International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

Deep Learning Architectures For Automatic Speech Emotion Recognition: A Review

Meera Mohan, Dr. P. Dhanalakshmi, Dr. R. Satheeshkumar
Ph.D. Scholar: Department of Computer Science and Engineering,Annamalai University
Professor: Department of Computer Science and Engineering, Annamalai University
Associate Professor: Department of Computer Science and Engineering Sahrdaya Engineering College, Trissur

Submitted: 20-03-2022 Revised: 27-03-2022 Accepted: 30-03-2022

ABSTRACT—Human-Computer Interaction (HCI) involves emotion detection from speech signals, which is a tough assignment. The on-demand and leveraging need for accurate and real-time Speech Emotion Recognition (SER) in human–computer interactions necessitates a comparison of available methods and databases in SER to arrive at feasible solutions and a better understanding of this open-ended problem. This study offers an outline of Deep Learning techniques and explores some recent research that has used these analytics to identify speech-based emotions. Deep learning approaches for SER with available datasets are examined in this review, accompanied by traditional machine learning strategies for speech emotion recognition. This review contributes to a better understanding of the field of discrete speech emotion recognition and presents a multi-aspect comparison between different deep learning approaches in speech emotion recognition.

Keywords—Speech Emotion Recognition; deep learning; feature extraction; emotional speech databases

I. INTRODUCTION

Speech Recognition is one of the most burgeoning research field in which attempts are made to identify speech signals.As a result, Speech Emotion Recognition (SER) is becoming a common research subject, with many potential applications in fields such as automatic translation systems, machine-to-human interaction etc. [1]-[2].Humans can easily sense the speaker's emotion. Many years of experience and observation are needed to accomplish this. Humans analyze various characteristics of a particular speech before recognizing the speaker's emotion based on previous experience or observation. A human-like system [3] that can detect emotions effectively and efficiently is demanded.The extraction of features or different characteristics from speech can be used to identify emotion, and quiet a lot of speech databases is required to be trained to make the system accurate. The first step in developing an emotion recognition system is to choose or implement an emotional speech corpora, after which emotion-specific features are drawn-out from those speeches and finally a classification model cast-off to recognize theemotions.

Prior to the widespread use of deep learning, SER relied on techniques such as hidden Markov models (HMM), Gaussian mixture models (GMM), and support vector machines (SVM), as well as extensive pre-processing and precise feature engineering.SER has begun to benefit from the tools made available by deep learning in order to solve all major problems in machine learning [2].

Speech emotion recognition has the potential to be very useful in vehicle safety features. It is capable of sensing the driver's mood and aiding in the prevention of collisions and disasters.Another use is in counselling sessions; by using SER, therapists would be able to understand their patients' current condition as well as any underlying latent emotions.Speech emotion detection in call centers can offer early warnings to customer care and managers about the caller's emotion in the service industry and e-commerce [2].

The organization of this paper is as follows: a background study on SER system and some of the traditional feature extraction techniques is described in section II. Section III aims to give a brief review on emotional speech databases and their comparison. Section IV focus

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

on different SER techniques (traditional and deep learning).Section V summarizes the review.

II. SPEECH EMOTION RECOGNITION SYSTEM

Signal pre-processing, feature extraction, and classification are the three basic components of emotion recognition systems based on digitized speech. To evaluate meaningful units of the signal, acoustic pre-processing such as de-noising and segmentation are used. To define the appropriate features in the signal, feature extraction is used. Finally, classifiers perform the mapping of derived feature vectors to specific emotions [1]. Speech signal processing, feature extraction, and classification are discussed in depth in this section. Fig.1 demonstrates a simplified method for understanding emotions based on expression.

Speech enhancement is carried out in the head start of speech-based signal processing [1], where the noisy components are eliminated. Feature extraction and feature selection [1] are the two sections of the second level. The appropriate features are drawn out from the pre-processed speech signal, and the extracted features are used to make the selection [1].

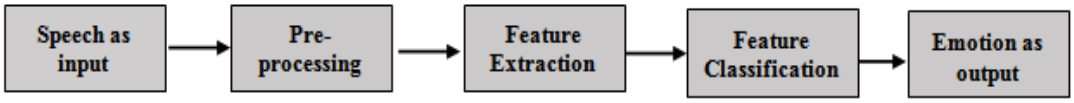


Fig. 1. peech Emotion Recognition System

A. Feature Extraction and Selection Techniquies In SER

During the recording process, the input data obtained for emotion recognition [1] is frequently distorted by noise. The feature extraction as-well-as classification become less reliable [1] as a result of these flaws. This means that in emotion detection and recognition systems, refining the input data is crucial.

Acoustic features, contextual knowledge, linguistic features, and hybrid features are the four types of speech features used in the SER. The most popular and widely used features of SER are acoustic features. Prosodic features (pitch, sound, and duration), as well as voice quality and spectral features, are among them. The harmonics-to-noise ratio, spectral power distribution, first three formants, jitter, and shimmer are all indicators of voice quality. Table I lists some of the factors that focus on acoustics and speech impulses or emotions.

Some of the common methods for feature extraction are Linear Predictive Cepstral Coefficients (LPCC), Fast Fourier Transform (FFT), Linear Predictive Analysis (LPC), Mel scale Cepstral Analysis (MEL), Perceptual Linear Predictive Coefficients (PLP) and Relative Spectra Filtering of log Domain Coefficients (RASTA) [16].

- **Linear Prediction Coefficients (LPC)**: It is a base model for human speech output that employs a typical source filter model. It estimates the concentration and frequency of the left-over residue by approximating the formants, extracting their effects from the speech signal, and analysing the speech signal. Each sample of the signal is claimed to be a direct incorporation of previous samples in the result. The formants are defined by the coefficients of the difference equation, so LPC must approximate these coefficients. LPC is a common formant estimation method and a strong speech analysis method.

Linear Prediction Cepstral Coefficients (LPCC), log area ratio (LAR), Reflection Coefficients (RC), Line Spectral Frequencies (LSF), and Arcus Sine Coefficients (ARCSIN) [118] are some of the other features that can be obtained from LPC. LPC is generally used for speech reconstruction. The LPC approach is widely used in musical and electrical companies to build mobile robots, as well as in telecommunications companies for tonal analysis of violins and other musical gadgets.

- **Mel Frequency Cepstral Coefficients (MFCC)**: It is one of the typical methods used for feature extraction which is being the most common in Automatic Speech Recognition (ASR).

While 10-12 coefficients are adequate for coding expression, it is dependent on the spectral type, which makes it more sensitive to noise. While temporal material is present in speech, this problem can be solved by using more information in speech signals periodicity. MFCC represents the real cepstral of a windowed short time fast Fourier transform (FFT) signal. The frequency is not linear. The audio function extraction MFCC technique [121] is used to extract parameters that are close to those used by humans when hearing speech. Other information is deemphasizes, and speech signals are separated using an arbitrary number of samples with time frames. Most systems use overlapping from frame to frame to smooth the transition, and

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

then hamming window to remove the discontinuities from each time frame.

- **Linear Prediction Cepstral Coefficients (LPCC)**: Emotion- specific information conveyed by vocal tract characteristics are captured by LPCC. The aim of using LPCCs is to take into account the speaker's vocal tract characteristics when doing automatic emotion recognition. LPCC are cepstral coefficients derived from the spectral envelope determined by LPC. LPC determines the spectral envelope, and LPCC are cepstral coefficients resulting from it. The letters LPCC represent for the coefficients of the Fourier transform illustration of LPC's logarithmic magnitude spectrum. Cepstral analysis is commonly used in the field of speech synthesis because of its capacity to separate the speech into its source and system components with a small number of features, without any a priori knowledge about source and / or system. Rosenberg and Sambur [117] discovered that adjacent predictor coefficients are highly correlated, meaning that representations with fewer correlated features are more efficient; LPCC is a good example of this. Atal proposed the relationship between LPC and LPCC in 1974. In the case of minimal phase signals, it is technically reasonably simple to convert LPC to LPCC [119].
- **Perceptual Linear Prediction (PLP)**: Hermansky [120] created a PLP model that models human speech using the psychophysics principle of hearing. PLP increases the speech recognition performance by filtering out irrelevant data. Only the transformation of spectral characteristics to human auditory system match distinguishes PLP from LPC. PLP approximates three major perceptual aspects: the intensity-loudness power-law relation, equal loudness curve, and critical-band resolution curves.
- **Mel scale Cepstral Analysis (MEL)**: PLP and MEL analysis [124] are similar in that they both use psychophysically dependent spectral transformations to change the spectrum.

