. Due to these impairments, the feature extraction and classification become less accurate [25]. This means that the enhancement of the input data is a critical step in emotion detection and recognition systems. In this preprocessing stage, the emotional discrimination is kept, while the speaker and recording variation is eliminated [26].

B. FEATURE EXTRACTION AND SELECTION IN SER

The speech signal after enhancement is characterized into meaningful units called segments [27]. Relevant features are extracted and classified into various categories based on the information extracted. One type of classification is short term classification based on short-period characteristics such as energy, formants and pitch [28]. The other is known as long term classification; mean and standard deviation are two of the often-used long term features [29]. Among prosodic features, the intensity, pitch, rate of spoken words and variance are usually important to identify various types of emotions from the input speech signal [30], [31]. A few of the characteristics based on acoustics emotions of speech are presented in Table 2.

C. MEASURES FOR ACOUSTICS IN SER

Information availability of emotions is encrypted in every aspect of language and the variations in it. The vocal parameters and their relation to emotion recognition are among the most researched topics in this field. Parameters such as intensity, pitch, and rate of spoken words and quality of voice are frequently considered [32]. Often, a straightforward view of emotion is considered, wherein emotions are assumed to exist as discrete categories. These discrete emotions sometimes have relatively clear relationships with acoustic parameters, for example, as indicated in Table 2 for a subset of emotions. Often, the intensity and pitch are correlated to activation, so that the value of intensity increases along with high pitch and gets low with low pitch [33]. Factors that affect the mapping from acoustic variables to emotion include whether the speaker is acting, there are high speaker variations, and the mood or personality of the individual.

TABLE 2. Summarized form of some acoustic variations observed based on emotions.

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

In HCI, emotions are usually spontaneous and are generally not the prototypical discrete emotions, rather they are often weakly expressed, mixed, and hard to distinguish from each other [34]. In the literature, emotional statements are termed as positive and negative based on emotions expressed by an individual [35]. Other experiments show that the listener-based acted emotions are much stronger and accurate than natural emotions, which may suggest that actors exaggerate the expression of emotions. According to the study in [35], the fundamental emotions can be described by areas within the space defined by the axes of arousal and valence are provided in Figure 2. Arousal represents the intensity of calmness or excitement, whereas the valence represents the effect of positivity and negativity in the emotions.

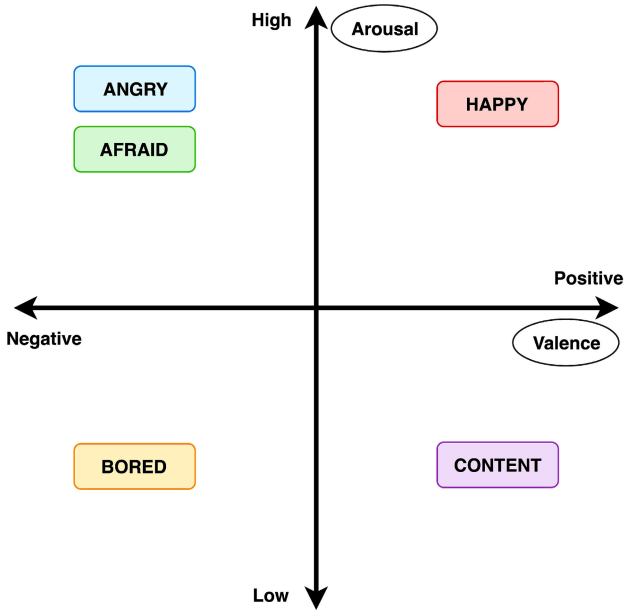

FIGURE 2. A two dimensional basic emotional space.

TABLE 3. Few linear and non-linear classifiers used for SER.

Classifiers	Linear/Non-Linear	References
Bayes Classifier	Linear	[36]
K-Nearest Neighbor classifier	Linear	[36]
GMM classifier	Non-Linear	[37]
HMM classifier	Non-Linear	[38] , [39]
PCA classifier	Linear/Non-Linear	[40]
SVM classifier	Linear/Non-Linear	[41], [42]
ELM classfifier	Linear/Non-Linear	[36]

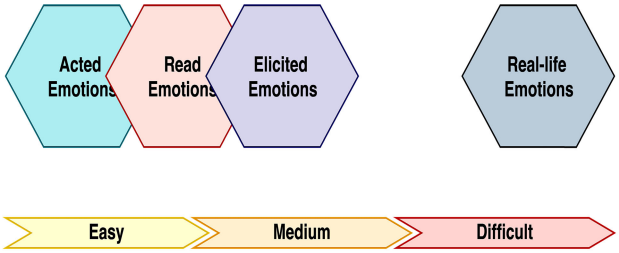

FIGURE 3. Emotion recognition databases and their difficulty level.

D. CLASSIFICATION OF FEATURES IN SER

In literature, various classifiers have been investigated to develop systems such as SER, speech recognition, and speaker verification, to name a few [36]–[39]. On the other hand, the justification for choosing a particular classifier to the specific speech task is often not mentioned in most of the applications. Typically, classifiers are selected on either rule of thumb or empirical evaluation of some indicators as mentioned earlier.

Normally, pattern recognition classifiers used for SER can be broadly be categorized into two main types, namely linear classifiers and non-linear classifiers. Linear classifiers usually perform classification based on object features with a linear arrangement of various objects [40]–[43]. These objects are mostly evaluated in the form of an array termed as a feature vector. In contrast, non-linear classifiers are utilized for object characterization in developing the non-linear weighted combination of such objects.

Table 3 depicts a few traditional linear and non-linear classifiers used for SER.

E. DATABASES USED FOR SER

Speech emotional databases are used by many researchers in a variety of research activities [52]. Quality of the databases utilized and performance achieved are the most important factors in evaluation for emotion recognition. The methods available and objectives in the collection of speech databases vary depending on the motivation for speech systems development. Table 4 provides the characteristics of various freely available emotional speech databases.

For development of emotional speech systems, speech databases are categorized into three main types. The categorization of databases can also be described according to the continuum shown in Figure 3.

- Simulated database:
  In these databases, the speech data has been recorded by well trained and experienced performers [44], [45]. Among all databases this one is considered as the simplest way to obtained the speech-based dataset of various emotions. It is considered that almost 60% of speech databases are gathered by this technique.
- Induced database:
  This is another type of database in which the emotional set is collected by creating an artificial emotional situation [50], [51]. This is done without the knowledge of the performer or speaker. As compared to actor-based database, this is a more naturalistic database. However, an issue of ethics may apply, because the speaker should know that they have been recorded for research-based activities.
- Natural database:
  While most realistic, these databases are hard to obtain due to the difficulty in recognition [52]. Natural emotional speech databases are usually recorded from the general public conversation, call center conversations and so on.

In early 1990s, when the research on speech-based emotion recognition developed in earnest, researchers often commenced with acted databases and later moved to realistic databases [51], [52]. In acted databases, the most commonly used databases are Berlin Emotional Speech Database (EmoDB) and Danish Emotional Speech Database (DES) that

TABLE 4. Characteristics of various free available emotional speech databases.

S.No.	Database	Language	Emotions	Size	Source	Access	References
1.	Berlin emotional database	German	Happiness, sadness, boredom, neutral, disgust and anger	800 utterances	Professional actors	Public and free	[44], [45], [46]
2.	Danish emotional databases	Danish	Anger, sadness, surprise, neutral, and joy	4 actors × 5 emotions	Non-professional actors	Free license	[44], [45]
3.	Interactive Emotional Dyadic Motion Capture	English	Happiness, anger, sadness, frustration, surprise, fear, disgust, excited and neutral state	10 actors (5 male and 5 female)	Professional actors	Free license	[44], [47]
4.	INTERFACE05	English, Spanish, French and Slovenian	Neutral, disgust, fear, joy, sadness	English=186, Spanish=184, French=175 and Slovenien=190 utterances, respectively	Actors	Commercially available	[44], [48], [49]
5.	LDC Emotional Speech and Transcripts	English	Despair, Sadness, Neutral, interest, joy, panic, anger, shame, contempt, elation, pride and cold anger	7 actors × 15 emotions × 10 utterances	Professional actors	Commercially available	[50], [51]

contains the recorded voices of 10 performers.This includes 4 persons for testing, who were asked to speak various sentences in 5 different emotional states. The data comprises the German-Aibo emotion and Smart-Kom data, where the actors' voices are recorded in a laboratory. Additionally, the call center conversations in a fully realistic environment from live recordings have been used.

Literature suggests there is a large variation between the databases in the number of emotions recognized and the number of performers, purpose, and methodology. Speech emotional databases are employed in psychological studies for knowing the patient's behavior as well as in situations where automation in emotion recognition is desired. The system becomes complex and emotion recognition is hard to achieve when the real-time data is employed.

III. NEED OF DEEP LEARNING TECHNIQUES FOR SER

Speech processing usually functions in a straightforward manner on an audio signal [53]. It is considered significant and necessary for various speech-based applications such as SER, speech denoising and music classification. With recent advancements, SER has gained much significance. However, it still requires accurate methodologies to mimic human-like behavior for interaction with human beings [54]. As discussed earlier, a SER system is made up of various components that include feature selection and extraction, feature classification, acoustic modeling, recognition per unit, and most importantly language-based modeling. The traditional SER systems typically incorporate various classification models such as GMMs and HMMs. The GMMs are utilized for illustration of acoustic features of sound units, while, the HMMs are utilized for dealing with temporal variations occurrence in speech signals.

