**SPEECH RECOGNIZATION USING LSTM CLASSIFICATION**

**18CSC206J/ SOFTWARE ENGINEERING AND PROJECT MANAGEMENT REPORT**

***Submitted by***

**JAANA SHRUTI B RA2111026020027**

**RITHIKA PJ RA2111026020041**

**Under the guidance of**

**DR. R. REGIN**

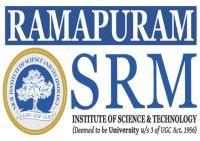
**(Assistant Professor, Department of Computer Science and Engineering)**

***in partial fulfilment for the award of the degree of***

**BACHELOR OF TECHNOLOGY**

***in***

**COMPUTER SCIENCE AND ENGINEERING WITH SPECIALIZATION IN ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING**

**FACULTY OF ENGINEERING AND TECHNOLOGY**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY RAMAPURAM CAMPUS, CHENNAI -600089**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY**

**(Deemed to be University U/S 3 of UGC Act, 1956)**

**BONAFIDE CERTIFICATE**

Certified that this project report titled **“SPEECH RECOGNIZATION USING LSTM CLASSIFICATION”** is the bonafide work of **JAANA SHRUTI B [RA2111026020027] and RITHIKA PJ [RA2111026020041]** who carried out the project work under my supervision. Certified further, that to the best of my knowledge, the work reported herein does not form any other project report or dissertation on the basis of which a degree or award was conferred on an occasion on this or any other candidate.

|  |  |
| --- | --- |
| SIGNATURE  **Dr. R. REGIN, M.E., Ph.D.**  **Assistant Professor**  Computer Science and Engineering,  SRM Institute of Science and Technology,  Ramapuram Campus, Chennai. | SIGNATURE  **Dr. Balika J Chelliah, M.E., Ph.D.**  **Head of Department**  Computer Science and Engineering,  SRM Institute of Science and Technology,  Ramapuram Campus, Chennai. |

Submitted for the project viva-voce held on at SRM Institute of Science and Technology, Ramapuram Campus, Chennai -600089.

**INTERNAL EXAMINER EXTERNAL EXAMINER**

**SRM INSTITUTE OF SCIENCE AND TECHNOLOGY RAMAPURAM, CHENNAI - 89**

**DECLARATION**

We hereby declare that the entire work contained in this project report titled “**SPEECH RECOGNIZATION USING LSTM CLASSIFICATION**” has been carried out by **JAANA SHRUTI B [RA2111026020027] and RITHIKA PJ [RA2111026020041]** at SRM Institute of Science and Technology, Ramapuram Campus, Chennai- 600089, under the guidance of **DR. R. REGIN, Assistant Professor**, Department of Computer Science and Engineering.

**Place: Chennai**

**Date: JAANA SHRUTI B**

**RITHIKA PJ**

**SPEECH RECOGNIZATION USING LSTM CLASSIFICATION ABSTRACT**

This project proposes the advancements in neural networks and the on-demand need for accurate and near real-time Speech Emotion Recognition (SER) .In human–computer interactions make it mandatory to compare available methods and databases in SER to achieve feasible solutions and a firmer understanding of this open-ended problem. The proposed method consists of feature extraction, feature selection and classification stages. Thereafter, a hybrid feature vector is created by combining acoustic and deep features. These datasets are Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Berlin (EMO-DB) and Interactive Emotional Dyadic Motion Capture (IEMOCAP). The Final results obtained in experiments, clearly, show that the proposed technique might be utilized to accomplish the task of speech emotion recognition efficiently. Moreover, when our technique is compared with those of methods used in the context, it is obvious that our method outperforms others in terms of classification accuracy rates. This project is implemented using Python Software.

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **CONTENTS** |  | **Page No.** |
|  |  |  |
| **ABSTRACT** |  | **4** |
|  |  |  |
| **LIST OF FIGURES**  **LIST OF ACRONYMS AND ABBREVIATIONS**  **LIST OF EQUATION**   1. **INTRODUCTION**    1. Introduction    2. Semi-Supervised learning    3. Axillary Tasks and Multitask Learning    4. Ladder Networks    5. Sentence Level and Frame Level Features    6. Relation to Prior Work    7. Advantages    8. Applications    9. Objective 2. **LITERATURE REVIEW**    1. Introduction 3. **EXISTING SYSTEM**    1. Introduction    2. Speech Emotion Recognition System    3. DSR (Distributed Speech Recognition Model)    4. Feature Extraction       1. Pitch       2. Energy       3. Pitch Difference and Energy Difference       4. Formants       5. Mel-Frequency Cepstrum Coefficients    5. Classification Algorithms       1. Sequential Minimal Optimization Algorithm       2. Support vector machine classifier       3. Decision Tree Algorithm       4. K Nearest Neighbor    6. Disadvantages of Existing System 4. **PROPOSED SYSTEM**    1. Introduction    2. Proposed system    3. Data Augmentation    4. Feature Extraction       1. RMS       2. Spectrum Centroid       3. Spectral Entropy       4. Skewness       5. MFCC       6. Attack time       7. Chroma       8. Zero-Crossing rate    5. Spectrogram Extraction    6. Convolutional Neural Networks    7. Deep convolutional neural networks    8. LSTM Networks 5. **RESULTS AND DISCUSSION**   5.1 Experimental Applications   1. **SOURCE CODE** 2. **CONCLUSION**   **References** |  | **9**  **11**  **13**  **14**  14  20  20  21  21  22  22  22  22  **24**  24  **37**  37  38  39  40  40  41  41  41  41  41  42  42  43  43  43  **45**  45  47  47  49  49  49  49  50  50  51  51  51  51  53  53  55  **60**  60  **67**  **80**  **81** |