TABLE I. ACOUSTIC VARIATIONS BASED ON SPEECH EMOTIONS

Emotions	**Pitch**	**Intensity**	**Speaking Rate**
Anger	Stress	Much higher	Marginally faster
Fear	Wide	Lower	Much faster
Joy	High mean, wide range	Higher	Faster
Sadness	Slight Narrow	Downward inflections	Lower

A spectrum is bundled in this system according to the MEL scale, while a spectrum is twisted in PLP according to the bark scale. The key difference between scale cepstral analysis of PLP and MEL is the output cepstral coefficients. The modified power spectrum is smoothed using the all pole model in PLP, and then output cepstral coefficients are computed using this model. In MEL scale cepstral analysis, on the other hand, the modified power spectrum is smoothed using cepstral smoothing. The Discrete Fourier Transform (DFT) is used to directly transform the log power spectrum into the cepstral domain.

- **Relative Spectra Filtering (RASTA)**: RASTA filtering [123] is available in the research library to compensate for linear channel distortions. In either the log spectral or cepstral realms, the RASTA filter may be used. Each function coefficient is effectively passed through the RASTA filter band. In both the log spectral and cepstral realms, linear channel distortions appear as an additive constant. The identical band pass filter's high-pass portion decreases the effect of convolutional noise introduced into the channel. Low-pass filtering assists in the smoothing of spectral variations from frame to frame.
- **Fast Fourier Transform (FFT)**: Speech signal power spectrum defines the frequency content of the signal over time and is one of the most common techniques for analysing speech signals. The speech signal's discrete Fourier Transform (DFT) is the first step in computing the power spectrum, which computes time domain signal equivalent frequency information. Fast Fourier Transform (FFT) can be used to improve the efficiency of real point values in speech signals.[122]

B. Feature classification in SER

Various classifiers have been discussed in the literature to build systems like SER, speech

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

recognition, and speaker verification. A classification system is a tool for assigning a particular emotion class to each speech based on the extracted features from the speech. For emotion recognition, various classifiers are available. When it comes to selecting a correct classifier, there are no thumb rules. The majority of the time, the classifier was selected based on previous experience. Each speech sample's features (feature vector) are fed into classifiers as an input using a linear combination of real weight vector W.

The weight vector is then modified using a proper training technique. An activation function is used to produce the output from the model, which mapped each input to a predefined emotion class. This activation function can be either linear or nonlinear. Classifiers can be classified into two groups based on the existence of their activation functions: linear classifiers and non-linear classifiers. If the feature vectors are linearly separable, the linear classifier can classify correctly. Since most feature vectors are not linearly separable in real life, a nonlinear classifiers used.

III. REVIEW ON EMOTIONAL SPEECH DATABASES

Mainly databases specifically designed for speech emotion recognition are classified into three, simulated, semi-natural, and natural speech collections [2].

- **Simulated Database**: The speech data in these databases was captured by well-trained and experienced performers [1].Of all databases, this one is considered to be the most comprehensive method of obtaining a speech-based dataset of various emotions [1]. This approach is believed to be responsible for approximately 60% of all speech databases.Simulated data sets, such as "EMO-DB (German), DES (Danish)[2], The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)[2],Toronto Emotional Speech Set (TESS)[2], and Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)"[2], are structured collections of emotions that allow for easy comparison of outcomes [2]. While they have a large number of distinct emotions, since they have manufactured emotions, they have over fitted models for recognizing emotions that are somewhat different from what happens in normal communication [2].
- **Semi-natural/ Induced database**: Another type of database is one in which the emotional collection is gathered by simulating an emotional situation [1]. This is achieved without the actor or speaker's knowledge. This database is more naturalistic than an actor-based database. However, there could be an ethical concern because the speaker should be conscious that they are being documented for research purposes. IEMOCAP, Belfast, and NIMITEK are three semi-natural emotion sets. This category has the benefit of being somewhat close to natural speech utterances. Despite the fact that they are based on scenarios and the speech happens in a context, they are artificially generated emotions, particularly when speakers are aware that they are being recorded for research purposes.

Natural database:These databases are difficult to obtain due to the difficulty in identifying them, despite the fact that they are the most realistic. Normal emotional expression databases are commonly assembled from public conversations, call centre conversations, and other similar circumstances. The natural corpora of emotional speech databases such as VAM, AIBO, and call centre data make up this category. These are totally real and can be used to build emotion recognition models without fear of being artificially produced. However, due to the continuousness of emotions and their complex variance throughout the course of the expression, the presence of concurrent emotions, and the presence of background noise, modelling and identification of emotions with this type of datasets can be difficult.

A. Studyof Different Types Of Databases

Table II shows a brief analogy of different types of databases, pointing out the differences in the features and some examples of each type [2]. Major examples of each type is discussed in this section.

- **Berlin Database of Emotional Speech (EMO-DB)[**2]:One among most commonly used databases for speech emotion detection is the Berlin Database of Emotional Speech (EMO-DB) [2] [55]. BJ Abbaschian et.al [2] observed that"EMO-DB is a simulated dataset made up of 10 German sentences, five short sentences, and five long sentences. The dataset was created using ten speakers, five females and five males. Each speaker had spoken ten sentences, five long and five short, voicing various emotions [2]. The whole dataset is containing 700 samples, of 10 sentences acted with seven emotions. The emotions chosen for this dataset were neutral, anger, fear, joy, sadness, disgust, and boredom" [2].
- **Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)**: IEMOCAP is a semi-natural English audio visual dataset of 1150 Utterances acted by ten speakers, five males and five females [2].Abbaschian et.al [2] observed"anger, happiness, sadness, and frustration were the original emotions in database scenarios.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

Later, they added four more categories to the data: disgust, fear, excitement, and surprise [2]. They also provided data that was labelled with continuous valence, activation, and dominance qualities [2]. The dataset is not open source and requires a license to use. Compared to the simulated databases, this semi-natural dyadic database emits more naturally emitted emotions. It will be one of the best datasets for deep learning applications, with each dialog lasting an average of 5 minutes" [2].

- **Vera am Mittag Database (VAM)**: VAM dataset is a real audio visual dataset based on dialogues from Vera am Mittag, the German TV talk show [2] [58]. It is made up of three components: valence, activation, and dominance, and it is used to recognize dimensional speech emotions. With 1018 audio utterances, 47 speakers from the show performed the audio portion of the dataset.

TABLE II. COMPARISON OF SPEECH EMOTION DATABASES

Database	**Language**	**Emotions**	**Source**
EMO-DB	German	Happiness, sadness, neutral, disgust, anger	Professional Actors
IEMOCAP	English	Anger, sadness, surprise, joy, frustration, fear, neutral, excited	Professional Actors
VAM	German	valence, activation, and dominance	TV shows, call centers
DES	Danish	neutral, surprise, happiness, sadness, and anger	Professional Actors
RAVDEES	English	happy, sad, angry, fearful, surprised, disgusted, calm, and neutral	Professional Actors
TESS	English	Angry, pleasantly surprised, disgusted, happy, sad, fearful, and neutral.	Professional Actors

IV. SPEECH EMOTION RECOGNITION METHODS

To recognize emotion from voice, a number of methods and algorithms are used [2]. Each of these approaches seeks to solve the problem from a different viewpoint, and each has its own package of advantages and disadvantages. In this section, we will briefly review some of the traditional SER methods followed by reviewing deep learning approaches for the problem of Speech Emotion Recognition [2].

A. Traditional Methods

These methods were designing foundations based on machine learning algorithms

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

that necessitated extensive feature engineering and a detailed understanding of the subject matter in order to infer the features that would be most useful in the calculations. Here three traditional methods SVM[17], GMM [47], and HMMs [48] are reviewed.