Deep learning methods are comprised of various non-linear components that perform computation on a parallel basis [55]. However, these methods need to be structured with deeper layers of architecture to overcome the limitations of other techniques. Deep learning techniques such as Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (RNN), Deep Belief Network (DBN), Convolutional Neural Networks (CNN) and Auto Encoder (AE) are considered a few of the fundamental deep learning techniques used for SER, that significantly improves the overall performance of the designed system.

Deep learning is an emerging research field in machine learning and has gained much attention in recent years. A few researchers have used DNNs to trained their respective models for SER. Figure 4 depicts the difference between traditional machine learning flow and deep learning flow mechanisms for SER. Table 5 shows a detailed comparative analysis of the traditional algorithms with Deep learning i.e., Deep Convolutional Neural Network(DCNN) algorithm in the context of measuring various emotions using IEMOCAP, Emo-DB and SAVEE datasets and recognized various emotions such as happiness, anger, and sadness [56]. It is deduced that deep learning algorithms perform well in emotion recognition as compared to traditional techniques.

In the next section, the paper aims to discuss various deep learning techniques in the context of SER. These methods provide accurate results as compared to traditional techniques but are computationally complex. This section

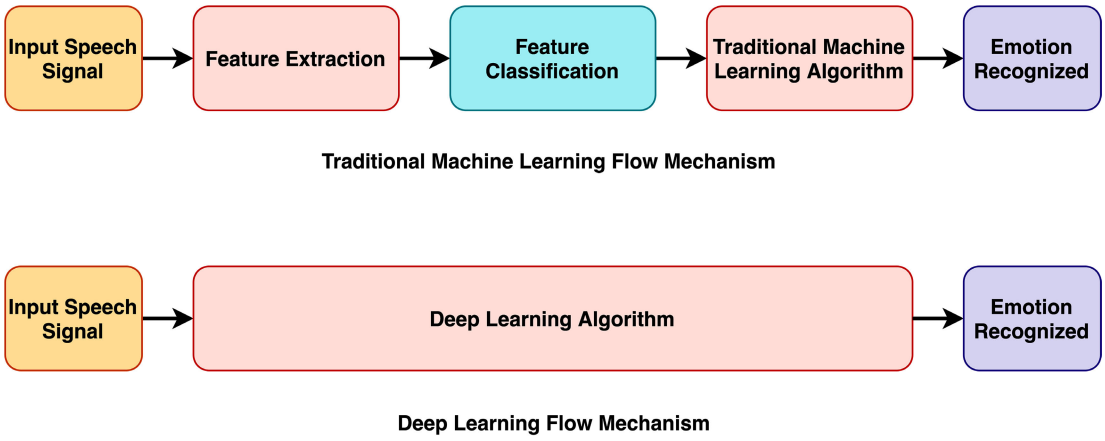

FIGURE 4. Traditional Machine Learning Flow vs Deep Learning Flow.

TABLE 5. Comparative analysis of different classifiers in SER [56].

Algorithms	Anger	Happy	Sad
k-nearest neighbor	93%	55%	77%
Linear discriminant analysis	68%	49%	72%
Support vector machine	74%	70%	93%
Regularized discriminant analysis	83%	73%	97%
Deep Convolutional neural network	99%	99%	96%

provides literature-based support to researchers and readers to assess the HCI feasibility and help them to analyze the user's emotional voice in the given scenario. The real-time applications of these techniques are much more complex, however, emotion recognition from speech input data is a feasible option [56]. These methods do have limitations, however, a combination of two or more of these classifiers results in a new step and possibly improve the detection of emotions.

IV. DEEP LEARNING TECHNIQUES FOR SER

Deep learning is derived from the family of machine learning, which is a broader learning technique for data representation such as emotions [57], [58]. This deep learning can un-supervised, semi-supervised or fully supervised. Figure 5 illustrates a generic layer-wise architecture for DNN.

Currently, deep learning is a fast-growing research area due to its multi-layered structure and efficient results delivery. These research areas include speech emotion recognition, speech and image recognition, natural language processing, and pattern recognition [59], [60]. In this section, various deep learning algorithms such as DBMs, DBNs, CNNs, RNNs, RvNNs, and AEs are discussed.

A. DEEP BOLTZMANN MACHINE (DBM)

DBMs are basically derived from Markov Random fields and are comprised of various hidden layers [61], [62]. These layers are based on randomly chosen variables and coupled with

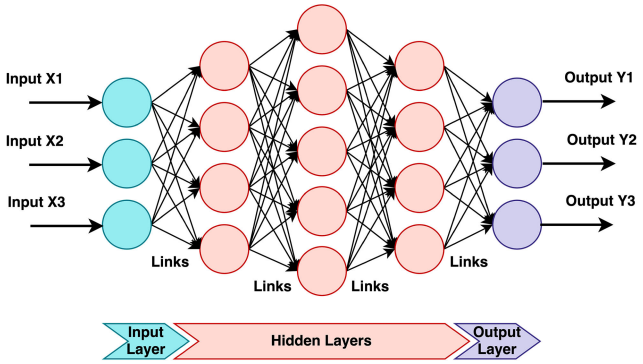

FIGURE 5. Generic layer-wise Deep Neural Network (DNN) Architecture.

stochastic entities. The domain of visible entities is given by $v_i\{0, 1\} \in^C$, with combination of hidden units $h_i^{(1)} \in \{0, 1\}^{b_1}$, $h_i^{(2)} \in \{0, 1\}^{C_2}, \cdots, h_i^{(L)} \in \{0, 1\}^{C_L}$ as shown in Figure 6 (a). Contrary, in a Restricted Boltzmann Machine (RBM), there is no inter-connection between the entities of the same layers. A generalized three-layer RBM network is shown in Figure 6 (b). The probability allotted to each vector component in v_i is given by

$$p(v_i) = \frac{1}{Z}\sum_{h_i}\{e^{\sum abWab^{(1)}v_i hb^{(1)}} + e^{\sum bcWbc^{(2)}h_b^{(2)}h_i^{(2)}} + e^{\sum cdWcd^{(3)}h_c^{(3)}h_i^{(3)}}\} \quad (1)$$

where $h = \{h_i^{(1)}, h_i^{(2)}, h_i^{(3)}\}$ represent the set of hidden layer entities and $\theta = \{W_i^{(1)}, W_i^{(2)}, W_i^{(3)}\}$ corresponds to symmetric interaction between visible and hidden units. This further denotes the visible-hidden and hidden-hidden vector interaction. If $W_i^{(2)} = W_i^{(3)} = 0$, the system is termed as RBM.

The main advantage of DBM is its tendency to learn quickly and provide efficient representation. It achieves this by layer to layer pre-training [63]. This is the reason that

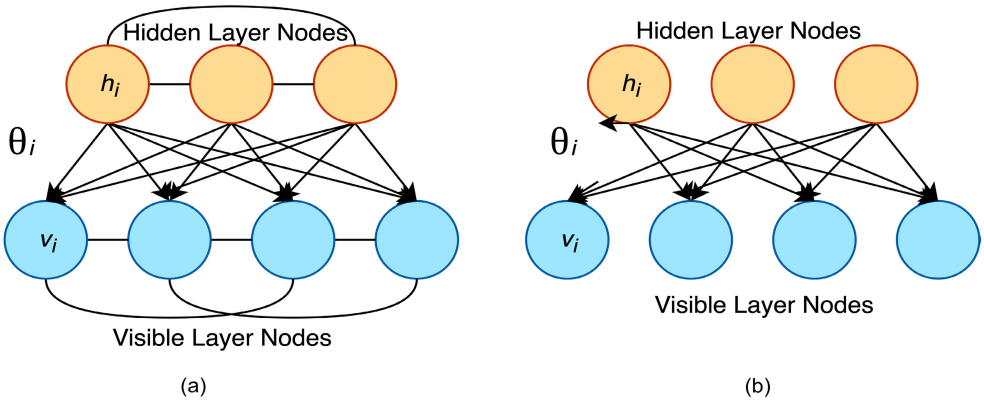

FIGURE 6. Graphical representation of (a) Deep boltzmann machine (DBM) and (b) Restricted boltzmann machine (RBM).

DBM can provide better results for emotion recognition when speech is provided as input. Along with this, DBM has some disadvantages as well, such as restricted effectiveness in certain scenarios [64].

B. RECURRENT NEURAL NETWORK (RNN)

RNN is a branch of neural network based on sequential information, where the outputs and inputs are interdependent [65]. Usually, this interdependency is useful in predicting the future state of the input. RNNs like CNNs require memory to store the overall information obtained in the sequential process of deep learning modeling, and generally works efficiently only for a few back-propagation steps. Figure 7 depicts the basic RNN architecture.

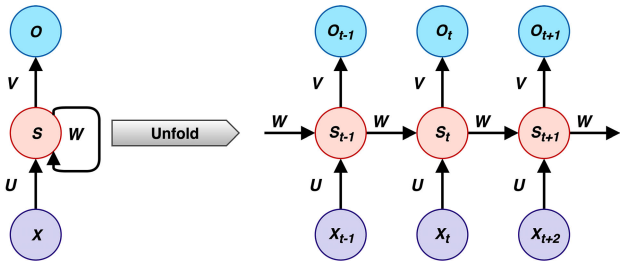

FIGURE 7. Basic architecture of Recurrent Neural Network.

where x_t is the input, s_t is the underlaying hidden state, and o_t is the output at time step t. The U, V, W are known as parameters for hidden matrices and their values may varies for every time step. The hidden state is calculated as $S_t = f(U_{(x_t)} + W_{s_{(t-1)}})$. According to [66], RNNs are suitable for speech emotion recognition due to short-time level framing for acoustic features.