**LIST OF FIGURES**

**Figure: 1.1** Relationship between emotion categories and sound types for the data in NNIME and BAUM-1

**Figure: 1.2** Framework of the proposed speech emotion recognition system

**Figure: 3.1** Algorithm

**Figure: 3.2** Speech Recognition model

**Figure: 3.3** Distributed Speech Recognition

**Figure: 4.1** Proposed Block Diagram

**Figure: 4.2** Steps of MFCC

**Figure: 4.3** Illustration of Speech Sound Signal and Spectrogram

(a) Speech Sound Signal,

(b) Spectrogram

**Figure: 5.1** Data Set

**Figure: 5.2** Waveplot And Spectrogram for Audio with Fear

**Figure: 5.3** Waveplot And Spectrogram for Audio with Angry Emotion

**Figure: 5.4** Waveplot And Spectrogram for Audio with Sad Emotion

**Figure: 5.5** Waveplot and Spectrogram for Audio with Happy Emotion

**Figure: 5.6** Simple Audio

**Figure: 5.7** Noise Injection

**Figure: 5.8** Stretching

**Figure: 5.9** Shifting

**Figure: 5.10** Pitch

**Figure: 5.11** Feature Extraction

**Figure: 5.12** Build Model

**Figure: 5.13** Model Loss And Accuracy

# 

# LIST OF ACRONYMS AND ABBREVIATIONS

# LSTM Long short-term memory

# SER Speech Emotion Recognition

# RAVDESS Ryerson Audio-Visual Database of Emotional Speech and Song

# IEMOCAP Interactive Emotional Dyadic Motion Capture

# NN Neural network

# LLD Low-level descriptors

# CNN Convolutional neural network

# ResNet Residual networks

# HMM Hidden Markov models

# GMM Gaussian mixture models

# SVM Support vector machine

# GAN Generative adversarial networks

# FBP Factorized bilinear pooling

# FCN Fully convolutional network

# MCIL Multi-classifier interactive learning

# MFCC Mel-frequency cepstral coefficient

# MLP Multi-layer perceptron

# KNN k-nearest neighbour

# SER Speech emotion recognition

# TRaSL Transfer subspace learning

# CMRN Collective Multi-view Relation Network

# R-CNN Residual Convolutional Neural Network

# ISLA Informed Segmentation and Labelling Approach

# MTL Multitask learning

# DNNs Deep neural networks

# GM Global-level emotion presenting module

# UM Utterance-level emotion presenting module

# LM Local-level emotion presenting module

**ADAN** Adversarial data augmentation network

**GANs** Generative adversarial networks

**LPC** Linear Prediction Coefficient

**SMO** Sequential Minimal Optimization

**DSR** DISTRIBUTED SPEECH RECOGNITION

**MFCC** MEL-FREQUENCY CEPSTRUM COEFFICIENTS

**SMO** SEQUENTIAL MINIMAL OPTIMIZATION

**AWS** Adjustable weighted segmentation

**RMS** Root Mean Square

**KNN**  K NEAREST NEIGHBOR

**HMM** Hidden Markov models

**GMM**  Gaussian mixture models

**ASR** Automatic speech recognition

**FFT** Fast Fourier Transform

**SIANNs** Shift invariant artificial neural networks

**DTPM** Discriminant temporal pyramid matching

**LFLB** Local feature learning block

**LMS**  Log Mel Spectrogram

**BN** Batch normalization

**ELU** Exponentiallinear unit

# LIST OF EQUATIONS

(3.1)

(3.2)

(4.1)

(4.2)

(4.3)

(4.4)

(4.5)

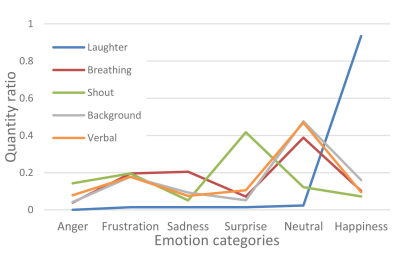
(4.6)

**CHAPTER-1**

**INTRODUCTION**

**1.1 INTRODUCTION**

In Recent Years, rapid progress of technology makes smart devices more attractive in our daily life. Intelligent services such as chatbots, psychological diagnosis assistants, intelligent healthcare, sales advertising and intelligent entertainment consider not only the completion of services, but also the humanization of communication between human and computer. How to implement an intelligent human-machine interface becomes an important issue. For the applications of the spoken dialog systems, the leading organizations use chatbots.



**Figure: 1.1** Relationship between emotion categories and sound types for the data in NNIME and BAUM-1

To improve their customer service and generate good business results for the organization. In contrast to customer engagement, empathy, which is highly related to emotion, has been incorporated into the design of a dialogue system for improving user experience in human-computer interaction. More importantly, being empathetic is a necessary step for the dialogue system to be perceived as a social character by users. Based on the above motivation to achieve empathetic human-machine interaction.

This study aims at improving the accuracy of emotion recognition. With the population growth of using mobile devices, speech is one of the most common ways for human-human and human machine communication. As people can realize others’ emotion by information involved in speech, speech emotion recognition is one of the most important technique for realizing computer affective intelligence. In general, research on speech emotion recognition mainly focuses on emotion feature extraction and recognition modelling.

Furthermore, the choice of emotional speech database is also important for developing a robust speech emotion recognizer. Recent emotion recognition systems, limited by the lack of speech data resources, seldom focused on nonverbal vocalizations in spontaneous communications. The relationship between sound types and emotion categories for the data in the NNIME and BAUM-1 corpora is estimated and shown in Fig: 1.1. As depicted in this figure, certain sound types are related to some specific emotion categories. For example, laughter is easily recognized as happiness, and shout is highly associated with surprise.

The existing literature on nonverbal emotion recognition described that certain nonverbal sounds (such as laughter and shout) are beneficial to the recognition of positive emotions (such as happiness and surprise). However, there are not many relevant experiments and discussions on other nonverbal sound types and the influence of obscure or negative emotions (such as sadness and frustration). In real life communications, nonverbal sounds within an utterance play an important role for people to recognize others’ emotion. When human brain analyses emotional speech, nonverbal sounds can effectively help the brain to gain the difference of emotion expression. Therefore, this study aims to apply the features of nonverbal sounds to speech emotion recognition. With more and more studies on affective computing, scholars have established several affective databases.

These databases contain the data in different forms (such as audio, video, text and motion) and different presentations (such as language, emotion label, and script). To analyse nonverbal sounds in emotional speech, a suitable emotion speech database is desirable for research. Different from other scripted emotional speech databases or segmented audio in dramas, the NTHU-NTUA Chinese Interactive Multimodal Emotion Corpus (NNIME), which is a spontaneous Chinese speech emotional database containing various emotional nonverbal sounds, such as laughter, sobs and voice sighs, meets the requirement of this study and is thus adopted.

In the mechanism of speech emotion recognition, how to represent emotion features extracted from speech signals is a key issue for effective emotion recognition. Many studies looked for appropriate audio features or suitable audio feature sets for emotion feature representation. Other studies tried to represent audio with different feature units. In recent years, with the progress of neural network (NN), many researchers tried to extract features from raw audio waveform or spectrum for emotion recognition. Some studies concluded that the existing acoustic feature set may lack subjective emotional features, because the parameters of the feature extraction algorithm were mostly manually adjusted.

They expected that the NN-based end-to-end emotion feature extraction methods can get information from speech. Furthermore, many studies have discussed which features are suitable for emotion classification. Types of emotions, such as sadness and disgust, can be distinguished by alternately using the intensity or pitch value of the first or last word. Tzinis and Potamianos concluded that the performance of feature selection was closely related to the duration of the recognition unit. For example, low-level descriptors (LLDs) has better performance in terms of phonemes. Compared to the result of sentiment classification, a word was better than a sentence to be selected as the speech unit.

The experimental results in showed that a relatively short snippet of a speech signal can be sufficient to perform emotion recognition with only a small accuracy loss. This is an important finding for the development of real-time applications which aim to make prediction while the user is still speaking. Based on different types of emotion features, the selection of speech emotion recognition models should consider the attributes of the extracted features. Long-short term memory (LSTM) is the one which can classify emotion types of features of continuous time attributes.

Proposed the attention-based long short-term memory (LSTM) recurrent neural network for speech emotion recognition using frame level speech features. Their experiments on the CASIA, eNTERFACE, and GEMEP emotion corpora demonstrated that the performance of the proposed approach outperformed the state-of-the-art algorithms reported to date. Convolutional neural network (CNN) can combine local features by the convolution step to get classification results of a larger range of combined information.

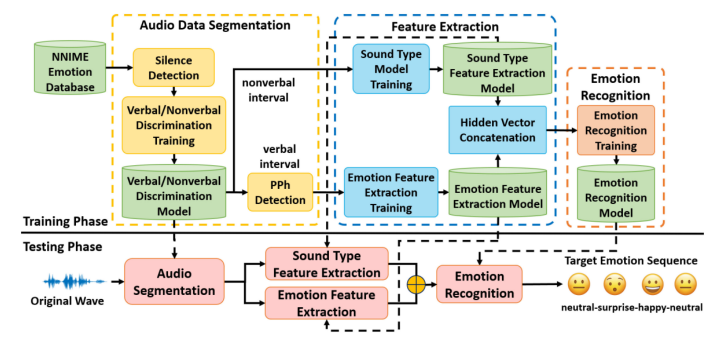
Used 1-D and 2-D CNNs with LSTM network for speech emotion recognition.1-D CNN was built to learn the local and global emotion-related features from speech, whereas the 2-D CNN was used to learn Mel spectrogram. The combination of CNN and LSTM, which was called CLDNN, expected to have benefits from two types of models and had better performance by involving attention. As the deep network architecture gradually matures, the use of deeper CNN-based models also has good performance for image recognition.

Using ResNet to extract features of the magnitude spectrum of the speech has also obtained many excellent results in speech emotion recognition. Based on the above descriptions, this study addresses the following issues to be resolved. The first issue is the selection of basic units for speech emotion recognition. When we recognize other people’s emotions in a conversation, we do not need to listen to the entire response of others for emotion recognition.