- **Hidden Markov Models (HMM)**:The Hidden Markov model is a double random process with a hidden underlying process that can be observed using a separate set of random processes that are responsible for generating the observed label series. In speech processing, HMM is commonly used. They're designed to create a scalable model that conforms to the temporal features of speech, allowing them to classify slight variations in an audio signal. HMMs is, of course, one of the first things we tried in SER [2]. MFCCs and LPCCs are two frequency representations that mostly commonly attempted by the researchers to solve the problem.
- **Gaussian Mixture Model (GMM)**:By combining linearly multivariate Gaussian distributions, the Gaussian Mixture Model is used to produce the probability density function of feature vector, ~x, which is a D dimensional continuous valued data vector. During the calculation of the log-likelihood, the classic GMM method assumes that all of the input features (MFCC filters, pitch, deltas, and delta-deltas) have equal weight. On both tasks, this is most likely not the case. The majority of speaker-specific information is carried in the upper part of the frequency range, and pitch holds more information than the lower part of the frequency band. This is also applicable when it comes to recognizing emotions.
- **Support Vector Machine (SVM)**:SVM is a well-known method for emotion recognition and is used as a classifier [2]. SVM is most commonly used for classification and regression. They do classification by building an N-dimensional hyper plane that divides the data into categories as efficiently as possible. The classification in the input feature space of the dataset is obtained using a linear or nonlinear separating surface. The main idea behind SVM is to use a kernel function to transform the original input set into a high-dimensional feature space and then achieve optimum classification in that space[2][42].

B. Deep Learning Methods in SER

Deep learning [1] is a modern machine learning research area that has gotten a lot of attention in recent years. Deep Neural Networks (DNNs) have been used by a few researchers to train their SER models. Table IIIdifferentiatesconventional algorithms and Deep learning [1] i.e. Deep Convolutional Neural Networks (DCNN) algorithms in the context of measuring various emotions using the IEMOCAP, Emo-DB [1] datasets and recognizing various emotions such as happiness, anger, and sadness. Deep learning algorithms outperform traditional techniques in emotion recognition, according to research. Deep learning approaches are made up of a variety of nonlinear components that perform parallel computations. To address the shortcomings of other approaches, these methods must be organized with deeper layers of architecture. Deep learning is currently a leveraging research area due to its multi-layered architecture and reliable results delivery.

In this section various deep learning algorithms such as DBMs, DBNs, CNNs, DCNNs, LSTM, RNNs, RvNNs, AEs, MTLs, GAN, Transfer Learning and Attention Mechanism [1] are discussed.

- **Deep Boltzmann Machine (DBM)**: DBMs [1] are made up of several hidden layers and are derived from "Markov Random fields". These layers are made up of stochastic entities and variables selected at random.Khalilet.al [1] shows "the domain of visible entities is given by$v_i\{0, 1\}\in^C$, with combination of hidden units $h_i^{(1)}\in\{0, 1\}^b{}_1$, $h_i^{(2)}\in\{0, 1\}^C{}_2$, · · ·, $h_i^{(L)}\in\{0, 1\}^C{}_{L.}$". This is shown in Fig.2 [1]. In a Restricted Boltzmann Machine (RBM), on the other hand, there is no interconnection between entities in the same layer.

The main benefit of DBM is that it has a proclivity for learning quickly and providing efficient representation [1]. It accomplishes this through layer-by-layer pre-training. When it comes to emotion recognition, DBM can facilitate when speech is used as an input. DBM also has several other functions. There are also drawbacks, such as limited effectiveness in certain scenarios.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

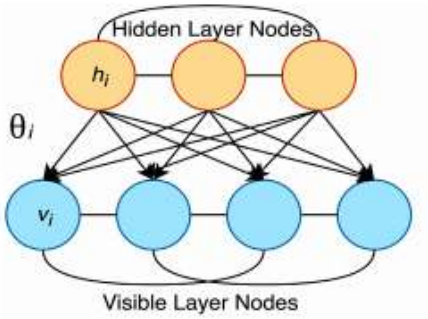


Fig. 2. Graphical representation of DBM

- **Recurrent Neural Network (RNN)**: RNN [1] is a type of neural network that is based on sequential information [1] in which the outputs and inputs are linked. This interdependence is usually helpful in predicting the state of the input in the future. Khalil et.al [1] summarise that "RNNs, like CNNs, require memory to save the overall informationgathered during the sequential training process of deep learning modelling, that generally only works well for a few back-propagation steps. Fig.3 [1] depicts the basic RNN architecture in which x_t is the input, $s_t$is the underlying 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:

$$S_t = f\big(U(x_t) + W_s(t-1)\big) \quad (1)$$

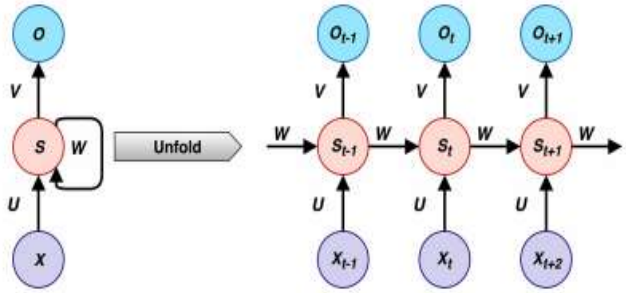


Fig. 3. Basic architecture of RNN

The key issue that has an impact on the RNN's overall performance is its vulnerability to the absence of gradients. In other words, during the training process, the gradients can decay exponentially and be multiplied by a lot of small or large derivatives. However, this sensitivity decreases, resulting in the forgetting of the initial inputs. To prevent this from occurring, Long Short-Term Memory (LSTM)[1] is used to provide a block between the recurrent connections. Khalil et.al [1] summarised "each memory block stores the network's temporal states and contains gated units that monitor the inflow of new data. Since residual connections are normally very large, they are useful for reducing gradient issues".

- **Recursive Neural Network(RvNN)**:RvNN[1]is a hierarchical deep learning technique that doesn't rely on a tree-structured input sequence. By splitting the input into small chunks, it can quickly learn the parse tree of the given data. Fig.4 depicts an RvNN [1]. The governing equation

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

where W is designed as n × 2n weighted matrix[1].

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

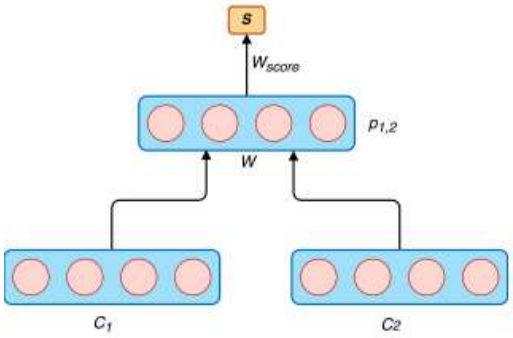


Fig. 4. Basic architecture of RvNN

RvNN [1] is mainly used for natural language processing (NLP), but its architecture enables it to handle a number of modalities, including speech recognition and SER. It begins by measuring the overall score of each possible pair in order to merge them up and create a syntactic tree. The highest-scoring pair is then combined with a vector called the compositional vector [1]. After the pair is combined,the RvNN produces several units, the vectors that represent regions, and labels for classification.

- **Deep Belief Network (DBN)**:The structure of DBN is even more complex, and it is made up of cascaded RBM structures [1]. Khalil et.al[1] discussed the summary as "DBN is a bottom-up extension of RBMs, in which RBMs are trained layer by layer [1] Due to their ability to learn the recognition parameters effectively, DBNs are widely used for speech emotion recognition, even though there are a large number of them. It also removes layer non-linearity. During preparation, DBNs are used to solve slow speed localized problems using back propagation algorithms. Fig.5 depicts the DBN's layer-wise architecture, in which RBMs are trained and evaluated layer by layer [1] from the bottom to the top"[1].