The main problem that affects the overall performance of the RNN is its sensitivity towards the disappearance of the gradients [67]. In other words, the gradients may decay exponentially during the training phase and get multiplied with lots of small or large derivatives. However, this sensitivity gets reduced over a while and results in forgetting the inputs provided at the initial level. To avoid such a situation, Long Short-Term Memory (LSTM) is utilized for providing a block between the recurrent connections. Each block of memory stores the network temporal states, and include gated units for controlling the inflow of new information. The residual connections are usually very deep and hence useful for reducing the gradient issue.

C. RECURSIVE NEURAL NETWORK (RvNN)

RvNN is a hierarchical deep learning technique with no dependency on the tree-structured input sequence on RvNN is a hierarchical deep learning technique with no dependency on the tree-structured input sequence [68]. It can easily learn the parse tree of the provided data by dividing the input into small chunks. Its governing equation is provided in (2) and Figure 8 depicts a RvNN architecture.

$$p_{1,2} = \tanh(W[c_1; c_2]) \tag{2}$$

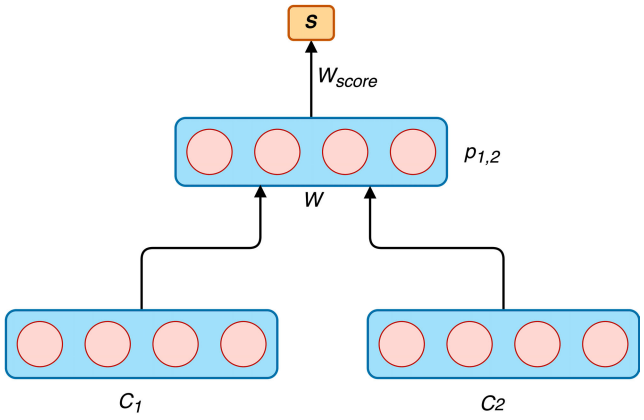

FIGURE 8. Basic architecture of Recursive Neural Network.

where W is designed as $n \times 2n$ weighted matrix. This structured network is well utilized for syntactic parsing for natural language processing, speech processing, speech emotion and pattern recognition in audio and visual input data.

RvNN is mostly used for natural language processing but its architecture is able to handle different modalities such as SER and speech recognition. According to the study in [69], the RvNN can be used for the classification of natural language sentences as well as natural image processing. It initially computes the overall score of the possible pair for merging them in pairs and to build a syntactic tree. The pair of highest score is further combined with a vector known as compositional vector. Once the pair gets merged, the RvNN then generates multiple units, the region representing vectors, and the classification labels.

D. DEEP BELIEF NETWORK (DBN)

DBN is much more complicated in structure and is built from cascaded RBM structures [70]. DBN is an extension of RBMs, in which RBMs are trained layer to layer in a bottom-up manner. The DBNs are usually used for speech emotion recognition due to their ability to learn the recognition parameters efficiently, no matter how a large number of parameters. It also avoids the non-linearity in layers [71]. DBNs are used to tackle slow speed localized problems using back propagation algorithms during training. Figure 9 represents the layer-wise architecture of the DBN in which the RBMs are trained and evaluated layer-wise from bottom to top.

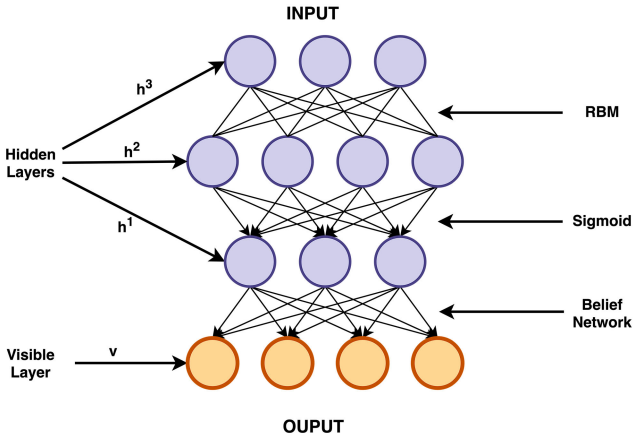

FIGURE 9. Layer-wise architecture of Deep Belief Network.

An RBM is usually generative stochastic because it provides probabilistic distributions as output for a given input. The configuration of RBM based on its energy level for determination of the output distribution function along with its weights and state vectors units y from its visible layer is expressed as

$$E(y, z) = -m^T y - n^T z - y^T W z \tag{3}$$

where z is termed as binary configuration units of the hidden layer, m and n refers to prejudices of both hidden and visible layers. The matrix W provides a connection between the weights of various layers. The probability between pair vectors of hidden and visible layers is given by the following equation

$$P(y, z) = \frac{e^{-E(y,z)}}{L} \tag{4}$$

where L is known as partition function defined for all possible configurations normalizing the probabilistic distribution to unity. As the RBM is unable to model the original input data, so DBN uses its greedy algorithm to improve the generative model by allowing all subnetwork to receive various representations of data. Moreover, with the addition of a new layer into DBN, the overall variational bounds on the deeper layer are further improved as compared to the previous RBM block.

The first main advantage of DBN is the un-supervised nature in pre-training techniques with large and unlabeled databases [71]. The second advantage of DBNs is that they can compute the required output weight of the variables using inference procedure approximation. There lies some limitation as well because the inference procedure of DBNs is only restricted to bottom-up pass. There exists a greedy layer, that learns features of a single layer and never re-adjusts with the remaining layers [72], [73].

E. CONVOLUTIONAL NEURAL NETWORK (CNN)

CNN is another type of Deep learning technique based solely on feed-forward architecture [74] for classification. CNNs are commonly used for pattern recognition and provide better data classification. These networks have small size neurons present on every layer of the designed model architecture that process the input data in the form of receptive fields [75]. Figure 10 provides the layer-wise architecture of a basic CNN network.

Filters are the base of local connections that are convolved with the input and share the same parameters (weight W^i and bias n^i) to generate i feature maps (z^i), each of size $a - b - 1$. The convolutional layers compute the dot product between the weights and provided inputs. So, the parameters for weight W^i and biasing n^i for generation of maps z^i for i features with sizes $a - b - 1$ can be given as

$$z^i = g(W^i * r + n^i) \tag{5}$$

An activation function f or a non-linear methodology needs to be applied to get the output of the convolution layers. It should be noted that inputs are very small regions of the original volumes as depicted in Figure 10. Down sampling is carried out at each subsampling layer to feature maps and decrease the parameters in the network. This, in turn, controls the overfitting and boosts the training process. The pooling process is carried out over $p \times p$ elements (also known as filter size) for adjoining expanse of all the feature maps. In the final stage, the layers need to be fully connected as in other neural networks. These later layers take the previous low-level and mid-level features and generate high-level abstraction form the input speech data. The last layer also known as SVM or Softmax is utilized to further generate the score of classification in probabilistic terms to relate to a certain class.

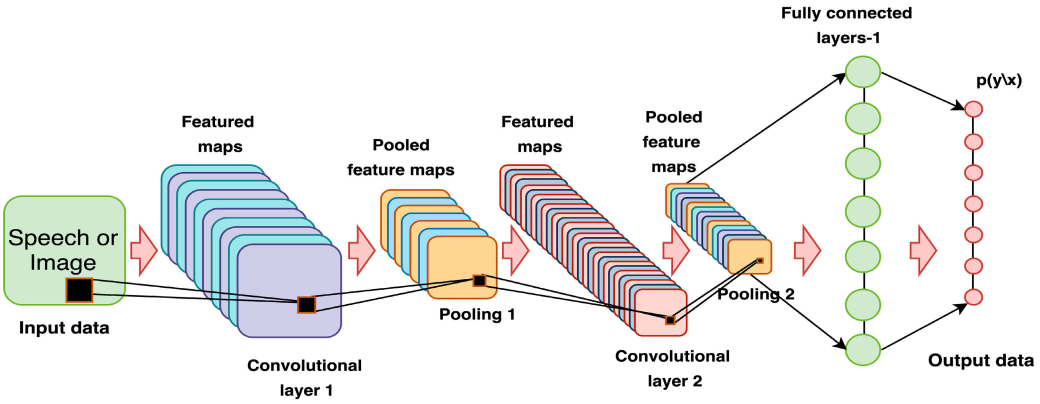

FIGURE 10. Layer-wise convolutional neural network architecture.

F. AUTO ENCODER (AE)

AE is a type of neural network with a detailed built-in model representation [76], [77]. The generalized architecture of AE is depicted in Figure 11.

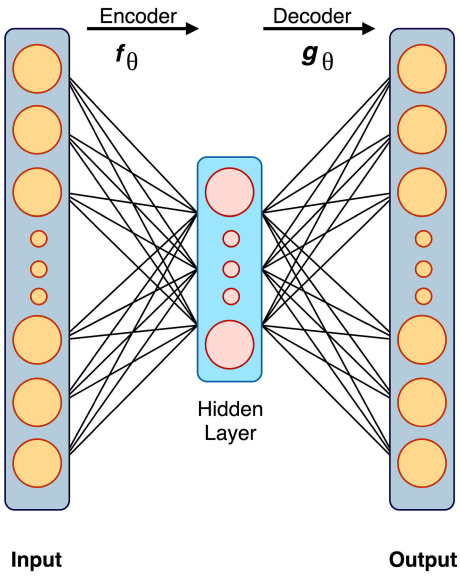

FIGURE 11. Auto Encoder architecture.