From the reports in the literature, sentence-level features were worse than speech segments that roughly correspond to the duration of a couple of words for emotion recognition. Previous study showed that recognition of emotions requires at least more than one second of speech signal. Studies in showed that discrete prosodic action units could represent verbal emotions. As pauses played the role of prosodic boundary markers in Mandarin speech, we chose prosodic phrases as the basic unit for emotion recognition.

For the second issue, there are many nonverbal sounds in spontaneous speech, such as laughter and cries, which can help distinguish the emotions of others. In the circumstance that we hear only snort sounds, we cannot determine others’ emotion. If we hear snort sounds involving in a dialogue turn with low energy and many pauses, we can identify that it is a sobbing sound in a sad emotion within a dialogue turn. The research in demonstrated that when people recognize the emotion of others, our brain processes the speech representation and emotion independently, and then obtains the difference of emotion more clearly through the voice effect.

As both verbal utterance and nonverbal vocalization could complement each other for speech emotion expression, this study uses two different deep residual networks (ResNet) to extract the emotional features of the speech utterance and the features of the nonverbal sounds for speech emotion recognition. For the last issue, we think that there may exist multiple emotional expressions in a dialogue turn. People express different degrees or different types of emotions in different dialogue turns. People show high arousal emotions in the content of interest, and they consider neutral emotions when describing normal daily life events.



**Figure: 1.2** Framework of the proposed speech emotion recognition system

For this issue, we divide the audio signal into verbal and nonverbal segments for feature extraction. The long short-term memory (LSTM) model is then used to characterize the sequential emotion expression of speech signals. That is, the LSTM model is constructed to obtain the speaker’s emotion changes in a dialogue from a series of segmented speech signals. Fig: 1.2 shows the proposed system framework. In the training phase, there are three main stages including verbal/nonverbal sound segmentation, feature embedding extraction and emotion model construction.

First, speech signals go through silence detection, verbal/nonverbal segment detection and prosodic-phrase segmentation procedures to obtain sound/speech segments. Secondly, sound/speech segments of verbal and nonverbal sounds are used to train the corresponding ResNet models for extracting the generic feature embeddings of emotion and sound. ResNet models which are trained for emotion/sound type classification are used as feature extractors by removing the output layer. Finally, the ResNet features of sound and emotion are used as the representative feature embedding of each segment. The feature embedding sequence is used for emotion recognition to obtain the emotion recognition results of each segment considering emotional changes in a temporal context.

**1.2 SEMI-SUPERVISED LEARNING**

Previous studies for semi-supervised learning have considered the inductive learning technique, where a classifier is first trained on the labelled samples. The trained classifier is then used on the unlabelled set to obtain predictions. The training set is then augmented with samples having highly confident predictions. The classifier is retrained with this augmented training set. This process is iterated a fixed number of times after which the performance often saturates. This inductive learning procedure for SER to leverage unlabelled data. They enhanced their supervised learning approach with this method, obtaining better predictions on labelled data. A similar strategy for facial expressions using probabilistic Bayesian classifiers.

**1.3 AUXILIARY TASKS AND MULTITASK LEARNING**

There are multiple studies that have analysed the regularizing benefits of auxiliary tasks for SER. The learning of emotional categories and emotional attributes. The primary task was the classification of emotional categories. The secondary task was either classification or regression of emotional attributes. To jointly predict arousal, valence and dominance scores using a MTL framework, where recognizing one of the attributes was the primary task and recognizing the other two attributes were the secondary tasks. The MTL framework learned the inherent correlation between the various emotional attributes leading to improvements over STL.

Similarly, an auxiliary task for a valence classifier. Similar improvements in performance for the prediction of time-continuous emotional attributes. Their system jointly predicted arousal and valence scores, obtaining the best performance for the affect sub-task in the audio/visual emotion challenge (AVEC) in 2017. A MTL framework for time-continuous attribute recognition. Their framework trained classifiers by discretizing attribute scores into discrete classes using the k-means algorithm with k ∈ {4, 6, 8, and 10}. The different classifiers were then learned together as multiple auxiliary tasks using MTL framework. (e.g., learning together a four-class problem and a six-class problem). for categorical emotion recognition by using a MTL framework for learning the dominant emotion (primary task) and secondary emotions also conveyed in the sentence (auxiliary task).

**1.4 LADDER NETWORKS**

This work showed the benefits of using lateral shortcut connections to aid deep unsupervised learning. Further extended this idea to support supervised learning. Classification and regression tasks were added to the unsupervised reconstruction of inputs through a denoising auto encoder. The various components that affected the ladder network, noting that lateral connections between encoder and decoder and the addition of noise at every layer of the network greatly contributed to the improved performance of this framework.

**1.5 SENTENCE-LEVEL AND FRAME-LEVEL FEATURES**

Conventionally, SER problems are formulated using sentence-level features over short speech segments. Previous studies often rely on statistics estimated over LLDs, where popular examples include the feature sets proposed for the paralinguistic challenges at Interspeech. An alternative approach is to directly use a sequence of features extracted at the low-level over short segments (e.g., 40 ms). We refer to these features as low-level features or frame-level features. Borrowed successful CNN architectures from the computer vision domain by treating speech spectrograms as images.

Performed SER in a two-step approach using a CNN architecture on low-level features. The first step learned features from unlabelled data and a sparse auto encoder. These features were then used for the recognition task. A CNN architecture to perform end-to-end SER that took raw speech waveforms as inputs. An attention based convolutional neural network for emotion recognition. arousal and valence using CNNs trained on spectrogram inputs. Likewise, to train 1-D CNNs on mel-filter bank energies to capture regional saliency for emotion recognition.

**1.6 RELATION TO PRIOR WORK**

This study presents important contributions with respect to previous studies, including our previous work. The use of ladder network for SER is appealing since the auxiliary task is unsupervised, so we can use data from the target domain without labels. This feature of the proposed approach is a key distinction between our work and most MTL studies, which use supervised auxiliary tasks. (1) Implements the ladder networks in a semi-supervised fashion instead of a supervised fashion, (2) demonstrates that the proposed architecture can be implemented with different features both at the sentence level, and the frame-level, and (3) evaluates the proposed architecture with extensive within-corpus and cross corpus evaluations.

**1.7 ADVANTAGES**

* Easy to implementation and more accurate.
* No training required for users.
* Providing users with subscription clue service with intelligent classification.

**1.8 APPLICATIONS**

* Mobile applications.
* Script singing system.
* Computer-aided language learning application.
* Business function applications in customer service.
* Industry applications.

**1.9 OBJECTIVE**

* To recognize emotions from speech, a method that is based on MFCC features and Gaussian mixture model classifier
* To classify seven different emotions in the dataset, 3 staged SVM classifier is used.
* To find how feature extraction through speech samples and then classification of the extracted features is performed under different emotions.