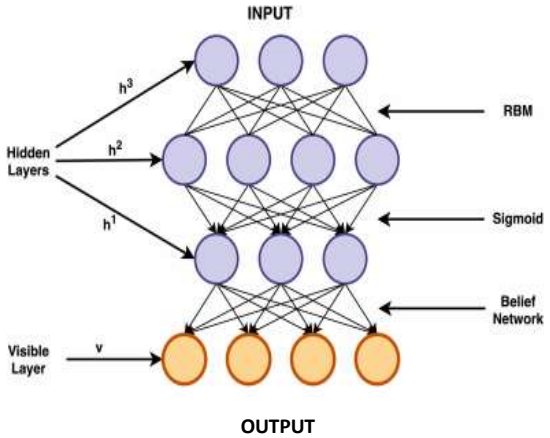


Fig. 5. Layer-wise architecture of DBN

The unsupervised design of pre-training techniques [1] with broad and unlabelled databases is the first big benefit of DBN. The second benefit of DBNs is that they can use inference method approximation to compute the necessary output weight of the variables [1]. Mohamed et.al [65] states "since DBNs' inference method is only limited to bottom-up pass, there are some limitations. A greedy layer exists that learns the features of a single layer and never re-adjusts with the other layers" [1].

- **Convolutional Neural Network (CNN)**: CNN is a form of Deep learning technique for classification that is solely based on feed-forward architecture. CNNs are widely used to enhance data classification and pattern recognition. The input data is interpreted in the form of receptive fields by these networks, which have small size neurons on every layer of the built model architecture [1].

Fig. 6[1]provides "the layer-wise architecture of a basic CNN network. Filters are the base of local

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

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 dot product between the weights and provided inputscomputed by the convolutional layers" [1].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:

$$z_i = g(W_i * r + n_i) \quad (3)$$

To obtain the output of the convolution layers, an activation function f or a non-linear methodology must be used. It should be noted that, as seen in Fig. 5, inputs are very tiny portions of the original volumes. To feature maps and reduce network parameters, down sampling is performed at each subsampling layer [1]. As a consequence, over fitting is minimized and the training phase is accelerated. For the adjoining expanse of all the function maps, the pooling process is carried out over p elements (also known as filter size). As with other neural networks, the layers must be completely linked in the final step.These later layers create high-level abstraction from the input speech data using the previous low- and mid-level features [1]. The final layer, also known as SVM or Softmax, [1] is used to produce a classification score in probabilistic terms that is connected to a particular class.

- **Deep Convolutional Neural Network (DCNN)**: Deep convolutional networks usually have many layers of convolution nodes, followed by one or more completely connected layers to complete the classification task [2].

Harar et al. [26] [2] have proposed "a method based on a deep neural network containing convolutional pulling and fully connected layers. They used the Berlin Database of Emotional Speech to test their system. They've confined their classes to angry, neutral, and sad to compare to previous studies [2]. They eliminated silence from their signals and then split the files into 20 ms chunks with no overlap in their method [2]. They have six layers of convolution in their network before any feature selection, followed by dropout layers with p values with 0.1, a lattice of two parallel feature selectors, and finally a sequence of completely connected layers [6]. Their system had a section accuracy of 77.51%, but a file level accuracy of 96.97%, with a confidence rate of 69.55%" [2]. Although the system's file-level accuracy was high, there is no indication to point a chunk of speech in real-world situations, and the system's independent detection needs to be improved [2].

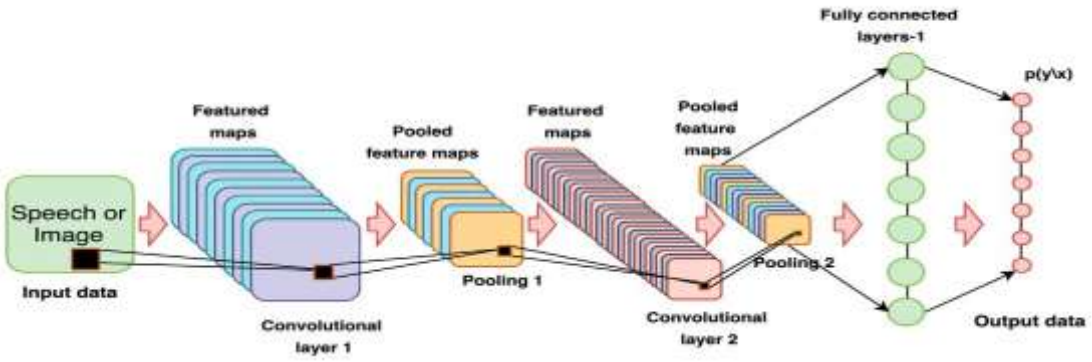


Fig. 6. Layer-wise architecture of CNN

Another Emotion recognition system based on a deep convolutional neural network is developed by Zhang et al. [25]. "They can identify three types of emotions (angry, sad, and happy) as well as a neutral category using this method" [2]. Furthermore, they have demonstrated that their device can achieve accuracies of over 80% using EMO-DB, which is around 20% higher than the baseline SVM standard [2]. Automatic feature selection in deep convolutional neural networks outperforms feature selection in shallow convolutional neural networks and statistical model-based methods like GMM and HMM [2], according to their framework [2]. The discriminant temporal pyramid matching (DTPM) strategy, which helps in concatenating the learned segment level feature [2] to form an utterance level feature representation, was one of the main features used in this system.

- **Long Short-Term Memory Network (LSTM)**: Forming short-term activation for recent events [2] in RNNs make them beneficial for

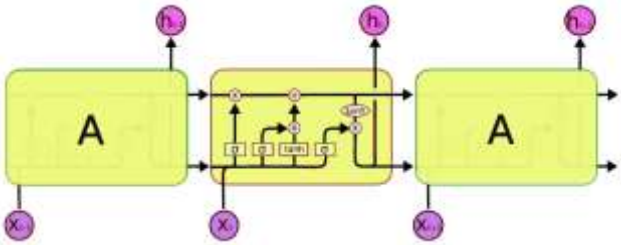
International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

applications in which time is an essential feature, like Speech Processing, music composition, and video description. However, depending on the size of the weights, error signals flowing backward in time [2] may either get bigger and bigger or disappear as they are trained using Back Propagation over Time. This would either result in oscillating weights or cause the network to train and converge slowly.

Hochreiter and Schmidhuber [80] developed a new architecture called Long ShortTerm Memory [2] in 1997 to integrate the short-term adaptation of RNNs and prevent the problems described above [2]. Even if the input sequences are incompressible and noisy, LSTM networks can bridge time intervals greater than 1000 steps. They're using a gradient-based algorithm that enforces constant error flow across individual units that are explicitly designed to handle the short-term, allowing them to truncate gradient computations at a given point without affecting long-term activations [2].

Fig.7 depicts architecture of LSTM network. Wöllmer et al. [28] proposed a multimodal LSTM-based classification network that takes into account acoustic, linguistic, and visual data [2]. They contrasted both unidirectional and bidirectional LSTM networks in their research.

Trigeorgis et al. [29] introduced a context-aware method for end-to-end emotion recognition in speech using CNNs and LSTM networks later in 2016. They do not pre-select features before training [2] the network, which is a major difference from other deep learning algorithms.



. Fig.7. Architecture of LSTM

A typical LSTM network is comprised of different memory blockscalledcells(the rectangles)**.** There are two states that are being transferred to the next cell; the **cell state** and the **hidden state**. The memory blocks are responsible for remembering things and manipulations to this memory is done through three major mechanisms, called **gates.**

Xie et al. [115] proposed a structure focused on two layers of modified LSTMs with 512 and 256 hidden units, followed by a layer of attention weighting on both time and function measurements, and two fully connected layers at the end in a paper published late in 2019 [2]. They tested five different variations of their proposed methods: LSTM with Time attention, LSTM with feature attention, LSTM with both time and feature attention, LSTM [2] with updated forget gate, and LSTM with both time and feature attention. Furthermore, according to the findings of their English speech dataset eNTERFACE, they have achieved UAR accuracy of 89.6%.

Due to their pattern history memorizing [2] capability, LSTM networks have proven to be very successful in time series data [2]. Speech emotion detection is one of the standard implementations of such a device. LSTM-based systems are highly capable of understanding the signal's spectral characteristics. They may form a competent framework to model and learn the samples when combined with CNNs to learn the temporal characteristics of the signal [2].