The encoder function f_θ is mapped to the provided input vector i in the h hidden layers. In function f_θ, $\theta = \{W_t, c\}$, where W_t is the weight metric, and c is the biasing vector. As far as the decoder function g_θ is concerned, the hidden layer represents h to the input d that is reconstructed via e_o.

There are two variations of AE, that include Stacked Auto Encoder (SAE) and Variational Auto Encoders (VAE). The VAE utilizes the log-likelihood of data and influences the lower bound estimator from a given graphical prototype with uninterrupted underlying variables [78]. The generative parameters θ (generative model parameter) and ϕ (variational parameter) assist the overall process of approximation. Another variation termed as Auto-Encoding Variational Bayes (AEVB) algorithm further optimizes the parameters θ and ϕ for various probabilistic encoders $q_\phi(j/i)$ in any defined neural network. This in turn leads to approximation of the generative model $p_\theta(i, j)$, where j is expressed as latent variable under a simplified distribution given as $N(0, I)$, and I is the identity matrix. The aim to enhance the probability of each i in the given training set with the underlying generative process given by

$$P_\theta(i) = \int p_\theta(j) p_\theta(i/j) dj \tag{6}$$

In the next section, a summary of the papers based on different deep learning techniques for speech-based emotion recognition is presented. Also, discussion is made on the layer-wise working of these deep learning techniques, and future directions are provided.

V. SUMMARY OF THE LITERATURE, DISCUSSION AND FUTURE DIRECTIONS

For SER, many deep learning algorithms have been developed [79]–[84]. However, there exist meaningful prospects and fertile ground for future research opportunities not only in SER but many other domains [85]–[87]. The layer-wise structure of neural networks adaptively learns features from available raw data hierarchically [88], [89]. The remainder of this section summarizes the literature on deep layer architectures, learning and regularization methodologies discussed in the context of SER.

Deep learning techniques utilize some key features during various applications such as SER, natural language processing (NLP) and sequential information processing described in Table 6. In the case of SER, most of these techniques use supervised algorithms during their implementation, however, there is a shift to semi supervised learning [90]. This will enhance the learning of real-world data without the need for manual human labels. Table 7 provides a summary of the Deep learning techniques used by researchers along with their respective descriptive key features, databases used, and results from accuracy outcome, and some commentary on future directions.

TABLE 6. Summary of deep learning techniques with descriptive key features.

Deep Learning Techniques	Descriptive Key Features	References
DBM	Unsupervised learning for RBM and directionless connections	[61]–[64]
RCNN	Efficient for sequential information processing such as NLP and SER	[65]–[67]
RvNN	Utilizes tree-like structure specially for NLP	[68], [69]
DBN	Unsupervised learning and directed connections	[70]–[73]
CNN	Basically designed for Image recognition but can be extended for NLP, computer vision and speech processing	[74], [75]
AE/SAE/VAE	Unsupervised learning based on Probabilistic graphical models	[76]–[78]

The traditional SER systems typically incorporate various classification models such as GMMs and HMMs. The GMMs are utilized for representation of the acoustic features of sound units. The HMMs, on the other hand, are utilized for dealing with temporal variations in speech signals [82]–[84]. The modeling process using such traditional techniques requires a larger dataset to achieve accuracy in emotion recognition, and hence, is time-consuming. In contrast, deep learning methods are comprised of various non-linear rudiments that perform computation on a parallel basis [85]. However, these methods need to be structured with deeper layered architectures.

A deep learning technique based on discriminative pre-training modality using DNN-HMM along with MFCC coefficients has been presented in [95]. The DNN-HMM has been combined with RBM utilizing unsupervised training to recognize different speech emotions. The Hybrid deep learning modality can achieve better results [122]. The same DNN-HMM has been presented and compared with the Gaussian Mixture Model (GMM). It is investigated along with restricted Boltzmann Machine (RBM) for a scenario where unsupervised and discriminative pre-training is concerned. The results obtained in both cases are then compared with those obtained for two layers and multilayers perception of GMM-HMMs and shallow-NN-HMMs. The hybrid DNN-HMMs with pre-training has accuracy using eNTERFACE05 dataset of 12.22% with unsupervised training, 11.67% for GMM- HMMs, 10.56% for MLP-HMMs and 17.22% for shallow- NN-HMMs respectively. This suggests multimodality as a fruitful avenue for research, and also, there is a span for improving the accuracy of emotion recognition, robustness, and efficiency of the recognition system [123].

The main problem that affects the overall performance of the RNN is its sensitivity towards the disappearance of gradients [84], [85]. In [95], an adaptive SER system based on deep learning technique known as DRNN is used for SER. The learning stage of the model includes both frame-level and short-time acoustic features, due to their similar structure. Another multi-tasking deep neural network with shared hidden layers named MT-SHL-DNN is utilized in [111], where the transformation of features is shared. Here, the output layers have an association separately with each data set used. The DNN also helps in measuring the SER based on the nature of the speaker and gender. When the DNNs are used for encoding of segments into length vectors that are fixed in nature, this is done by using pooling of various hidden layer over the specified time. The design of the feature encoding procedure is done in such a manner that it can be used jointly with segmental level classifier for efficient classification.

Convolutional Neural Network (CNN) also uses the layer-wise structure and can categorize the seven universal emotions from the defined speech spectrograms [111]. In [114], a technique for SER that is based on spectrograms and deep CNN is presented. The model consists of three fully connected convolutional layers for extracting emotional features from the spectrogram images of the speech signal. Another adaptation, where a technique that shares priors between related source and target classes (SPRST) is carried out in [115]. The two-layered neural network where speech data is collected from various sources and scenarios usually led to mismatching and hence degrades the overall performance of the system. Initially, a pre-training of the weights is carried out for the first layer and then classification parameters of the second layer are enforced between the two taken classes. These classes with less labeled data in the target domain can borrow information from the source associated domain to compensate for the deficiencies.

The tendency of DNNs to learning the specific features from various auditory emotion recognition systems is analyzed in [117]. These features include voice and music-based recognition. Further, the utilization of cross-channel architecture can improve the general performance in a complex environment. The model provided some good results for human speech signal and music signal; however, the results for generalized auditory emotion recognition are not optimal [110]. The purpose of this cross-channel hierarchy is to extract specific features and combine them into a much more generalized scenario. Also, these models can be coupled with visual-based DNNs to improve automatic SER. RNNs utilization in such a scenario can further boost the performance for the input data with time-dependent constraints.

According to the study in [118], the evaluation of CNN has been assessed using a more autonomous scenario known as Human-Robot Interaction (HRI). The HRI used a humanoid robotic head that resulted in the desired

TABLE 7. Summary of literature on deep learning techniques for SER with discussion and future directions.

S. No.	References	Emotion recognized	Databases used	Deep Learning approach used	Contribution towards Emotion recognition and accuracy	Future Direction
1.	J. Zhao et. al (2019) [91]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Berlin EmoDB and IEMOCAP	CNN and DBN with four LFLBs and one LSTM	Deep 1D and 2D CNN LSTM to achieve 91.6% and 92.9% accuracy	The presented model can be extended to multimodal emotion recognition
2.	S. Tripathi et. al (2018) [92]	Anger, Happiness, Sadness, Neutral	IEMOCAP database	LSTM based RNN with 3 layers	Model is tested on MoCap data with overall accuracy of 71.04%	Future work may lead to inclusion of further layers to get improved results
3.	P. Yenigalla et. al (2018) [93]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	IEMOCAP database	2D CNN with Phoneme data input data	Achieve accuracy in SER for about 4% above average	It can be tested in other applications such conversational Chatbot and other databases
4.	E. Lakomkin et. al (2018) [94]	Anger, Happiness, Neutral and Sadness	IEMOCAP database	RNN and CNN	Combine RNN-CNN for iCLub robot and in-domain data with 83.2% accuracy	Future work may lead to use of generative models for real-time data input
5.	D. Tang et. al (2018) [95]	Anger, Happiness, Neutral and Sadness	EmotAsS (EMOTional Sensitivity Assistance System for people with disabilities) dataset	Combine CNN and RNN with ResNEt	The CNN+RNN model achieves 45.12% on the improvement dataset, and 42.27% on the test dataset.	Future work may include the learning of features and augmentation
6.	C. W. Lee et. al (2018) [96]	Sadness, Happiness, Anger, Disgust, Surprise, and Fear	CMU-MOSEI dataset	LSTM based CNN	The proposed model gives overall 83.11% accuracy in SER	Future work may lead to testing on multimodalilty
7.	S. Sahu et. al (2018) [97]	Open Smile features recognized	IEMOCAP database	Adversarial auto-encoders (AAE)	Accuracy with UAR of approximately 57.88%	Future work may lead to recognizing other emotions
8.	S. Latif et. al (2018) [98]	Motherese, Joyful, Neutral, Rest, Angry, Touchy, Emphatic and Sadness	FAU-AIBO, IEMOCAP, EMoDB, SAVEE and EMOVO databases	RBM based DBN	DBN is used for transfer learing for cross-corpus and cross-language	Future work may include application of other neural networks
9.	M. Chen et. al (2018) [99]	anger, Sadness, Happy, Neutral, Fear, Disgust and Bored	IEMOCAP and Emo-DB databases	3-D CNN with LSTM to learn discriminative features	The model achieves an overall of 86.99%	Future work includes the testing on different databases
10.	M. Sarma et. al (2018) [100]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	IEMOCAP database	TDNN-LSTM for SER	TDNN-LSTM with time-restriction for self-attention, and achieve a weighted accuracy of 70.6%, versus 61.8% previously	Future work includes investigation in multi-dimensional space
11.	S. E. Eskimez et. al (2018) [101]	Anger, Frustration, Neutral, and Sadness	USC-IEMOCAP audio-visual dataset	CNN with VAE, AAE and AVB	Better achievement on F-1 score level as 47%	Future work may include RNN
12.	J. Zhao et. al (2018) [102]	Joy, Happiness, Neutral, Sadness, Disgust, Fear, and Anger	Berlin EmoDB and IEMOCAP	Deep Convolutional Neural Network (DCNN)	Merged Deep 1D and 2D CNN for high-level learning of features from input audio and log-mel spectrograms with 92.71% accuracy	Future work may include addition of LSTM for better SER
13.	W. Zhang et. al (2017) [103]	Fear, Anger, Neutral, Joy, Surprise and Sadness	CAS emotional speech database	Features fusion with SVM and Deep Belief Network for SER is carried out	DBM provides accuracy of 94.6% as compared to SVM that is 84.54%	Future direction may leads to more train DBN with combination of lexical features and audio features
14.	Z. Lianzhang et. al (2017) [104]	Anger, Fear, Happy, Neutral, Sad, Surprise, and Average	Chinese Academy of Sciences emotional speech database	Combined SVM and DBN for SER	The combined model achieves 94.6% accuracy	Future work includes testing on other databases