**CHAPTER-2**

**LITERATURE REVIEW**

**2.1 INTRODUCTION**

In Recent years, deep learning has made remarkable progress in many areas, such as speech recognition, image recognition, and genomics. Generally, deep learning models are complex and require a large amount of data to achieve accurate predictions or classifications. Unfortunately, the data collection process is often expensive and time-consuming, which makes acquiring labelled data a big challenge. This problem is particularly acute in speech emotion recognition because an utterance may contain ambiguous or multiple emotions. Multiple annotators are often employed to label the utterances in speech emotion corpora to increase the annotation reliability. Nevertheless, in some cases, even professional annotators may not be unanimous in their decisions. Therefore, it is important to address the data sparsity problem. Emotion recognition plays a key role in natural human–computer interaction. Traditional speech emotion recognition systems consist of a feature extractor in the front end and a classifier at the back end. For the latter, the hidden Markov models (HMMs) and the Gaussian mixture models (GMMs) have been used to classify the instantaneous and global features extracted from the front end. Another approach is to use prosodic features to train a support vector machine (SVM) for classification. However, these handcrafted features may not be optimal for emotion recognition. Handcrafted features are not robust in that their performance is highly dependent on the evaluation set. With the development of deep learning, using deep neural networks (DNNs) to extract features gradually replaces manual feature engineering. In particular, the convolutional neural networks (CNNs) and the long short-term memory recurrent neural networks (LSTM-RNN) have been used to exploit the dynamic structure of frame-based features. Motivated by the success of CNNs in image classification, more and more researchers applied CNNs to extract features from spectrograms. For instance, a CNN-based feature learning method to extract emotion-salient features that are invariant to nuisance factors. CNNs with recurrent neural networks (RNNs) to extract language information from spectrograms. An HSF-CRNN system that combines the handcrafted high-level statistic functional (HSF) features and the features learned by a convolutional RNN (CRNN). A similar strategy was trained to extract emotion features from handcrafted features and a CNN was trained to extract emotion features from spectrograms. The combined features were classified by an extreme learning machine. Recent work tends to apply an end-to-end scheme to tackle the speech emotion recognition tasks. Typically, spectrograms and class labels are, respectively, used as the input and output of the end-to-end systems. In feature maps produced by convolutional filters are divided into time-specific and frequency-specific. To assign higher weights to emotion-related parts in the spectrogram, top–down attention and bottom–up attention were applied to the last convolutional layer. With the widespread applications of deep learning in emotion recognition, many effective solutions to the data sparsity problem have been investigated. Transfer learning is a popular solution to the insufficient-data problem. In particular, domain adaptation (a subset of transfer learning) can leverage labelled data from the source domain to learn a model for the unlabelled data in the target domain. Motivated by the achievements in image classification, this technique has been gradually applied to speech emotion recognition as well. For example, a feature transfer learning method in which source-domain data are transformed to the target-domain through a sparse auto encoder trained from the target-domain training data. Transformed data from the source domain are then used for training an emotion classifier. The authors found that the classifier trained by the transformed data can significantly improve the performance of emotion classification on the target data set, even if the target data set is small. In addition to transferring knowledge within the same task (emotion recognition), another study demonstrates that transferring from speaker recognition to emotion recognition could also help improve emotion classification. The introduction of generative adversarial networks (GANs) creates new possibilities for tackling the insufficient data problem. A typical GAN consists of a generator and a discriminator. Both are neural networks that act like two players competing with each other in a zero-sum game. The generator learns to map an arbitrary distribution to the data distribution to confuse the discriminator, and the discriminator is trained to distinguish whether a sample comes from the data distribution (i.e., genuine) or from the generator (i.e., fake). Some studies show that GAN-based data augmentation techniques can help improve the performance of image recognition. An improved GAN to generate high-dimensional representations and showed that GAN-based data augmentation outperforms conventional data augmentation techniques. We proposed an adversarial data augmentation network (ADAN) that combines an auto encoder with a GAN to perform data augmentation. The ADAN not only overcomes the gradient vanishing problem that often occurs in vanilla GANs but also produces real-like samples that share common latent representation with the real data. In this article, we will further explain the effectiveness of ADAN and extend by replacing the adversarial loss with Wasserstein divergence. Unlike we used the whole IEMOCAP data set and used a newer feature set in this article. Moreover, in-depth analyses have been added. These analyses include comparisons of different types of GANs, distances between the distributions of real and synthetic data, the effect of varying the number of augmented samples, and the efficiency of the proposed models. Our experimental results demonstrate that the proposed data augmentation approach can further improve the recognition performance on the EmoDB and IEMOCAP data sets. The rest of this article is organized as follows. Section II presents a review of common solutions to the data sparsity problem. Section III describes the design of the proposed network and provides a theoretical analysis. Section IV introduces the details of the experiments, including descriptions of data, features, experimental setup, and evaluation protocol. Section V presents and analyses the experimental results.

**Hengshun Zhou et al [2021]** explained the Multimodal emotion recognition is a challenging task in emotion computing as it is quite difficult to extract discriminative features to identify the subtle differences in human emotions with abstract concept and multiple expressions. Moreover, how to fully utilize both audio and visual information is still an open problem. We propose a novel multimodal fusion attention network for audio-visual emotion recognition based on adaptive and multi-level factorized bilinear pooling (FBP). First, for the audio stream, a fully convolutional network (FCN) equipped with 1-D attention mechanism and local response normalization is designed for speech emotion recognition. Next, a global FBP (G-FBP) approach is presented to perform audio-visual information fusion by integrating self-attention based video stream with the proposed audio stream. To improve G-FBP, an adaptive strategy (AG-FBP) to dynamically calculate the fusion weight of two modalities is devised based on the emotion-related representation vectors from the attention mechanism of respective modalities. Tested on the IEMOCAP corpus for speech emotion recognition with only audio stream, the new FCN method outperforms the state-of-the-art results with an accuracy of 71.40%. Moreover, validated on the AFEW database of EmotiW2019 sub-challenge and the IEMOCAP corpus for audio-visual emotion recognition, the proposed AM-FBP approach achieves the best accuracy of 63.09% and 75.49% respectively on the test set.

**Ying Zhou et al [2022]** proposed a speech emotion recognition technology is of great significance in widespread applications such as call centres, social robots and health care. Thus, the speech emotion recognition has been attracted much attention in both industry and academic. Since emotions existing in an entire utterance may have varied probabilities, speech emotion is likely to be ambiguous, which poses great challenges to recognition tasks. However, previous studies commonly assigned a single-label or multi-label to each utterance in certain. Therefore, their algorithms result in low accuracies because of the inappropriate representation. Inspired by the optimally interacting theory, we address the ambiguous speech emotions by proposing a novel multi-classifier interactive learning (MCIL) method. In MCIL, multiple different classifiers first mimic several individuals, who have inconsistent cognitions of ambiguous emotions, and construct new ambiguous labels (the emotion probability distribution). Then, they are retrained with the new labels to interact with their cognitions. This procedure enables each classifier to learn better representations of ambiguous data from others, and further improves the recognition ability.

**Reem Hamed Aljuhani et al [2021]** explained the machine learning algorithms has become an active research topic lately as a result of the demand for more human interactive applications. Emotion recognition systems are mostly implemented in German, English, Spanish, Dutch, Danish, and other European and Asian languages due to the availability of datasets for these languages. However, for Arabic, there is an extremely limited number of available speech emotion datasets. Therefore, in this paper studies emotion recognition based on Arabic Saudi dialect spoken data. The dataset was created from freely available YouTube videos and labelled using four perceived emotions: anger, happiness, sadness, and neutral. Various spectral features such as the mel-frequency cepstral coefficient (MFCC) and Mel spectrogram, were extracted, and then the classification methods support vector machine (SVM), multi-layer perceptron (MLP), and k-nearest neighbour (KNN) were applied. The results were discussed, analyzed, and compared between the three models using different feature extractions.

**Chenghao Zhang et al [2021]** proposed an important part of the human-computer interaction process is speech emotion recognition (SER), which has been receiving more attention in recent years. However, although a wide diversity of methods has been proposed in SER, these approaches still cannot improve the performance. A key issue in the low performance of the SER system is how to effectively extract emotion-oriented features. In this paper, we propose a novel algorithm, an auto encoder with emotion embedding, to extract deep emotion features. Unlike many previous works, instance normalization, which is a common technique in the style transfer field, is introduced into our model rather than batch normalization. Furthermore, the emotion embedding path in our method can lead the auto encoder to efficiently learn a priori knowledge from the label. It can enable the model to distinguish which features are most related to human emotion. We concatenate the latent representation learned by the auto encoder and acoustic features obtained by the open SMILE toolkit. To improve the generalization of our method, a simple data augmentation approach is applied. Two publicly available and highly popular databases, IEMOCAP and EMODB, are chosen to evaluate our method.