- **Auto Encoder (AE)**: One of the most important goals of feature extraction, which is one of the most important tasks in classification, is to find a stable data representation in the presence of noise. AE is a set of unsupervised machine learning methods that can be used in this way. In general, an AE network consists of two components: an encoder and a decoder. The encoder learns to create a copy of the input that is as similar to the output as possible; hence, the input and output dimensions are the same. In literature, several versions of autoencoders have been proposed. Among them, Variational Auto Encoder (VAE), Denoising Auto Encoder (DAE), Sparse Auto Encoder (SAE), Adversarial Auto Encoder (AAE) are very popular and useful in SER. The generalized architecture of AE is depicted in Fig.8[1].

In 2018, Latif et al. [31] were the first to suggest VAEs [2] as a method for evaluating the latent representation [1] of speech signals and using this representation to intuitively identify emotions using deep learning. The VAE takes advantage of data's log-likelihood and affects the lower bound

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

estimator from a given graphical prototype [1] with unbroken underlying variables.

In the same year, Eskimez et al. [32] used feature learning strategies such as Denoising AutoEncoder (DAE), Variational AutoEncoder (VAE), Adversarial AutoEncoder (AAE), and Adversarial Variational Bayes (AVB) to boost the efficiency of CNN-based SER systems. Since autoencoder is an unsupervised learning process, it is not constrained by the number of labeled data samples accessible.

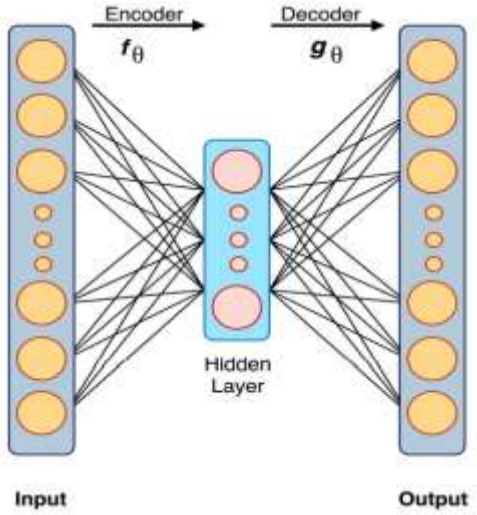


Fig.8. Architecture of Auto Encoder

They came to the conclusion that autoencoder frameworks could increase the F1 score and unweighted accuracy rating of automatic SER systems for SVM and CNN frameworks.

- **Multitask Learning(MTL)**:By exchanging representations between related auxiliary tasks at the same time, the Multitask learning (MTL) approach helps the model to generalize better on our primary task.

Fig.9depicts Example of a MTL network architecture with two tasks and two outputs. In SER, speaker characteristics such as gender and age may affect how emotion can be expressed; therefore, they can be considered as Meta-information in MTL.For the first time, Kim et al. [33] used gender and naturalness as auxiliary tasks for deep neural networks in an MTL manner. Their proposed approach provides high-level feature representation, allowing for the identification of discriminative emotional clusters. They present the results of their practice with six corpora, both within and across corpora.

- **Generative Adversarial Networks (GANs)**: Since 2014, when Goodfellow et al. [85] proposed them for the first time to learn and mimic an input data distribution, GANs have been considered as data augmentation, data representation, and denoising resources in deep learning.Latif et al. [86] used generative adversarial networks to boost the SER system's robustness in 2018.

A generative adversarial network (GAN) has two parts:The generator learns to generate plausible data. The generated instances become negative training examples of the discriminator.

The discriminator learns to distinguish the generator's fake data from real data. The discriminator penalizes the generator for producing implausible results [125].

The generator G(z) produces synthetic data from an input, z, which is a sample from the probability distribution P(z). The discriminator, on the other hand, takes the information and decides whether the input data is real or artificial.

Finally, both networks reach an equilibrium in a way that they have a value function that one agent seeks to maximize, and the other tries to minimize as an objective function shown in the equation below:

$$\min_G \max_D V(D,G) = \mathop{E}_{x\sim P(x)}[\log D(x)] + \mathop{E}_{z\sim P(z)}[\log(1 - D(G(z)))] \quad (4)$$

where D(x) and D (G(z)) are the probabilities that x and G(z) are inferred to be real samples by the discriminator.

The GAN [2]“functions unsupervised and separately from the class mark of the real data in this system. Following that, Conditional GAN [88] was suggested, which is based on the premise that

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

GAN can be conditional by using Class labels and data from various modality or portions of the data". In conditional GAN, the objective function of a two-player minimax game would be as follows [2]:

$$\min_G \max_D V(D,G) = \mathop{E}_{x\sim P(x)}[\log D(x|y)] + \mathop{E}_{z\sim P(z)}[\log(1 - D(G(z)|y))] \quad (5)$$

In this equation, "y" is the class label of the data. One of the most significant shortcomings of GAN approaches is their dependency on data and initialization for convergence. However, the GAN is initialized using a pertrained auto encoder to overcome this limitation and achieve faster convergence [89]. Furthermore, their proposed method for learning common features between minority and majority classes. Both classes are used to fine-tune GAN.

- **Transfer Learning**: It is a popular approach in deep learning where pre-trained models are used as the starting point on computer vision and natural language processing tasks given the vast compute and time resources required to develop neural network models on these problems and from the huge jumps in skill that they provide on related problems [126].

The cross-domain difficulty of SER, i.e., test corpora do not fit train corpora, can be overcome with Transfer learning. Song et al. [91] use transfer learning to practice dimension reduction and Maximum Likelihood Estimation in a cross-corpus speech emotion recognition task. Mean difference embedding optimization and SVM as a classifier tool were used to obtain two adjacent latent feature spaces for the source and goal corpora.

EMO-DB was used as the source corpus, with five emotion categories, and a Chinese emotion dataset with the same emotion categories as the test corpus.

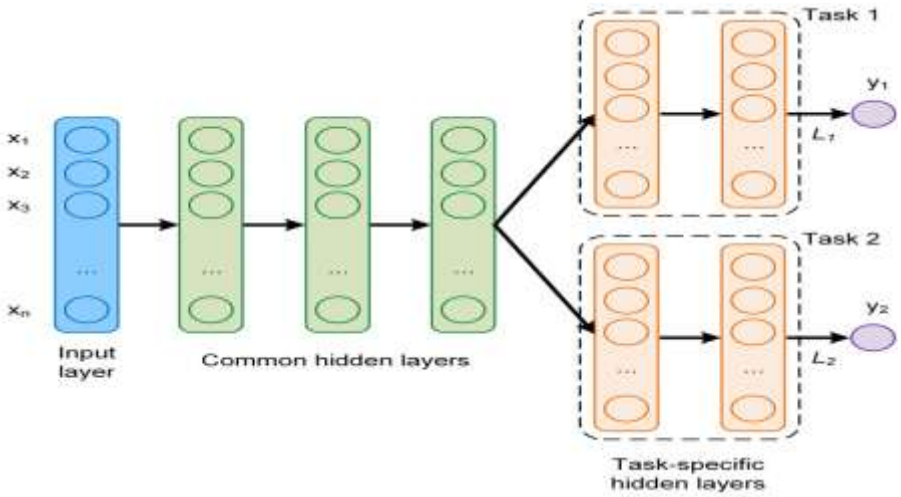


Fig.9. Architecture of MTL

Principal component analysis (PCA) and local preserving projection were used to minimize dimension in two of the proposed models. As a consequence, neutral has the highest acceptance rate, followed by satisfaction and fear. However, the proposed method outperformed the automatic recognition approach in terms of recognition rate.

One important constraint in transfer learning is the size of training and test data, i.e., the number of training data for transfer learning should be small enough to avoid over fitting the model. Furthermore, since train and test datasets are not distinct and identically distributed in fact (i.i.d.), algorithms like PCA and LDA perform poorly.

Song [94] offered"Transfer Linear Subspace Learning (TLSL), generalized linear subspace learning, and transfer learning to consider thedifference between train and test set, and application on several benchmark datasets". The system has been applied cross-corpus to the EMO-DB, eNTERFACE, and FAU Aibo databases in six different ways, and the results show that using TLSL enhances efficiency over baseline methods. Unlike previous approaches to transfer learning, the emphasis is not exclusively on informative elements, and less informative parts are not ignored.