TABLE 7. *(Continued.)* **Summary of literature on deep learning techniques for SER with discussion and future directions.**

S. No.	References	Emotion recognized	Databases used	Deep Learning approach used	Contribution towards Emotion recognition	Future Direction
15.	P. Tzirakis et. al (2017) [105]	Anger, Happiness, Sadness, Neutral	Spontanous emotional RECOLA and AVEC 2016 Database	Convolutional Neural Network (CNN) and ResNet of 50 layers for both audio-visual modality along with LSTM	End-to-end methodology is used to recognize various emotions with 78.7% accuracy	Future work involves the application of proposed model on other databases
16.	H.M. Fayek et. al (2017) [106]	Anger, Happy, Neutral, Sad and Silence	IEMOCAP Database	Feed forward in combination with Recurrent Neural Network (RNN) and CNN to recognize emotion from speech	Proposed SER technique relies on minimal speech processing and frame based end-to-end deep learning 64.78% accuracy	Future work may be to apply the same model to other databases
17.	Q. Mao et. al (2017) [107]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	INTERSPEECH 2009 Emotion Challenge, ABC and Emo-DB	Two layers Emotion-discriminative and Domain-invariant Feature Learning Method (EDFLM)	the proposed method effectively transferred knowledge and enhanced the classification performance with 65.62% and 61.63% accuracy	Future direction may leads to application on other databases
18.	S. Zhang et. al (2017) [108]	Anger, Fear, Happy, Neutral, Sad, Surprise, and Average	EMoDB, eNTERFACE05, RML and BAUM-1s	DCNN model with DTPM strategy in combination with temporal pyramid matching and Lp-norm pooling for feature representation	The model provides accuracy of 87.31% for EMo-DB, 69.70% for RML, 76.56% for eNTERFACE05 and 44.61% for BAUM-1s	Future directions may lead to CNN with LSTM modality for SER
19.	J. Deng et. al (2017) [109]	Neutral, Anger, Fear, Disgust, Sadness, Boredom and Happiness	ABC, EMoDB and Geneva Whispered Emotion Corpus (GeWEC)	Unsupervised Universum Autoencoders adaptive model	The model learn discriminative data and incorporate the unlabeled learning with 62% , 63.3% and 62.8% accuracy	The future works may leads to paralinguistic computations
20.	S. Mirsamadi et. al (2017) [110]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	IEMOCAP databases	Deep Recurrent Neural Network (RNN)	Deep RNNs for frame-level characterization achieves +5.7% and +3.1% accuracy in WA and UA, respectively as compared to SVM	The presented model can be tested on other deep learning techniques and databases
21.	Y. Zhang et. al (2017) [111]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	ComParE and eGeMAPS feature set	multi-tasking DNNs with few shared hidden layers (MT-SHL-DNN)	multi-tasking DNN has been used for the very first time with 93.6% accuracy	Can use multiple Softmax in combination with linear layers for classification and regression tasks
22.	Y. Zhao et. al (2017) [112]	Neutral, Anger, Fear, Disgust, Sadness, Joy and Happiness	IEMOCAP for emotion recognition and TIMIT for phoneme recognition	Recurrent Convolutional Neural Network (RCNN)	This hybrid model has been applied for SER for the very first time with 83.4% accuracy	Possibility of making this model more generic and efficient cross modal deep learnings
23.	Z. Q. Wang and I. Tashev (2017) [113]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Mandarin dataset	DNN based kernel extreme learning machine (ELM)	DNN-ELM approach provides 3.8% weighted accuracy and 2.94% unweighted accuracy in SER	Future works include testing on other corpus
24.	J. Han et. al (2017) [114]	Anger, Happiness, Sadness, Neutral	Spontanous emotional RECOLA Database	Memory enhance Recurrent Neural Network (RNN) with baseline of LSTM	Novelty here is Reconstruction-Error-based (RE-based) framework for emotion recognition in a continuous data	Future work involves the utilization of BLSTM-RNN with regression like Support Vector Regression

TABLE 7. ***(Continued.)*** **Summary of literature on deep learning techniques for SER with discussion and future directions.**

S. No.	References	Emotion recognized	Databases used	Deep Learning approach used	Contribution towards Emotion recognition	Future Direction
25.	Abdul Malik et. al (2017) [115]	Anger, Boredom, Disgust, Joy, Sadness and Neutral	Berlin Emotions Database	Deep Convolutional Neural Network (CNN)	Emotion recognition has been done using spectrogram of the speech signal	The model need to be trained enough to recognize the emotion for fear. Also it can be test with other emotional databases
26.	Qirong Mao et. al (2016) [116]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	INTERSPEECH 2009 Emotion Challenge, FAU-AEC, Emo-DB	Two layers DNN for Sharing Priors between Related Source and Target classes (SPRST)	the proposed method effectively transferred knowledge and enhanced the classification performance for emotion recognition	Further direction may lead from single to many layers architecture.
27.	Pablo et. al (2016) [117]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	GTZAN database, SAVEE database and EMotiW corpus databases	Cross-channel architecture based Deep Neural Network (DNN)	Model perform better for speech, music and complex audio signals from recorded clips	Model need training to achieve optimum results for generalized auditory scenarios.
28.	Lim Wootaek et. al (2016) [118]	Neutral, Anger, Fear, Disgust, Sadness, Boredom and Happy	Berlin Emotional Database (Emo-DB)	Concatenation of CNNs and RNNs with no traditional hand-crafted features	The deep hierarchical feature extraction architecture of CNNs is combined with LSTM network layers for better emotion recognition	The future works may leads to a more concatenated CNNs and using it for multimodal (audio/video) based emotion recognition.
29.	W. Q. Zheng et. al (2015) [119]	Neutral, Anger, Fear, Sadness, Joy and Happiness	Interactive emotional dyadic motion capture database (IEMOCAP)	Principle component analysis (PCA) based Deep Convolutional Neural Network (CNN)	Better results obtained using CNN with 40% more accuracy. Works efficiently with SVM	Input audio data is fixed, need extension to variable length input data.
30.	Pablo et. al (2015) [120]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Cohn-Kanade dataset for emotions, CAM3D corpus for spontaneous response	Deep Convolutional Neural Network (CNN)	Obtained emotional expressions are spontaneous and thus can easily be classified into positive or negative	Improvement can be done to achieve accurate decisions for various emotions.
31.	Fayek et. al (2015) [121]	Anger, Boredom, Disgust, Joy, Sadness and Neutral	eNTERFACE and SAVEE database	End-to-End architecture based on DNN	60.53% accuracy for eNTERFACE dataset (6 emotions) and 59.7% accuracy for SAVEE dataset (all 7 emotions)	Model can be modified for the real-time input speech data.
32.	Qirong Mao et. al (2014) [122]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Emo-DB, MES, SAVEE, DES databases	Sparse Auto Encoder (SAE) and Salient discriminative feature analysis (SDFA) based Convolutional Neural Network (CNN)	Trained the CNN with learn affect-salient features and achieved robust emotion recognition with variational speaker, language and environment	Model can be modified for the naturalistic speech input data to obtain real-time emotion recognition
33.	Z. Huang et. al (2014) [123]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	Emo-DB, MES, SAVEE, DES databases	Unsupervised and Semi-Supervised CNN	Feature learning and emotion-salient features learning using semi-CNN	Model can be modified for the naturalistic speech input data to obtain real-time emotion recognition.
34.	Jianwei et. al (2014) [124]	Anger, boredom, disgust, fear, joy, sadness, neutral	TIMIT Corpus database	DNN based acoustic features such as MFCCs, PLPs and FBANKs	3 Hidden layers based DNN with acoustics of MFCCs, PLPs, FBANKs and achieved 8.2% emotion recognition	Future work may involves the increase in hidden layers of the considered DNN