**Jia-Hao Hsu et al [2021]** explained the real-life communication, nonverbal vocalization such as laughter, cries or other emotion interjections, within an utterance play an important role for emotion expression. In previous studies, only few emotion recognition systems consider nonverbal vocalization, which naturally exists in our daily conversation. In this work, both verbal and nonverbal sounds within an utterance are considered for emotion recognition of real-life affective conversations. Firstly, a support vector machine (SVM)-based verbal and nonverbal sound detector is developed. A prosodic phrase auto-tagger is further employed to extract the verbal/nonverbal sound segments. For each segment, the emotion and sound feature embedding’s are respectively extracted using the deep residual networks (ResNets). The extracted feature embedding for the entire dialog turn are fed to an attentive long short-term memory (LSTM)-based sequence-to-sequence model to output an emotional sequence as recognition result. The NNIME corpus (The NTHU-NTUA Chinese interactive multimodal emotion corpus), which consists of verbal and nonverbal sounds, was adopted for system training and testing. 4766 single speaker dialogue turns in the audio data of the NNIME corpus were selected for evaluation.

**Na Liu et al [2021]** proposed a speech emotion recognition model learned on a source (training) domain but applied to a novel target (testing) domain degenerates even significantly due to the mismatch between the two domains. Aiming at learning a better speech emotion recognition model for the target domain, the paper investigates this interesting problem, i.e., unsupervised cross-corpus speech emotion recognition (SER), in which the training and testing speech signals come from two different speech emotion corpora. Meanwhile, the training speech signals are labelled, while the label information of the testing speech signals is entirely unknown. To deal with this problem, we propose a simple yet effective method called transfer subspace learning (TRaSL). TRaSL aims at learning a projection matrix with which we can transform the source and target speech signals from the original feature space to the label space. The transformed source and target speech signals in the label space would share similar feature distributions. Consequently, the classifier learned on the labelled source speech signals can effectively predict the emotional states of the unlabeled target speech signals. To evaluate the performance of the proposed TRaSL method, we carry out extensive cross-corpus SER experiments on four speech emotion corpora including IEMOCAP, EmoDB, eNTERFACE, and AFEW 4.0. Compared with recent state-of-the-art cross-corpus SER methods, the proposed TRaSL can achieve more satisfactory overall results.

**Mixiao Hou et al [2021]** proposed an Automatic emotion recognition from speech plays a fundamental role towards advanced emotional intelligence in human-machine interaction systems. The discriminative knowledge from speech for effective emotion recognition may come from multiple physical properties such as energy spectrum, frequency, prosody, which could be collected as multi-view representations. However, the current works fail to fully explore the underlying interactive relations among multiple speech representations for emotion recognition. We propose a novel Collective Multi-view Relation Network (CMRN) to exploit the intrinsic characteristics of multi-view speech representations for discriminative speech emotion recognition. Generally, the proposed CMRN consists of three sub-networks, i.e., view-specific attention network, multi-view shared attention network and collective relation network. Specifically, the view-specific attention network is designed to excavate the distinguishable view-specific features deduced from the original speech. By contrast, the multi-view shared attention network is conceived to capture the collaborative knowledge from multiple views. Moreover, a well-designed collective relation network is explicitly constructed to characterize the shared-specific correlations, which could reflect the underlying physical interaction capabilities. As such, the decision phase can comprehensively leverage the shared and view-specific information of multiple representations, such that the final privileged deciding principle can aggregate the heterogeneous information of multi-view features to make accurate emotion recognition. Extensive experiments on two benchmark datasets demonstrate the superb performance of the proposed method in comparison with some state-of-the-art methods.

**Ting-Wei Sun [2020]** explained the proper selection of speech acoustic features. We propose a novel emotion recognition algorithm that does not rely on any speech acoustic features and combines speaker gender information. We aim to benefit from the rich information from speech raw data, without any artificial intervention. In general, speech emotion recognition systems require manual selection of appropriate traditional acoustic features as classifier input for emotion recognition. Utilizing deep learning algorithms, and the network automatically select important information from raw speech signal for the classification layer to accomplish emotion recognition. It can prevent the omission of emotion information that cannot be direct mathematically modelled as a speech acoustic characteristic. We also add speaker gender information to the proposed algorithm to further improve recognition accuracy. The proposed algorithm combines a Residual Convolutional Neural Network (R-CNN) and a gender information block. The raw speech data is sent to these two blocks simultaneously. The R-CNN network obtains the necessary emotional information from the speech data and classifies the emotional category. **Yelin Kim et al [2017]** proposed an essential part of human interaction. Automatic emotion recognition can greatly benefit human-cantered interactive technology, since extracted emotion can be used to understand and respond to user needs. However, real-world emotion recognition faces a central challenge when a user is speaking: facial movements due to speech are often confused with facial movements related to emotion. Recent studies have found that the use of phonetic information can reduce speech-related variability in the lower face region. However, methods to differentiate upper face movements due to emotion and due to speech have been underexplored. This gap leads us to the proposal of the Informed Segmentation and Labelling Approach (ISLA). ISLA uses speech signals that alter the dynamics of the lower and upper face regions. We demonstrate how pitch can be used to improve estimates of emotion from the upper face, and how this estimate can be combined with emotion estimates from the lower face and speech in a multimodal classification system. We also demonstrate how emotion estimates from different modalities correlate with each other, providing insights into the differences between posed and spontaneous expressions.

**Srinivas Parthasarathy et al [2020]** briefs that the Speech emotion recognition (SER) systems find applications in various fields such as healthcare, education, and security and defense. A major drawback of these systems is their lack of generalization across different conditions. For example, systems that show superior performance on certain databases show poor performance when tested on other corpora. This problem can be solved by training models on large amounts of labelled data from the target domain, which is expensive and time-consuming. Another approach is to increase the generalization of the models. An effective way to achieve this goal is by regularizing the models through multitask learning (MTL), where auxiliary tasks are learned along with the primary task. These methods often require the use of labelled data which is computationally expensive to collect for emotion recognition (gender, speaker identity, age or other emotional descriptors). This study proposes the use of ladder networks for emotion recognition, which utilizes an unsupervised auxiliary task. The primary task is a regression problem to predict emotional attributes. The auxiliary task is the reconstruction of intermediate feature representations using a denoising auto encoder. This auxiliary task does not require labels so it is possible to train the framework in a semi-supervised fashion with abundant unlabelled data from the target domain.

**Reza Lotfian et al [2019]** proposed to design a curriculum for machine-learning to maximize the efficiency during the training process of deep neural networks (DNNs) for speech emotion recognition. Previous studies in other machine-learning problems have shown the benefits of training a classifier following a curriculum where samples are gradually presented in increasing level of difficulty. For speech emotion recognition, the challenge is to establish a natural order of difficulty in the training set to create the curriculum. We address this problem by assuming that, ambiguous samples for humans are also ambiguous for computers. Speech samples are often annotated by multiple evaluators to account for differences in emotion perception across individuals. While some sentences with clear emotional content are consistently annotated, sentences with more ambiguous emotional content present important disagreement between individual evaluations. We propose to use the disagreement between evaluators as a measure of difficulty for the classification task. We propose metrics that quantify the inter-evaluation agreement to define the curriculum for regression problems and binary and multi-class classification problems.