- **Attention Mechanism**: A neural network is considered to be an effort to mimic human brain actions in a simplified manner. In deep neural networks, the Attention Mechanism is an effort to apply the same behavior by selectively focusing on

International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 3 Mar 2022, pp: 1282-1301 www.ijaem.net ISSN: 2395-5252

a few important items while avoiding others. Usually, all positions of a given utterance receive equal attention in deep learning methods for SER; however, emotion is not uniformly distributed throughout the utterance for each sample. The attention mechanism considers the precise positions of given samples based on the attention weights allocated to each section of the data that contains an emotionally significant component.

Instead of conventional low-level descriptors (LLD) and high-level statistical aggregation functions (HSF), Mirsamadi et al. [91] used "bidirectional LSTM with a weighted-polling" approach to find more insightful features about emotion. The attention mechanism [2], which allows the network to concentrate on emotionally relevant parts of a sentence while ignoring silent frames of utterance, was the inspiration for this process.

V. CONCLUSION AND FUTURE DIRECTIONS

In this literature, various emotional SER methodologies and the associated speech databases have been reviewed.Several SER publications also reviewed and tried to cover all the major deep learning techniques used for the task of SER, from DBMs to LSTMs to attention mechanisms. Deep learning techniques have many disadvantages, including a broad layer-wise internal architecture, lower efficiency for temporally varying input data, and over-learning during layer-wise information memorization [1].

Table III gives a brief comparison of all the algorithms reviewed, containing the highest accuracy reported for each dataset, all the features used to train the system, methods used and if applicable and number of the layers in each method [2].

TABLE III. COMPARISON OF SPEECH EMOTION DATABASES

Methodology/ no of layers	Features Extracted	Dataset & Accuracy	References
HMM	• 1st and 2nd derivative of F0, • 1st derivative of F1 • 2nd derivative of MBE4, 2nd derivative of MBE5 • MFCC [2]	• DES: 99.5%	[6]
SVM	• 1st and 2nd derivative of F0, • 1st derivative of F1 • 2nd derivative of MBE4MFCC [2]	• EMO-DB: 93.75%	[17]
CNN/2	• PCM	• TEDLIUM2: 66.1%	[21]
DCNN/6	• MFCC	• SAVEE: 65.83% • RAVDESS: 75.83% • THAI: 96.60%	[79]
DCNN/10	• PCM	• EMO-DB: 96.97%	[26]
DCNN, LSTM/4	• PCM	• RECOLA: 68.4%	[29]
DCNN, LSTM/5	• PCM • Log-Mel Spectrogram	• EMO-DB: 95.33% IEMOCAP: 86.16%	[30]
VAE, LSTM/ 2, 4	• Log-Mel Spectrogram	• IEMOCAP: 64.93%	[31]
CNN, VAE/5, 6, 4, 10, 5	• Log-Mel Spectrogram	• IEMOCAP: 48.54%	[32]
LSTM, MTL/3, 3, 2	• F0, voice probability, • zero-crossing-rate • 12 MFCCs with energy and their first-time derivatives [2]	• EMO-DB: 92.5% • LDC: 56.4% • IEMOCAP: 56.9%	[33]

LSTM, GAN/2	• eGeMAPS features	• IEMOCAP: 53.76%	[86]
GAN, SVM	• 1582-dimensional openSMILE feature space	• IEMOCAP: 60.29%	[88]
LSTM, ATTN/4, 3, 3, 4, 4, 4	• 57-dimensional magnitude FFT vectors • F0, voice probability, frame energy, ZCR • 12 MFCCs and Delta	• IEMOCAP: 63.5%	[91]
CNN BLSTM, ATTN/2, 2	• 800 point STFT Mel scale spectrogram, Deltas, Delta deltas [2]	• IEMOCAP: 82.8%	[92]

ACKNOWLEDGMENT

The authors acknowledge with thanks Department of Computer Science and Engineering, Annamalai University for the technical and intellectual support provided for the successful completion of this study.

REFERENCES

[1] Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech Emotion Recognition Using Deep Learning Techniques: A Review. IEEE Access 2019, 7, 117327–117345.

[2] Babak Joze Abbaschian, Daniel Sierra-Sosa,Adel Elmaghrabhy. "Deep Learning Techniques for Speech Emotion Recognition, from Databases to models", Sensors, 2021.

[3] 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.

[4] 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.

[5] M.Chen,P.Zhou,andG.Fortino,''Emotioncommunicationsystem,'' IEEE Access, vol. 5, pp. 326–337, 2016.

[6] K. Elissa, "Title of paper if known," unpublished.

[7] Lin, Y.L.; Wei, G. Speech emotion recognition based on HMM and SVM. In Proceedings of the 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 18–21 August 2005; Volume 8, pp. 4898–4901.

[8] 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.

[9] 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.

[10] 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.

[11] K.R.Scherer,''Whatareemotions?Andhowcantheybemeasured?'' Social Sci. Inf., vol. 44, no. 4, pp. 695–729, 2005.

[12] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, ''Emotion analysis in man-machine interaction systems,'' inProc.Int.WorkshopMach.Learn.MultimodalInteract.Springer,2004, pp. 318–328.

[13] 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.

[14] O. Kwon, K. Chan, J. Hao, T. Lee, ''Emotion recognitionby speech signal,'' in Proc. EUROSPEECH, Geneva, Switzerland, 2003, pp. 125–128.

[15] R. W. Picard, ''Affective computing,'' Perceptual Comput. Sect., Media Lab., MIT, Cambridge, MA, USA, Tech. Rep., 1995.

[16] 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.

[17] Chavhan, Y.; Dhore, M.; Pallavi, Y. Speech Emotion Recognition Using Support Vector Machines. Int. J. Comput. Appl. 2010, 1, 86–91

[18] M.ElAyadi,M.S.Kamel,andF.Karray,''Surve yonspeechemotion recognition: Features, classification schemes, and databases,'' Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011.

[19] A.D.DileepandC.C.Sekhar,''GMM-basedintermediatematching kernelforclassificationofvaryinglengthpatter nsoflongdurationspeech using support vector machines,'' IEEE Trans. neural Netw. Learn. Syst., vol. 25, no. 8, pp. 1421–1432, Aug.2014.

[20] L.DengandD.Yu,''Deeplearning:Methodsan dapplications,''Found. Trends Signal Process., vol. 7, nos. 3–4, pp. 197–387, Jun. 2014.

[21] Bertero, D.; Fung, P. A first look into a convolutional neural network for speech emotion detection. ICASSP 2017, 5115–5119.

[22] J. Schmidhuber, ''Deep learning in neural networks: An overview,''Neural Netw., vol. 61, pp. 85–117, Jan. 2015.

[23] T. Vogt andE.André,''Comparingfeaturesetsforacted andspontaneous speech in view of automatic emotion recognition,'' in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2005, pp.474–477.

[24] C.-N.Anagnostopoulos,T.Iliou,andI.Giannoukos ,''Featuresandclassi- fiers for emotion recognition from speech: A survey from 2000 to 2011,'' Artif. Intell. Rev., vol. 43, no. 2, pp. 155–177,2015.

[25] Zhang, S.; Zhang, S.; Huang, T.; Gao, W. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Trans. Multimed. 2018, 20, 1576–1590.

[26] Harár, P.; Burget, R.; Kishore Dutta, M. Speech Emotion Recognition with studies. In Proceedings of the 4th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 2–3 February 2017; pp. 137–140.

[27] E.Mower,M.J.Mataric,andS.Narayanan,''Afr ameworkforautomatic human emotion classification using emotion profiles,'' IEEETrans. Audio, Speech, Language Process., vol. 19, no. 5, pp. 1057–1070, Jul. 2011.

[28] Wöllmer, M.; Kaiser, M.; Eyben, F.; Schüller, B.; Rigoll, G. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. Image Vis. Comput. 2013, 31, 153–163.

[29] Trigeorgis, G.; Ringeval, F.; Brueckner, R.; Marchi, E.; Nicolaou, M.A.; Schüller, B.; Zafeiriou, S. Adieu Features? End-To-End Speech Emotion Recognition Using A Deep Convolutional Recurrent Network. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016.