TABLE 7. *(Continued.)* **Summary of literature on deep learning techniques for SER with discussion and future directions.**

S. No.	References	Emotion recognized	Databases used	Deep Learning approach used	Contribution towards Emotion recognition	Future Direction
35.	L. Li et. al (2013) [125]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	eNTERFACE'05 and Berlin Database	HMM based Hybrid DNN (DNN-HMM)	Using DNN-HMM provides 12.22% better results for eNTERFACE'05, 11.67% for GMM-HMM, 10.56% for MLP-HMM, and 17.22% for shallow-NN-HMM, respectively	The future direction shows the authors interest to make their designed system an audio-visual based SER.
36.	André et. al (2011) [126]	Happiness, Sadness, Neutral, Surprise, Disgust, Fear, and Anger	eNTERFACE database and EMOD database	Generalized Discriminant Analysis (GerDA) based Deep Neural Network (DNN)	Comparative analysis of GerDA based DNN with SVM and obtained better speech emotion recognition	this comparison can be modeled and extended towards multi-modality.
37.	L. Fu et. al (2008) [127]	Anger, boredom, disgust, fear, joy, sadness, neutral	Berlin Database and Mandarin Database	Hybrid features based Artificial Neural Network (ANN)	Achieved better results for relative features as compared to absolute features for emotion recognition	Exploration and extraction of more rationalized method for feature extraction.

emotional feedback. It works with several processes simultaneously: feature extraction based on shape, facial characteristics and information regarding the subject motion. The reception of shunting inhibitory fields internally increased the depth of the network for emotion extraction. Later on, cross-channel learning is used to correlate the static and dynamic streams in the same scenario. As the model works efficiently for a person performing spontaneous real-time expressions, it can be further extended to a multimodal system, where visual stimuli can also be used as an input with audio.

A hybrid deep learning modality may inherit the underlying properties of RNN with CNN, with convolutional levels implanted with RNN [94], [105], [118]. This enables the model to obtain both frequency and temporal dependency in a given speech signal. Sometimes, a memory enhanced reconstruction-error-based RNN for continuous speech emotion recognition can also be used [97]. This RNN model uses two components, first an auto-encoder for the reconstruction of features, and second for the prediction of emotions. It can also be used to obtain further insights into the behavior of BLSTM-based RNN using regression models such as SVR [95].

SER algorithms based on CNNs and RNNs have been investigated in [118]. The deep hierarchical CNNs architecture for feature extraction has been combined with LSTM network layers. It was found that CNNs have a time-based distributed network that provides results with greater accuracy. Similarly, in [103], a system based on a deep convolutional network (DCNN) that used audio data as input is presented, named PCA-DCNNs-SER. It contained 2 convolutional layers and 2 pooling layers. Before computing the log-spectrogram, the background interferences have been eliminated by using Principal Component Analysis (PCA) scheme. The noise-free spectrogram is divided into non-overlapping components. Using hand-crafted acoustic features, this model also performed well for the SVM based classification. As emotional expressions are usually spontaneous, their classification into the positive and negative domain is quite easy.

Preprocessing is an aditionaltask usually required to be carried out before recognition of emotion from the speech signal [120]. To eliminate these drawbacks, a real-time SER system is needed that can work on end-to-end deep learning architecture. An example based spectrogram processing using DNN is given in [122]. It requires a deep hierarchical architecture, regularization, and augmentation of data.

Emotion recognition from speech using salient features has also been researched, where a CNN based on affective learning is used [121]. Initially, the CNN is trained with unlabeled samples for learning localized invariant features (LIF) using a Sparse auto-encoder (SAE). A well-known variation of auto-encoder is the Variational Auto-encoder (VAE) [120]. Afterward, the LIF is utilized as input for the extraction of features using SDFA. The reason here is to learn salient features that are discriminative and orthogonal for speech emotion recognition [123]. The results obtained with these experiments are more stable, accurate and robust in recognition in complex scenarios where there is variation in language and speaker, and other environmental distortion. Emotions such as happiness, joy, sadness, neutral, surprise, boredom, disgust, fear, and anger can easily be recognized. But it becomes hard to do so when real-time emotion recognition is desired. A variation of this salient features-based emotion learning has been presented in [123], where

a semi-Convolutional neural network (semi-CNN) is presented for SER. The same two-stage training has been carried out. However, the experimental results obtained are robust, stable and accurate for emotion recognition in scenarios such as a distorted speech environment. The experiments are carried out on synthetic data.

A comparison of DNN with GMM for SER has been presented in [124], where both the techniques are tested for the six universal emotions. The comparative analysis of GMM and DNN classifiers showed that the system's performance could be improved after the introduction of deep learning. For system modeling, various sizes of deep hidden layers (512, 1024, 2048 and so on) and acoustic features such as MFCCs, PLPs, FBANKs are used. Combining these three acoustic features showed better performance and provide recognition rate of 92.3% as compared to single acoustic features that were 92.1% when using MFCCs. This study concluded that emotion recognition systems that are based on deep learning architecture provide better performance compared to the systems using GMMs as classifiers. A DNN based on Generalized Discriminant Analysis (GerDA) for SER is proposed in [26], [125], [127]. The model learns features that are discriminative for low dimensions and optimized for faster classification of large acoustic feature set. A comparative analysis is carried out for GerDA and its competitor linear classifiers such as SVMs to analyze its performance.

Deep learning is mostly used for natural language processing but its architecture is able to handle different modalities such as SER and speech recognition. According to [128], the RvNN can be used for the classification of natural language sentences as well as natural image processing. It initially computes the overall score of the possible pair for merging them in pairs and to build a syntactic tree. The pair of highest score is further combined with a vector known as compositional vector [129]–[131]. Once the pair gets merged, the RvNN then generates multiple units, the region representing vectors, and the classification labels. It should be noted that the RvNN tree structure is the compositional representation of vectors for the entire considered region.

As a final remark, deep learning is rapidly becoming a method of choice over traditional techniques for SER. Also, most of the research is evolving towards multimodal and unsupervised SER, speech recognition and NLP [132], [133]. Multimodal emotion recognition can use input data such as audio-visual at the same time, in an efficient way.

VI. CONCLUSION

This paper has provided a detailed review of the deep learning techniques for SER. Deep learning techniques such as DBM, RNN, DBN, CNN, and AE have been the subject of much research in recent years. These deep learning methods and their layer-wise architectures are briefly elaborated based on the classification of various natural emotion such as happiness, joy, sadness, neutral, surprise, boredom, disgust, fear, and anger. These methods offer easy model training as well as the efficiency of shared weights. Limitations of deep learning techniques include their large layer-wise internal architecture, less efficiency for temporally-varying input data and over-learning during memorization of layer-wise information. This research work forms a base to evaluate the performance and limitations of current deep learning techniques. Further, it highlights some promising directions for better SER systems.

ACKNOWLEDGMENT

This article was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah. The authors therefore, acknowledge with thanks DSR for technical and financial support.

REFERENCES

[1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018.
[2] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data," *Inf. Fusion*, vol. 49, pp. 69–78, Sep. 2019.
[3] M. Chen, P. Zhou, and G. Fortino, "Emotion communication system," *IEEE Access*, vol. 5, pp. 326–337, 2016.
[4] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?" in *Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl.*, 2015, pp. 117–122.
[5] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet, and A. Larsson, "Speech emotion recognition in emotional feedbackfor human-robot interaction," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 2, pp. 20–27, 2015.
[6] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2013, pp. 216–221.
[7] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.
[8] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in *Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC)*, Oct. 2014, pp. 1–4.
[9] K. R. Scherer, "What are emotions? And how can they be measured?" *Social Sci. Inf.*, vol. 44, no. 4, pp. 695–729, 2005.
[10] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.* Springer, 2004, pp. 318–328.
[11] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.
[12] O. Kwon, K. Chan, J. Hao, T. Lee, "Emotion recognition by speech signal," in *Proc. EUROSPEECH*, Geneva, Switzerland, 2003, pp. 125–128.
[13] R. W. Picard, "Affective computing," Perceptual Comput. Sect., Media Lab., MIT, Cambridge, MA, USA, Tech. Rep., 1995.
[14] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review," *Int. J. speech Technol.*, vol. 15, no. 2, pp. 99–117, 2012.
[15] 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, 2011.
[16] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Trans. neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1421–1432, Aug. 2014.
[17] L. Deng and D. Yu, "Deep learning: Methods and applications," *Found. Trends Signal Process.*, vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.
[18] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, Jan. 2015.
[19] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2005, pp. 474–477.