**Yi Lei et al [2022]** explained the synthetic speech is essential for many human-computer interaction and audio broadcast scenarios, and thus synthesizing expressive speech has attracted much attention in recent years. Previous methods performed the expressive speech synthesis either with explicit labels or with a fixed-length style embedding extracted from reference audio, both of which can only learn an average style and thus ignores the multi-scale nature of speech prosody. We propose MsEmoTTS, a multi-scale emotional speech synthesis framework, to model the emotion from different levels. Specifically, the proposed method is a typical attention-based sequence-to-sequence model and with proposed three modules, including global-level emotion presenting module (GM), utterance-level emotion presenting module (UM), and local-level emotion presenting module (LM), to model the global emotion category, utterance-level emotion variation, and syllable-level emotion strength, respectively. In addition to modeling the emotion from different levels, the proposed method also allows us to synthesize emotional speech in different ways, i.e., transferring the emotion from reference audio, predicting the emotion from input text, and controlling the emotion strength manually. Extensive experiments conducted on a Chinese emotional speech corpus demonstrate that the proposed method outperforms the compared reference audio-based and text-based emotional speech synthesis methods on the emotion transfer speech synthesis and text-based emotion prediction speech synthesis respectively. **Yue Xie et al [2019]** proposed an Automatic speech emotion recognition has been a research hotspot in the field of human-computer interaction over the past decade. However, due to the lack of research on the inherent temporal relationship of the speech waveform, the current recognition accuracy needs improvement. To make full use of the difference of emotional saturation between time frames, a novel method is proposed for speech recognition using frame-level speech features combined with attention-based long short-term memory (LSTM) recurrent neural networks. Frame-level speech features were extracted from waveform to replace traditional statistical features, which could preserve the timing relations in the original speech through the sequence of frames. To distinguish emotional saturation in different frames, two improvement strategies are proposed for LSTM based on the attention mechanism: first, the algorithm reduces the computational complexity by modifying the forgetting gate of traditional LSTM without sacrificing performance and second, in the final output of the LSTM, an attention mechanism is applied to both the time and the feature dimension to obtain the information related to the task, rather than using the output from the last iteration of the traditional algorithm.

**Lu Yi et al [2020]** briefs that the training data are scarce, it is challenging to train a deep neural network without causing the over fitting problem. For overcoming this challenge, this article proposes a new data augmentation network namely adversarial data augmentation network (ADAN) based on generative adversarial networks (GANs). The ADAN consists of a GAN, an auto encoder, and an auxiliary classifier. These networks are trained adversarially to synthesize class-dependent feature vectors in both the latent space and the original feature space, which can be augmented to the real training data for training classifiers. Instead of using the conventional cross-entropy loss for adversarial training, the Wasserstein divergence is used in an attempt to produce high-quality synthetic samples. The proposed networks were applied to speech emotion recognition using EmoDB and IEMOCAP as the evaluation data sets. It was found that by forcing the synthetic latent vectors and the real latent vectors to share a common representation, the gradient vanishing problem can be largely alleviated. Also, results show that the augmented data generated by the proposed networks are rich in emotion information.

**Norbert Braunschweiler et al [2022]** explained the recognition performance of deep learning models is influenced by multiple factors such as acoustic condition, textual content, style of emotion expression (e.g. acted, natural), etc. multiple factors are analysed by training and evaluating state-of-the-art deep learning models using the input modalities speech, text, and their combination across 6 emotional speech corpora. A novel deep learning model architecture is presented that further improves the state-of-the-art in multimodal emotion recognition with speech and text on the IEMOCAP corpus. Results from models trained on individual corpora show that combining speech and text improves performance only on corpora where the text of utterances varies across different emotions, while it reduced performance on corpora with fixed text expressed in different emotions, where the speech-only models performed better. Further, cross-corpus investigations are presented to understand the robustness to changing acoustic and textual content.

**CHAPTER-3**

**EXISTING SYSTEM**

**3.1 INTRODUCTION**

Emotion is defined as the positive or negative state of a person’s mind which is related with a pattern of physiological activities. Emotions describe the mental state of a person. Paul Ekman’s research work led him to categorize emotion into six basic classes: HAPPINESS, ANGER, FEAR, DISGUST, SURPEISE and SADNESS. These six basic emotions blend to form complex emotions. For ex: - disgust and anger unify to take a new form of emotion that is contempt. Human Machine interaction has become an issue of common concern over the past few decades.

Speech is the most natural and efficient way of communication. Earlier, traditional methods used manual processing of parameters from speech signals which were time consuming and costly. There are many applications which are earning profit by using emotion classification technique. In medical fields for online assessment of patients’ disorders, Smartphone interface personalized to automatically choose song based on the current emotional state of a person. A number of speech processing technologies are there which helps in analysing emotions from speech. Speech signals contain various information like age, sex, physical state of a person, etc.

Which can easily identify emotional state of a person. The classification of emotion from speech is done in a series of stages from extracting information from speech to classification of emotional content from speech. The stages involved in speech emotion classification are: the speech signal is pre-processed and segmented by converting it into a wave file, after pre-processing of speech signal features extraction algorithm like Linear Prediction Coefficient (LPC), Cepstrum Coefficients algorithms are applied to extract emotional content from speech signal and at last the classification algorithms are applied to classify the emotional content.

The paper discuss about the Speech Emotion Recognition System, the various features contained in a speech signal and the different classification algorithms applied to classify the emotional content of speech. The features like pitch, energy, fundamental frequency, formants, etc., are discussed how these are helpful in extracting the emotional information from speech signals. The various classification algorithms like Support Vector Machine (SVM) algorithm, Sequential Minimal Optimization (SMO) algorithm, Decision Tree algorithm, etc. are used for classification and the algorithm which provides the best emotion recognition rate is identified.



**Figure: 3.1** Algorithm

**3.2 SPEECH EMOTION RECOGNITION SYSTEM**

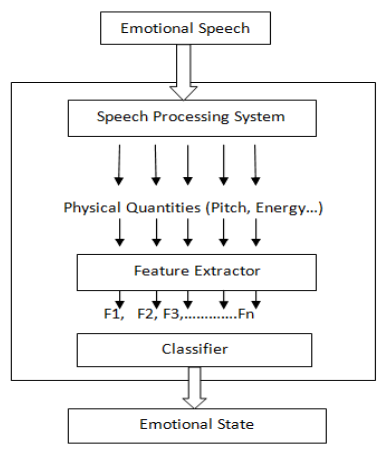
Speech is the most natural form through which humans communicate. Speech Recognition involves transformation of speech signal into a sequence of words with the help of an algorithm. Speech recognition is the capability of a machine to acknowledge the speech samples. The speech emotion recognition system is dependent on the naturalness of the database which contains speech as the input signal.

The accuracy of the emotion recognition rate in SER system is dependent on the database used. The database which will be used as an input source for the system should hold real time world emotions. The basic architecture for SER system.

(a) A speech processing system is used for extracting suitable features from the speech signal like pitch or energy,

(b) The features extracted are then summarized into reduced set of features,

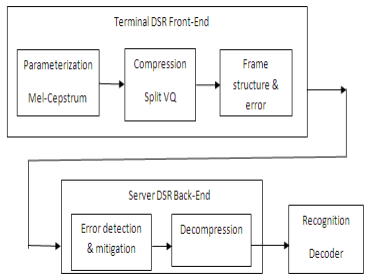
(c) A classifier is trained using supervised learning by providing example data how to correlate the features to the emotions.



**Figure: 3.2** Speech Recognition model

**3.3 DSR (DISTRIBUTED SPEECH RECOGNITION) MODEL**

ETSI published the first DSR in February, 2000. In this model, FRONT-END is used to transform the digitized speech into a stream of feature vectors and then the extracted features are sent to the BACK-END for further processing. The diagram below shows the working of front-end and back-end of the DSR model.



**Figure: 3.3** Distributed Speech Recognition

**3.4 FEATURE EXTRACTION**

The speech is partitioned into small intervals known as frames. The process of partitioning speech into frames based on the information they are carrying about emotion is known as feature extraction. Feature Extraction is a vital step in SER (Speech Emotion Recognition) System. Some of the features which helps to figure out emotions from the speech are,

**3.4.1 PITCH**

It is the main component of any speech which is defined as the lowness or highness of a voice as identified by the human ears. Pitch is dependent on the number or vibrations per second. The value of pitch parameter is extracted by using cepstrum in the frequency domain. Pitch helps in identifying the neutral and angry emotions from speech samples.