[30] Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D and 2D CNN LSTM networks. Elsevier Biomed. Signal Process. Control 2019, 47, 312–323.

[31] Latif, S.; Rana, R.; Qadir, J.; Epps, J. Variational Autoencoders for Learning Latent Representations of Speech Emotion: A Preliminary Study. arXiv 2018, arXiv:1712.08708.

[32] Eskimez, S.E.; Duan, Z.; Heinzelman, W. Unsupervised Learning Approach to Feature Analysis for Automatic Speech Emotion Recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 5099–5103.

[33] Kim, J.; Englebienne, G.; Truong, K.P.; Evers, V. Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning. arXiv 2017, arXiv:1708.03920

[34] J.Han,Z.Zhang,F.Ringeval,andB.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.

[35] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, ''A survey of affect recognitionmethods:Audio,visual,andsponta neousexpressions,''IEEE Trans. Pattern Anal. Mach. Intell., vol. 31, no. 1, pp. 39–58, Jan.2009.

[36] T. Vogt, E. André, and J. Wagner, ''Automatic recognition ofemotions fromspeech:Areviewoftheliteratureandrecom mendationsforpractical realisation,'' in Affect and Emotion in Human-Computer Interaction. Springer, 2008, pp. 75–91.

[37] J. Deng, S. Frühholz, Z. Zhang, and B. Schuller, ''Recognizingemotions fromwhisperedspeechbasedonacousticfeatur etransferlearning,''IEEE Access, vol. 5, pp. 5235–5246,2017.

[38] S.DemircanandH.Kahramanlı,''Featureextraction fromspeechdata for emotion recognition,'' J. Adv. Comput. Netw., vol. 2, no. 1, pp. 28–30, 2014.
[39] F. Dellaert, T. Polzin, and A. Waibel, ''Recognizing emotion inspeech,'' in Proc. 4th Int. Conf. Spoken Lang. Process., Oct. 1996, pp. 1970–1973.
[40] Y.Zhou,Y.Sun,J.Zhang,andY.Yan,''Speechemotionrecognitionusing both spectral and prosodic features,'' in Proc. IEEE Int. Conf. Inf. Eng. Comput. Sci. (ICIECS), Dec. 2009, pp.1–4.
[41] S.Haq,P.J.Jackson,andJ.Edge,''Audio-visualfeatureselectionand reduction for emotion classification,'' in Proc. Int. Conf.Auditory-Visual Speech Process. (AVSP), Tangalooma, QLD, Australia,2008.
[42] M.Alpert,E.R.Pouget,andR.R.Silva,''Reflections ofdepressionin acoustic measures of the patient's speech,'' J. Affect. Disorders, vol. 66, no. 1, pp. 59–69, 2001.
[43] D.VerveridisandC.Kotropoulos,''Emotionalspeechrecognition:Resources, features, and methods,'' Speech Commun., vol. 48, no. 9, pp. 1162–1181, 2006.
[44] S. Mozziconacci, ''Prosody and emotions,'' in Proc. Int. Conf.Speech Prosody, 2002, pp. 1–9.
[45] J.Hirschberg,S.Benus,J.M.Brenier,F.Enos,andS.Friedman,''Distinguishing deceptive from non-deceptive speech,'' in Proc. 9th Eur. Conf. Speech Commun. Technol. (INTERSPEECH), 2005, pp.1833–1836.
[46] O. Pierre-Yves, ''The production and recognition of emotions inspeech: Features and algorithms,'' Int. J. Hum.-Comput. Stud., vol. 59, nos. 1–2, pp. 157–183, 2003.
[47] D. Neiberg, K. Elenius, and K. Laskowski, ''Emotion recognitioninspontaneous speech using GMM,'' in Proc. Int. Conf. Spoken Lang. Process. (ICSLP), 2006, pp. 809–812.
[48] A. D. Dileep and C. C. Sekhar, ''HMM based intermediate matchingkernelforclassificationofsequentialpatternsofspeechusingsupportvector machines,'' IEEE Trans. Audio, Speech, Language Process., vol. 21, no. 12, pp. 2570–2582, Dec.2013.
[49] G.Vyas,M.K.Dutta,K.Riha,andJ.Prinosil,''Anautomaticemotion 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.
[50] S. Wang, X. Ling, F.Zhang, and J. Tong, ''Speech emotionrecognition 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.
[51] Y. Pan, P. Shen, and L. Shen, ''Speech emotion recognition usingsupport vector machine,'' Int. J. Smart Home, vol. 6, no. 2, pp. 101–108, 2012.
[52] 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.
[53] B. Schuller, G. Rigoll, and M. Lang, ''Speech emotion recognitioncombining 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.
[54] 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.
[55] D.VerveridisandC.Kotropoulos,''Astateoftheartreviewonemotional speech databases,'' in Proc. 1st Richmedia Conf., 2003, pp.109–119.
[56] F.Burkhardt,A.Paeschke,M.Rolfes,W.F.Sendlmeier,andB.Weiss, ''AdatabaseofGermanemotionalspeech,''inProc.9thEur.Conf.Speech Commun. Technol., 2005, pp. 1–4.
[57] J.A.CoanandJ.J.Allen,HandbookofEmotionElicitationandAssessment. London, U.K.: Oxford Univ. Press, 2007.
[58] 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.
[59] M.Valstar,J.Gratch,B.Schuller,F.Ringeval,D.Lalanne,M.Torres,S.Scherer,G.Stratou,R.Cowie,andM.Pantic,''AVEC2016: 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.
[60] P. Jackson and S. Haq, Surrey Audio-Visual Expressed Emotion(SAVEE) Database. Guildford, U.K.: Univ. Surrey, 2014.
[61] 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.

[62] K. Poon-Feng, D.-Y. Huang, M. Dong, and H. Li, ''Acousticemotion 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.

[63] V. Nair and G. E. Hinton, ''Rectified linear units improverestrictedBoltzmann machines,'' in Proc. 27th Int. Conf. Mach. Learn. (ICML), 2010, pp. 807–814.

[64] G. E. Hinton, S. Osindero, and Y.-W. Teh, ''A fast learning algorithmfor deep belief nets,'' Neural Comput., vol. 18, no. 7, pp. 1527–1554, 2006.

[65] A.-R. Mohamed, G. Dahl, and G. Hinton, ''Deep belief networks for phonerecognition,''inProc.NIPSWorkshopDeepLearn.SpeechRecog- nit. Rel. Appl., Vancouver, BC, Canada, 2009, vol. 1, no. 9, p.39.

[66] J.LeeandI.Tashev,''High-levelfeaturerepresentationusingrecurrentneural network for speech emotion recognition,'' in Proc. 16th Annu. Conf. Int. Speech Commun. Assoc., 2015, pp. 1537–1540.

[67] V.ChernykhandP.Prihodko,''Emotionrecognitionfromspeechwith recurrent neural networks,'' 2017, arXiv:1701.08071. [Online]. Avail- able: https://arxiv.org/abs/1701.08071.

[68] Y. Kamp and M. Hasler, Recursive Neural Networks forAssociativeMemory. 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 usingdeepbeliefnetworks,''IEEETrans.Audio,Speech,LanguageProcess.,vol.20, no. 1, pp. 14–22, Jan.2012.

[72] C. Huang, W. Gong, W. Fu, and D. Feng, ''A research of speechemotion 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,andH.Larochelle,''Greedylayer 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 ofspeech 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 sentenceclassification,'' 2014, arXiv:1408.5882. [Online]. Available: https://arxiv.org/ abs/1408.5882

[76] W.Fei,X.Ye,Z.Sun,Y.Huang,X.Zhang,andS.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, ''Introducingshared 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,andB.Schuller,''Sparseautoencoder based feature transfer learning for speech emotion recognition,'' in Proc. IEEE Humaine Assoc. Conf. Affect. Comput. Intell. Interact., Sep. 2013, pp. 511–516.

[79] Mekruksavanich, S.; Jitpattanakul, A.; Hnoohom, N. Negative Emotion Recognition using Deep Learning for Thai Language. In Proceedings of the Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT and NCON), Pattaya, Thailand, 11–14 March 2020; pp. 71–74.