[20] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011," *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155–177, 2015.
[21] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in *Emotion-Oriented Systems*. Springer, 2011, pp. 71–99.
[22] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 5, pp. 1057–1070, Jul. 2011.
[23] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion recognition in speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5005–5009.
[24] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Reconstruction-error-based learning for continuous emotion recognition in speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2367–2371.
[25] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[26] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: A review of the literature and recommendations for practical realisation," in *Affect and Emotion in Human-Computer Interaction*. Springer, 2008, pp. 75–91.
[27] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning," *IEEE Access*, vol. 5, pp. 5235–5246, 2017.
[28] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," *J. Adv. Comput. Netw.*, vol. 2, no. 1, pp. 28–30, 2014.
[29] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. 4th Int. Conf. Spoken Lang. Process.*, Oct. 1996, pp. 1970–1973.
[30] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in *Proc. IEEE Int. Conf. Inf. Eng. Comput. Sci. (ICIECS)*, Dec. 2009, pp. 1–4.
[31] S. Haq, P. J. Jackson, and J. Edge, "Audio-visual feature selection and reduction for emotion classification," in *Proc. Int. Conf. Auditory-Visual Speech Process. (AVSP)*, Tangalooma, QLD, Australia, 2008.
[32] M. Alpert, E. R. Pouget, and R. R. Silva, "Reflections of depression in acoustic measures of the patient's speech," *J. Affect. Disorders*, vol. 66, no. 1, pp. 59–69, 2001.
[33] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Commun.*, vol. 48, no. 9, pp. 1162–1181, 2006.
[34] S. Mozziconacci, "Prosody and emotions," in *Proc. Int. Conf. Speech Prosody*, 2002, pp. 1–9.
[35] J. Hirschberg, S. Benus, J. M. Brenier, F. Enos, and S. Friedman, "Distinguishing deceptive from non-deceptive speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol. (INTERSPEECH)*, 2005, pp. 1833–1836.
[36] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. Hum.-Comput. Stud.*, vol. 59, nos. 1–2, pp. 157–183, 2003.
[37] D. Neiberg, K. Elenius, and K. Laskowski, "Emotion recognition in spontaneous speech using GMM," in *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, 2006, pp. 809–812.
[38] A. D. Dileep and C. C. Sekhar, "HMM based intermediate matching kernel for classification of sequential patterns of speech using support vector machines," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2570–2582, Dec. 2013.
[39] G. Vyas, M. K. Dutta, K. Riha, and J. Prinosil, "An automatic emotion recognizer using MFCCs and hidden Markov models," in *Proc. IEEE 7th Int. Congr. Ultra Mod. Telecommun. Control Syst. Workshops (ICUMT)*, Oct. 2015, pp. 320–324.
[40] S. Wang, X. Ling, F. Zhang, and J. Tong, "Speech emotion recognition based on principal component analysis and back propagation neural network," in *Proc. Int. Conf. Measuring Technol. Mechatron. Automat. (ICMTMA)*, vol. 3, Mar. 2010, pp. 437–440.
[41] Y. Pan, P. Shen, and L. Shen, "Speech emotion recognition using support vector machine," *Int. J. Smart Home*, vol. 6, no. 2, pp. 101–108, 2012.
[42] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *Int. J. Comput. Appl.*, vol. 1, no. 20, pp. 6–9, 2010.
[43] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 1, May 2004, p. I-577.
[44] 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, 2018.
[45] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *Proc. 1st Richmedia Conf.*, 2003, pp. 109–119.
[46] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, 2005, pp. 1–4.
[47] J. A. Coan and J. J. Allen, *Handbook of Emotion Elicitation and Assessment*. London, U.K.: Oxford Univ. Press, 2007.
[48] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE' 05 audio-visual emotion database," in *Proc. IEEE 22nd Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2006, p. 8.
[49] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. 6th Int. Workshop Audio/Vis. Emotion Challenge (AVEC)*, New York, NY, USA, no. 8, 2016, pp. 3–10.
[50] P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*. Guildford, U.K.: Univ. Surrey, 2014.
[51] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *Proc. IEEE 10th Int. Conf. Workshops Autom. Face Gesture Recognit. (FG)*, Apr. 2013, pp. 1–8.
[52] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond emotion archetypes: Databases for emotion modelling using neural networks," *Neural Netw.*, vol. 18, no. 4, pp. 371–388, 2005.
[53] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, p. 436, 2015.
[54] W. Wang, Ed., *Machine Audition: Principles, Algorithms and Systems*. Hershey, PA, USA: IGI Global, 2010.
[55] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Commun.*, vol. 53, nos. 9–10, pp. 1062–1087, Nov./Dec. 2011.
[56] M. Sidorov, S. Ultes, and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4803–4807.
[57] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Proc. INTERSPEECH*, Makuhari, Japan, 2010, pp. 2362–2365.
[58] Y. Niu, D. Zou, Y. Niu, Z. He, and H. Tan, "A breakthrough in speech emotion recognition using deep retinal convolution neural networks," 2017, *arXiv:1707.09917*. [Online]. Available: https://arxiv.org/abs/1707.09917
[59] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[60] M. W. Bhatti, Y. Wang, and L. Guan, "A neural network approach for human emotion recognition in speech," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, vol. 2, May 2004, p. II-81.
[61] K. Poon-Feng, D.-Y. Huang, M. Dong, and H. Li, "Acoustic emotion recognition based on fusion of multiple feature-dependent deep Boltzmann machines," in *Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Sep. 2014, pp. 584–588.
[62] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn. (ICML)*, 2010, pp. 807–814.
[63] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.
[64] A.-R. Mohamed, G. Dahl, and G. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learn. Speech Recognit. Rel. Appl.*, Vancouver, BC, Canada, 2009, vol. 1, no. 9, p. 39.
[65] 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. 1537–1540.
[66] V. Chernykh and P. Prihodko, "Emotion recognition from speech with recurrent neural networks," 2017, *arXiv:1701.08071*. [Online]. Available: https://arxiv.org/abs/1701.08071
[67] F. Weninger, F. Ringeval, E. Marchi, and B. W. Schuller, "Discriminatively trained recurrent neural networks for continuous dimensional emotion recognition from audio," in *Proc. IJCAI*, 2016, pp. 2196–2202.
[68] Y. Kamp and M. Hasler, *Recursive Neural Networks for Associative Memory*. Hoboken, NJ, USA: Wiley, 1990.
[69] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng, "Parsing natural scenes and natural language with recursive neural networks," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 129–136.

[70] 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.
[71] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.
[72] C. Huang, W. Gong, W. Fu, and D. Feng, "A research of speech emotion recognition based on deep belief network and SVM," *Math. Problems Eng.*, vol. 2014, Aug. 2014, Art. no. 749604.
[73] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 153–160.
[74] W. Q. Zheng, J. S. Yu, and Y. X. Zou, "An experimental study of speech emotion recognition based on deep convolutional neural networks," in *Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII)*, Sep. 2015, pp. 827–831.
[75] Y. Kim, "Convolutional neural networks for sentence classification," 2014, *arXiv:1408.5882*. [Online]. Available: https://arxiv.org/abs/1408.5882
[76] W. Fei, X. Ye, Z. Sun, Y. Huang, X. Zhang, and S. Shang, "Research on speech emotion recognition based on deep auto-encoder," in *Proc. IEEE Int. Conf. Cyber Technol. Automat., Control, Intell. Syst. (CYBER)*, Jun. 2016, pp. 308–312.
[77] J. Deng, R. Xia, Z. Zhang, Y. Liu, and B. Schuller, "Introducing shared-hidden-layer autoencoders for transfer learning and their application in acoustic emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 4818–4822.
[78] J. Deng, Z. Zhang, E. Marchi, and B. Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. IEEE Humaine Assoc. Conf. Affect. Comput. Intell. Interact.*, Sep. 2013, pp. 511–516.
[79] L. Cen, W. Ser, and Z. L. Yu, "Speech emotion recognition using canonical correlation analysis and probabilistic neural network," in *Proc. 7th Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2008, pp. 859–862.
[80] X. Zhou, J. Guo, and R. Bie, "Deep learning based affective model for speech emotion recognition," in *Proc. Ubiquitous Intell. Comput., Adv. Trusted Comput., Scalable Comput. Commun., Cloud Big Data Comput., Internet People, Smart World Congr. (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)*, Jul. 2016, pp. 841–846.
[81] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Proc. 5th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 223–227.
[82] K.-C. Huang and Y.-H. Kuo, "A novel objective function to optimize neural networks for emotion recognition from speech patterns," in *Proc. IEEE 2nd World Congr. Nature Biolog. Inspired Comput. (NaBIC)*, Dec. 2010, pp. 413–417.
[83] E. M. Albornoz, M. Sánchez-Gutiérrez, F. Martinez-Licona, H. L. Rufiner, and J. Goddard, "Spoken emotion recognition using deep learning," in *Proc. Iberoamer. Congr. Pattern Recognit.* Cham, Switzerland: Springer, Nov. 2014, pp. 104–111.
[84] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks—Studies on speech recognition tasks," 2013, *arXiv:1301.3605*. [Online]. Available: https://arxiv.org/abs/1301.3605
[85] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[86] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 30–42, Jan. 2012.
[87] A. Satt, S. Rozenberg, and R. Hoory, "Efficient emotion recognition from speech using deep learning on spectrograms," in *Proc. INTERSPEECH*, 2017, pp. 1089–1093.
[88] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Dec. 2011, pp. 24–29.
[89] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," 2014, *arXiv:1404.2188*. [Online]. Available: https://arxiv.org/abs/1404.2188
[90] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 5200–5204.
[91] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomed. Signal Process. Control*, vol. 47, pp. 312–323, Jan. 2019.
[92] S. Tripathi and H. Beigi, "Multi-modal emotion recognition on IEMOCAP dataset using deep learning," 2018, *arXiv:1804.05788*. [Online]. Available: https://arxiv.org/abs/1804.05788
[93] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognitionusing spectrogram & phoneme embedding," in *Proc. Interspeech*, 2018, pp. 3688–3692.
[94] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, "On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 854–860.
[95] D. Tang, J. Zeng, and M. Li, "An end-to-end deep learning framework with speech emotion recognitionof atypical individuals," in *Proc. Interspeech*, Sep. 2018, pp. 162–166.
[96] C. W. Lee, K. Y. Song, J. Jeong, and W. Y. Choi, "Convolutional attention networks for multimodal emotion recognition from speech and text data," 2018, *arXiv:1805.06606*. [Online]. Available: https://arxiv.org/abs/1805.06606
[97] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, and C. Espy-Wilson, "Adversarial auto-encoders for speech based emotion recognition," 2018, *arXiv:1806.02146*. [Online]. Available: https://arxiv.org/abs/1806.02146
[98] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, "Transfer learning for improving speech emotion classification accuracy," 2018, *arXiv:1801.06353*. [Online]. Available: https://arxiv.org/abs/1801.06353
[99] M. Chen, X. He, J. Yang, and H. Zhang, "3-D convolutional recurrent neural networks with attention model for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 25, no. 10, pp. 1440–1444, Oct. 2018.
[100] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, and N. Dehak, "Emotion identification from raw speech signals using DNNs," in *Proc. Interspeech*, 2018, pp. 3097–3101.
[101] S. E. Eskimez, Z. Duan, and W. Heinzelman, "Unsupervised learning approach to feature analysis for automatic speech emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5099–5103.
[102] J. Zhao, X. Mao, and L. Chen, "Learning deep features to recognise speech emotion using merged deep CNN," *IET Signal Process.*, vol. 12, no. 6, pp. 713–721, 2018.
[103] 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, 2017.
[104] 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, 2017.
[105] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 8, pp. 1301–1309, Dec. 2017.
[106] H. M. Fayek, M. Lech, and L. Cavedon, "Evaluating deep learning architectures for speech emotion recognition," *Neural Netw.*, vol. 92, pp. 60–68, Aug. 2017.
[107] Q. Mao, G. Xu, W. Xue, J. Gou, and Y. Zhan, "Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition," *Speech Commun.*, vol. 93, pp. 1–10, Oct. 2017.
[108] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Trans. Multimedia*, vol. 20, no. 6, pp. 1576–1590, Oct. 2017.
[109] J. Deng, X. Xu, Z. Zhang, and S. Frühholz, and B. Schuller, "Universum autoencoder-based domain adaptation for speech emotion recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 4, pp. 500–504, Apr. 2017.
[110] 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. (ICASSP)*, Mar. 2017, pp. 2227–2231.
[111] Y. Zhang, Y. Liu, F. Weninger, and B. Schuller, "Multi-task deep neural network with shared hidden layers: Breaking down the wall between emotion representations," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 4990–4994.
[112] Y. Zhao, X. Jin, and X. Hu, "Recurrent convolutional neural network for speech processing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5300–5304.