**3.4.2 ENERGY**

The intensity of the speech defines the energy level of speech. Energy level for each frame is calculated as: first the square of all the sample amplitudes is done and then summing up the values of all the squared sample amplitudes.

**3.4.3 PITCH DIFFERENCE and ENERGY DIFFERENCE**

The difference between the values of pitch or energy level of neighbouring segments is used to categorize the speech parameters into emotions. The more the fluctuation the more it is easier to reveal the lively emotions like happiness and anger.

**3.4.4 FORMANTS**

Formants are governed by the shape of the vocal tract and are manipulated by different emotions. For example, the state of excitement results in obtaining the higher mean values of the first formant frequency. The fundamental frequency (F0) helps in identifying the happy emotion from speech samples.

**3.4.5 MEL-FREQUENCY CEPSTRUM COEFFICIENTS (MFCC)**

MFCC is the most vital parameter in speech which best describes the emotional state by using simple calculations. MFCC also provides good frequency resolution when the speech frequency is low. MFCC based parameters show the energy migration in frequency domain and also helps in identifying phonetic characteristics of speech.

**3.5 CLASSIFICATION ALGORITHMS**

The Speech emotion classification systems are trained using the various classification algorithms. The system is trained using some data sets through machine learning algorithms. Various machine learning algorithms are used to recognize the basic human emotions from the given speech samples. The recognition rate is also achieved by applying various machine learning algorithms.

It has been generalized that the recognition rate for audio alone is 75% and for that of video alone is 70%. But if the speech is a joint audio-visual then the recognition rate achieved is 97%. Some of the algorithms which have been used to train our systems are as,

**3.5.1 SEQUENTIAL MINIMAL OPTIMIZATION (SMO) ALGORITHM**

SMO algorithm is used for solving the optimization problem which occurs when training the support vector machines. John Platt discovered the SMO algorithm in 1998 at Microsoft. LIBSVM tool is used to implement the SMO algorithm. The SMO algorithm is as,

* A Lagrange multiplier is found that contravene the KKT conditions for the optimization problem.
* Next multiplier is picked up for optimizing the pair.
* Steps 1 and 2 are repeated until convergence.

**3.5.2 SUPPORT VECTOR MACHINE (SVM) CLASSIFIER**

The main motive of the SVM classifier is to track down the hyper-planes with maximum obtainable margin that sets apart the data points into classes by identifying a weight vector and an offset. Support Vector Machine (SVM) classifier uses binary classification based on statistical learning theory. SVM transforms the original input set to a high dimensional feature space with the help of kernel function. This renovation can also be used for transforming non-linear problems.

SVM can have a very good classification performance even when there is a limited training data set. SVM has the capability to generalize new and accurate data by using the trained models designed in the learning phase. An adjustable weighted segmentation (AWS) is proposed to improve the accuracy rate of SVM classifier. AWS is a very simple approach in which each segment is assigned with a weight vector based on the type of emotion and the weights assigned are adjustable according to the input data.

**3.5.3 DECISION TREE ALGORITHM**

It is a hierarchical classifier in which each node signifies a choice between a number of alternatives and each leaf node signifies a decision to be taken. This classifier is similar to the if-then-else structure. In this algorithm, the mean and the Root Mean Square (RMS) of each and every feature in emotion class are calculated.

(3.1)

(3.2)

**3.5.4 K NEAREST NEIGHBOR (KNN)**

KNN is the most simplest and traditional classification algorithm which does not use any parameters i.e. it is a non-parametric method. The value of K nearest neighbours is provided as input from the feature space. The classification under KNN is done on the basis of number of nearest neighbours of an object. The output obtained is a set of classes. The object is assigned to a particular class on the basis of majority of votes of its nearest neighbours. The classification of samples is done by calculating the distance of that particular sample to the training set. KNN algorithm is independent of prior assumptions. Optimal value of K is chosen so that better results are obtained. But it is observed that as the value of K grows larger the effect of noise is reduced while training the data set.

**3.6 DISADVANTAGES OF EXISTING SYSTEM**

* Cepstral domain features or transforms for feature extraction has high complexity and thereby the run time increases.
* Accuracy level is low.

**CHAPTER-4**

**PROPOSED SYSTEM**

**4.1 INTRODUCTION**

Speech emotion recognition is the task of recognizing emotions from speech signals; this is very important in advancing human computer interaction: Human computer interaction is characterized as consisting of five major areas of study: research into interactional hardware and software, research into matching models, research at the task level, research into design, and research into organizational impact. Understanding one’s feelings at the time of communication is constructive in comprehending the conversation and responding appropriately.

Currently, this part of human computer interaction has not yet entirely been solved, and except for a limited number of applications, there is no general solution to this problem. Along with all major problems in machine learning, SER has started to gain an advantage from the tools made available by deep learning. Before the extensive employment of deep learning, SER was relying on methods like hidden Markov models (HMM), Gaussian mixture models (GMM), and support vector machines (SVM) along with lots of pre-processing and precise feature engineering.

However, with deep learning making up most of the new literature, the results are going up from around 70% accuracy to the upper 90s in controlled environments. Automatic SER helps smart speakers and virtual assistants to understand their users better, especially when they recognize dubious meaning words. For example, the term “really” can be used to question a fact or emphasize and stress out a statement in both positive and negative ways. Read the following sentences in different ways: “I really liked having that tool.” The same application can help translate from one language to another, especially as other languages have different ways of projecting emotions through speech.

SER is also beneficial in online interactive tutorials and courses. Understanding the student’s emotional state will help the machine decide how to present the rest of the course contents. Speech emotion recognition can also be very instrumental in vehicles’ safety features. It can recognize the driver’s state of mind and help prevent accidents and disasters. Another related application is in therapy sessions; by employing SER, therapists will understand their patients’ state and possibly underlying hidden emotions as well.

It has been proven that in stressful and noisy environments like aircraft cockpits. The application of SER can significantly help to increase the performance of automatic speech recognition systems. The service industry and e-commerce can utilize speech emotion recognition in call centres to give early alerts to customer service and supervisors of the caller’s state of mind. In addition, speech emotion recognition has been suggested to be implemented in interactive movies to understand viewers’ emotions. The interactive film could then go along different routes and have different endings.

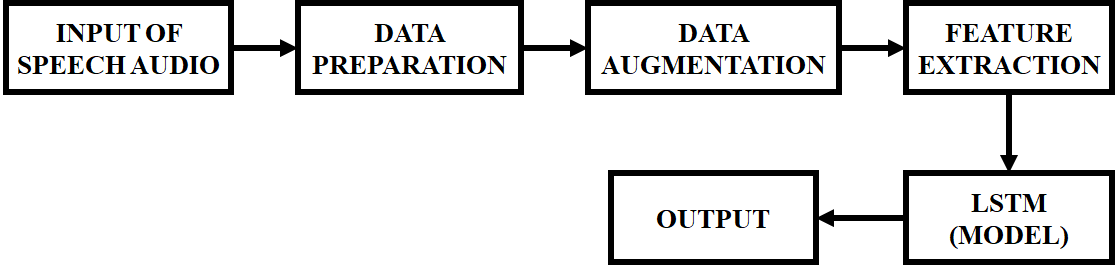
To train machine learning algorithms to classify emotions, we need to have training datasets. For SER tasks. There are generally three types of training datasets, natural, semi-natural, and simulated. The natural datasets are extracted from available videos and audios, either broadcasted on TV or online. There are also databases from call centres and similar environments. Semi-natural datasets are made by defining a scenario for professional voice actors and asking them to play them.

The third and most widely used type, the simulated datasets, are similar to semi-naturals. The difference is that the voice actors are acting the same sentences with different emotions. Traditionally SER used to follow the steps of automatic speech recognition (ASR), and methods based on HMMS, GMMs, and SVMs were widespread. Those approaches needed lots of feature engineering and any changes in the features usually required restructuring the entire architecture of the method.