[80] Sepp Hochreiter, J.S. Long Short-Term Memory. Neural Comput. 1997, 9, 1735–1780.

[81] Kim, J.; Englebienne, G.; Truong, K.P.; Evers, V. Towards Speech Emotion Recognition "in the wild" using Aggregated Corpora and Deep Multi-Task Learning. arXiv 2017, arXiv:1708.03920.

[82] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative

Adversarial Networks. arXiv 2014, arXiv:1406.2661.

[83] Latif, S.; Rana, R.; Qadir, J. Adversarial Machine Learning Additionally, Speech Emotion Recognition: Utilizing Generative Adversarial Networks For Robustness. arXiv 2018, arXiv:1811.11402.

[84] Mirza, M.; Osindero, S. Conditional Generative Adversarial Nets. arXiv 2014, arXiv:1411.1784.

[85] Sahu, S.; Gupta, R.; Espy-Wilson, C. On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks. arXiv 2018, arXiv:1806.06626.

[86] Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Adversarial Networks. arXiv 2014, arXiv:1406.2661.

[87] Latif, S.; Rana, R.; Qadir, J. Adversarial Machine Learning Additionally, Speech Emotion Recognition: Utilizing Generative Adversarial Networks For Robustness. arXiv 2018, arXiv:1811.11402.

[88] Song, P.; Jin, Y.; Zhao, L.; Xin, M. Speech Emotion Recognition Using Transfer Learning. IEICE Trans. Inf. Syst. 2014, 97, 2530–2532Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[89] Sahu, S.; Gupta, R.; Espy-Wilson, C. On Enhancing Speech Emotion Recognition Using Generative Adversarial Networks. arXiv 2018, arXiv:1806.06626.

[90] Chatziagapi, A.; Paraskevopoulos, G.; Sgouropoulos, D.; Pantazopoulos, G.; Nikandrou, M.; Giannakopoulos, T.; Katsamanis, A.; Potamianos, A.; Narayanan, S. Data Augmentation Using GANs for Speech Emotion Recognition. In Proceedings of the INTERSPEECH 2019: Speech Signal Characterization 1, Graz, Austria, 15–19 September 2019

[91] Huang, C.W.; Narayanan, S.S. Deep convolutional recurrent neural network with attention mechanism for robust speech emotion recognition in Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017.

[92] Song, P.; Jin, Y.; Zhao, L.; Xin, M. Speech Emotion Recognition Using Transfer Learning. IEICE Trans. Inf. Syst. 2014, 97, 2530–2532.

[93] Hsiao, P.W.; Chen, C.P. Effective Attention Mechanism in Dynamic Models for Speech Emotion Recognition in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018.

[94] Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.

[95] Li, Y.; Zhao, T.; Kawahara, T. Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning inProceedings of the INTERSPEECH 2019: Training Strategy for Speech Emotion Recognition, Graz, Austria, 15–19 September 2019.

[96] X.Zhou,J.Guo,andR.Bie,''Deeplearningbase daffectivemodelfor 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.

[97] Song, P. Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition. IEEE Trans. Affect. Comput. 2019, 10, 265–275.

[98] K.Han,D.Yu,andI.Tashev,''Speechemotionre cognitionusingdeep neural network and extreme learning machine,'' in Proc. 5th Annu. Conf. Int. Speech Commun. Assoc., 2014, pp. 223–227.

[99] K.-C.HuangandY.-H.Kuo,''Anovelobjectivefunctiontooptimize neural networks for emotion recognition from speech patterns,'' in Proc. IEEE 2nd World Congr. Nature Biolog. Inspired Comput. (NaBIC), Dec. 2010, pp. 413–417.

[100] 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.

[101] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, ''Featurelearning in deep neural networks—Studies on speech recognition

tasks,'' 2013, arXiv:1301.3605. [Online]. Available: https://arxiv.org/abs/1301.3605

[102] 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.

[103] G.E.Dahl,D.Yu, L.Deng,andA.Acero,''Context-dependentpretrained deep neural networks for large-vocabulary speech recognition,'' IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 1, pp. 30–42, Jan. 2012.

[104] A.Satt,S.Rozenberg,andR.Hoory,''Efficiente motionrecognitionfrom speech using deep learning on spectrograms,'' in Proc. INTERSPEECH, 2017, pp. 1089–1093.

[105] F.Seide,G.Li,X.Chen,andD.Yu,''Featureengineeringincontext dependentdeepneuralnetworksforconversationalspeechtranscription,'' in Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU), Dec. 2011, pp. 24–29

[106] 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.

[107] S. Tripathi and H. Beigi, ''Multi-modal emotion recognitiononIEMOCAP dataset using deep learning,'' 2018, arXiv:1804.05788. [Online]. Available: https://arxiv.org/abs/1804.05788

[108] P.Yenigalla,A.Kumar,S.Tripathi,C.Singh,S. Kar,andJ.Vepa,''Speechemotion recognitionusing spectrogram & phoneme embedding,'' in Proc. Interspeech, 2018, pp. 3688–3692.

[109] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S.Wermter,''On the robustness of speech emotion recognition for human-robot inter- action with deep neural networks,'' in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. 2018, pp. 854–860.

[110] D.Tang,J.Zeng,andM.Li,''Anend-to-enddeeplearningframeworkwith speech emotion recognitionof atypical individuals,'' in Proc. Inter- speech, Sep. 2018, pp. 162–166.

[111] 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

[112] 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

[113] M.Chen,X.He,J.Yang,andH.Zhang,''3-Dconvolutionalrecurrent neural networks with attention model for speech emotion recognition,''IEEE Signal Process. Lett., vol. 25, no. 10, pp. 1440–1444, Oct. 2018.

[114] 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.

[115] S.E.Eskimez,Z.Duan,andW.Heinzelman,''Unsupervisedlearning approach to feature analysis for automatic speech emotion recognition,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Apr. 2018, pp.5099–5103.

[116] Xie, Y.; Liang, R.; Liang, Z.; Huang, C.; Zou, C.; Schüller, B. Speech Emotion Classification Using Attention-Based LSTM. IEEE/ACM Trans. Audio Speech Lang. Process. 2019, 27, 1675–1685.

[117] M. R. Sambur, A. E. Rosenberg, L. R. Rabiner and C. A. McGonegal, "On reducing the buzz in LPC synthesis", J. Acoust. Soc. Amer., vol. 63, no. 3, pp. 918-924, 1978.

[118] Kumar R, Ranjan R, Singh SK, Kala R, Shukla A, Tiwari R. Multilingual speaker recognition using neural network. in Proceedings of the Frontiers of Research on Speech and Music, FRSM. 2009. pp. 1-8

[119] Holambe R, Deshpande M. Advances in Non-Linear Modeling for Speech Processing. Berlin, Heidelberg: Springer Science & Business Media; 2012

[120] Hermansky H. Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America. 1990;87(4):1738-1752.

[121] Chakroborty S, Roy A, Saha G. Fusion of a complementary feature set with MFCC for improved closed set text-independent speaker identification. In: IEEE International Conference on Industrial Technology, 2006. ICIT 2006. pp. 387-390.

[122] Greg Hopper, Reza Adhami, "An FFT-based speech recognition system", Journal of Franklin Institute, vol.329, no.3, pp.555-565, May 1992.

[123] Hermansky H., Morgan N., Bayya A. & Kohn P.: RASTA-PLP Speech Analysis. Technical Report (TR-91-069), International Computer Science Institute, Berkeley, CA., 1991.
[124] Shrawankar, Urmila & Thakare, V. M.. (2013). Techniques for Feature Extraction In Speech Recognition System : A Comparative Study.
[125] "Generative Adversarial Network." developers.google.com. https://developers.google.com/machine-learning/gan/gan_structure, 24 May. 2019. Web. 25 Mar. 2021.
[126] Brownlee, Jason . "A Gentle Introduction to Transfer Learning for DeepLearning."emarsys.com.https://machinelearningmastery.com/transfer-learning-for-deep-learning/, 20 Dec. 2017. Web. 25 Mar. 2021.