[113] Z.-Q. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 5150–5154.
[114] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *Proc. IEEE Int. Conf. Platform Technol. Service (PlatCon)*, Feb. 2017, pp. 1–5.
[115] Q. Mao, W. Xue, Q. Rao, F. Zhang, and Y. Zhan, "Domain adaptation for speech emotion recognition by sharing priors between related source and target classes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 2608–2612.
[116] P. Barros, C. Weber, and S. Wermter, "Learning auditory neural representations for emotion recognition," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 921–928.
[117] 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. (APSIPA)*, Dec. 2016, pp. 1–4.
[118] P. Barros, C. Weber, and S. Wermter, "Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction," in *Proc. IEEE-RAS 15th Int. Conf. Humanoid Robots (Humanoids)*, Nov. 2015, pp. 582–587.
[119] H. M. Fayek, M. Lech, and L. Cavedon, "Towards real-time speech emotion recognition using deep neural networks," in *Proc. IEEE 9th Int. Conf. Signal Process. Commun. Syst. (ICSPCS)*, Dec. 2015, pp. 1–5.
[120] Q. Mao, M. Dong, Z. Huang, and Y. Zhan, "Learning salient features for speech emotion recognition using convolutional neural networks," *IEEE Trans. Multimedia*, vol. 16, no. 8, pp. 2203–2213, Dec. 2014.
[121] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using CNN," in *Proc. ACM 22nd Int. Conf. Multimedia*, 2014, pp. 801–804.
[122] J. Niu, Y. Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," in *Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Sep. 2014, pp. 128–132.
[123] 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. (ACII)*, Sep. 2013, pp. 312–317.
[124] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5688–5691.
[125] L. Fu, X. Mao, and L. Chen, "Relative speech emotion recognition based artificial neural network," in *Proc. IEEE Pacific–Asia Workshop Comput. Intell. Ind. Appl. (PACIIA)*, vol. 2, Dec. 2008, pp. 140–144.
[126] D. Le, Z. Aldeneh, and E. M. Provost, "Discretized continuous speech emotion recognition with multi-task deep recurrent neural network," in *Proc. INTERSPEECH*, 2017, pp. 1108–1112.
[127] N. Morgan, "Deep and wide: Multiple layers in automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 7–13, Jan. 2012.
[128] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, "'Of all things the measure is man' automatic classification of emotions and inter-labeler consistency [speech-based emotion recognition]," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 1, Mar. 2005, pp. I/317–I/320.
[129] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proc. Speech Prosody*, 2006.
[130] I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang, "Semisupervised learning of classifiers: Theory, algorithms, and their application to human-computer interaction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 12, pp. 1553–1566, Dec. 2004.
[131] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. Courville, P. Vincent, R. Memisevic, C. Pal, and Y. Bengio, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *J. Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.
[132] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2015, pp. 467–474.
[133] H. Ranganathan, S. Chakraborty, and S. Panchanathan, "Multimodal emotion recognition using deep learning architectures," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–9.

RUHUL AMIN KHALIL received the bachelor's and master's degrees in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2013 and 2015, respectively. He is currently pursuing the Ph.D. degree in electrical engineering with the University of Engineering and Technology, Peshawar, Pakistan.

He is also a Lecturer with the Department of Electrical Engineering, University of Engineering and Technology, Peshawar, Pakistan. His research interests include audio signal processing and its applications, pattern recognition, machine learning, and wireless communication.

EDWARD JONES received the B.E. degree in electronics engineering (Hons.), and the Ph.D. degree in electronics engineering from the National University of Ireland (NUI), Galway. His Ph.D. research topic was on the development of computational auditory models for speech processing. He is currently a Professor with the Department of Electrical and Electronics Engineering, NUI, Galway. From 2010 to 2016, he was a Vice-Dean of the College of Engineering and Informatics, with responsibility for performance, planning, and strategy. He also has a number of years of industrial experience in senior positions, in both indigenous start-up and multinational companies.

His current research interests include DSP algorithm development and embedded implementation for applications in connected and autonomous vehicles, biomedical engineering, speech and audio processing, and environmental/agriculture applications.

MOHAMMAD INAYATULLAH BABAR received the B.Sc. degree in electrical engineering from the University of Engineering and Technology (UET), Peshawar, Pakistan, in 1997, and the master's and Ph.D. degrees from the School of Engineering and Applied Sciences, George Washington University, Washington, DC, USA, in 2001 and 2005, respectively.

He also taught a number of telecommunications engineering courses at graduate level with the School of Engineering, Stratford University, Virginia, USA, as an Adjunct Faculty. He is currently a Professor with the Department of Electrical Engineering, supervising postgraduate scholars in the field of wireless communications network. He has authored and coauthored more than 50 publications in reputable engineering conferences and journals. He is a member of ACM, USA.

TARIQULLAH JAN received the bachelor's degree in electrical engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2002, and the Ph.D. degree in electronic engineering from the University of Surrey, U.K., in 2012. He is currently an Associate Professor with the Department of Electrical Engineering, Faculty of Electrical and Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan.

His research interests include blind signal processing, machine learning, blind reverberation time estimation, speech enhancement, multimodal-based approaches for the blind source separation, compressed sensing, and non-negative matrix/tensor factorization for the blind source separation.

MOHAMMAD HASEEB ZAFAR received the Ph.D. degree in electronic and electrical engineering (EEE) from the University of Strathclyde, in 2009. He is currently a Professor with the Faculty of Computing and IT, King Abdulaziz University, Saudi Arabia. He is also a Visiting Researcher with the Department of Electronic and Electrical Engineering (EEE), Centre for Intelligent Dynamic Communications (CIDCOM), University of Strathclyde, Glasgow, U.K.

His main research interests include performance analysis of diverse computer and wireless communication networks and systems. He is particularly interested in design, deployment, and analysis of wireless sensor networks (WSNs), mobile ad hoc networks (MANETs), wireless mesh networks, wireless personal area networks (WPANs), the Internet of Things (IoT), routing, network traffic estimation, software defined networks, machine-2-machine communications, femtocells, and intelligent transportation systems.

THAMER ALHUSSAIN received the master's and Ph.D. degrees in information and communication technology from Griffith University, Australia.

He is currently an Associate Professor with the Department of E-Commerce, Saudi Electronic University (SEU), Saudi Arabia. He is currently a Vice President for Academic Affairs, Saudi Electronic University (SEU). His current research interests include the success of information systems, measuring IS effectiveness, mobile services, and e-learning.