However, lately, by the development of deep learning tools and processes, solutions for SER can be changed as well. There is a lot of effort and research on employing these algorithms to recognize emotions from the speech. In addition to deep learning, more recently, along with improvements in recurrent neural networks and the use of long short-term memory (LSTM) networks, auto encoders, and generative adversarial models, there has been a wave of studies on SER using these techniques to solve the problem.

The rest of the paper is organized as follows: In Section 2, we define SER, and in Section 3, we present some related studies. In Section 4, we provide a review of existing emotional speech datasets, and Section 5 is where we review several traditional and deep learning methods used in SER. Finally, in the last chapter, we discuss and conclude our work while proposing direction for future actions in SER.

**4.2 PROPOSED SYSTEM**

****

**Figure: 4.1** Proposed Block Diagram

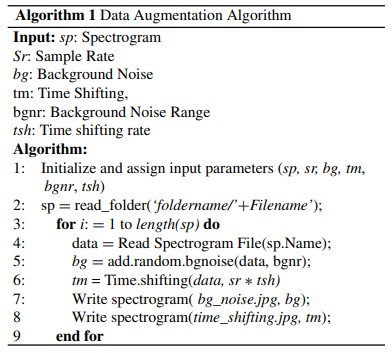
**4.3 DATA AUGMENTATION**

In situations where the original data size is limited, data augment is needed to surmount this issue of data shortage. Data augment is the production of extra training data samples by performing a series of deformations on the data in the training dataset. The essential basis in data enhancement is that the labels of the new data created by deformations applied to the tagged data are not changed. There are many methods for data augment, such as rotating the image at different angles, horizontally rotating and vertically rotating, adding noise and colour manipulation to the image.

In this study, 2 different processes were applied to the signal before extracting the spectrogram and these processes are explained below,

* Background Noise: Added random noise in the range of [0.1, 0.5] to the sound samples.
* Time Shifting: The sound is shifted from the starting point and the original length is preserved. Each sample shifted from the starting point to 0.3 seconds.

The sound is shifted from the starting point and the original length is preserved. Each sample shifted from the starting point to 0.3 seconds.



The sound samples obtained as a result of increasing data augmentation were added to the original dataset as additional. Training examples. In Algorithm 1, the pseudo-code of the data augmentation algorithm is given.

**4.4 FEATURE EXTRACTION**

The purpose of the acoustic analysis is to separate the speech signal into its components and to present parametric measurements of these components. Acoustic features are physical properties in terms of frequency, loudness and amplitude. Speech signals are pre-processed before acoustic features are extracted. While recording, samples of speech may contain unwanted information such as noise, depending on environmental factors. In this study, the Butterworth filter is used to remove noise in speech samples. Also, the speech signal is divided into frames of 30ms with an overlap of 15ms. LibROSA toolbox is used to extract acoustic sound features from different groups. LibROSA is a widely used library for music and sound analysis. Extracted acoustic features are Root Mean Square energy (RMS), MFCC, Chroma, Spectrum centroid, Spectral entropy, Skewness, Attack time and Zero crossing rate. The size of the acoustic property vector obtained is 32.

**4.4.1 RMS**

It is the measure of the loudness of an audio signal. It is found by calculating the square root of the sum of the mean squares of the amplitudes of the sound samples. RMS formula is given in Equation: 4.1.

(4.1)

**4.4.2 SPECTRUM CENTROID**

It is usually associated with a measure of the brightness of a sound and is a measure of where the centre of mass of the spectrum is. Higher centroid values indicate higher frequency values.

**4.4.3 SPECTRAL ENTROPY**

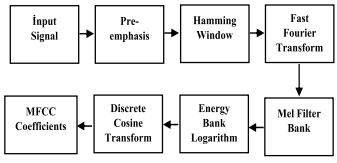
The probabilities of the power spectrum components of the signal are taken into account when calculating this value. The normalized power distribution in the frequency domain of the signal is evaluated as the probability distribution.

**4.4.4 SKEWNESS**

Indicates the degree of asymmetry of a distribution around its mean, and it is the average skewness coefficient of the spectral distribution in the Lower frequency bands.

**4.4.5 MFCC**

It is based on human hearing perception and is one of the most used feature extraction methods in the field of sound processing. MFCC features are based on obtaining distinctive values for speakers by imitating the frequency selectivity of the human ear. The steps to be taken to extract MFCC features are given in Figure: 4.2.



**Figure: 4.2** Steps of MFCC

Conversion between Mel scale (M) and frequency scale (Hz) can be done using equations 2 and 3 given below,

(4.2)

(4.3)

One way to express this perceptual structure is to use a triangular band-pass mel filter bank. Mel Frequency Kepstrum coefficients are obtained by applying discrete cosine transform after filter bank. MFCC is calculated according to equation: 4.4.

(4.4)

**4.4.6 ATTACK TIME**

It is the estimation of the time it takes for a signal to reach to its peak. A simple way to define and calculate this feature is to predict the time duration of the range of the phase where the signal’s amplitude rises.

**4.4.7 CHROMA**

The notes relate to the energy density around them and provide important information about the harmonic content of the sound. There are 7 notes in western music, and since the two notes are divided into two equal parts, except for the E and F notes, 12 features can be obtained by taking the sounds in between.

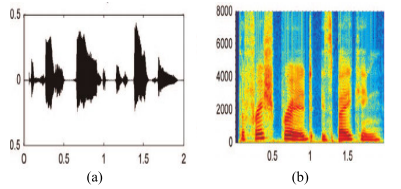
**4.4.8 ZERO-CROSSING RATE**

It is the rate of the transition of a signal through the zero lines, that is, the change of the sign. The X-axis shows how many times the signal has passed, it can be used as an indication of noise as well as frequency.

**4.5 SPECTROGRAM EXTRACTION**

There may be silence at some moments during the conversation and no emotion may arise. This is a factor that makes emotion recognition difficult and the moments when no emotion is felt should be filtered. However, since the duration of the speech sound recordings in the datasets used is not very long, the entire part of the speech sound recordings was taken into account in spectrogram extraction. Since speech signals are not static signals, the signal must be processed in small frames. Speech signals are first divided into 30 ms frames in order to obtain the spectrograms.

Also, each frame comes out in such a way that it overlaps a part of the previous frame. The overlap ratio of the frames is chosen as 50% of a frame. Windowing is applied after framing is applied on a signal. The aim is to prevent discontinuity that may occur at the extreme ends of each frame. The widely preferred ‘‘Hamming window’’ technique is employed in this study. Hamming Window minimizes unwanted radiation from the extreme-ends of the regions of the signal. It is also the function that makes the signal convenient for the Fourier transform. The algorithm of the hamming function is as follows: The hamming window is multiplied by the framed sound signal and then the windowed signal is finally obtained. Hamming window formula is given in equation: 4.5.



**Figure: 4.3** Illustration of Speech Sound Signal and Spectrogram (a) Speech Sound Signal, (b) Spectrogram

(4.5)

Fourier transform is performed on the signal after the Hamming window is performed. The Fourier transform transforms the signal from the time domain to the frequency domain. Fast Fourier Transform (FFT) is used in this study. In this step, each frame consisting of N samples is passed from the time domain to frequency domain by performing Fast Fourier transform. A set with N samples is defined as in equation: 4.6. At the last stage, the power spectrograms of the signals which Fourier transform is applied, are extracted. An example speech audio signal and spectrogram image are given in Figure: 3.3.

(4.6)

**4.6 CONVOLUTIONAL NEURAL NETWORKS**

Convolutional neural networks (CNNs) or shift invariant artificial neural networks (SIANNs) are particular types of neural networks that, in their hidden layer they have different filters or regions that respond to a specific feature of the input signal. Their design is based on the research by Hubel and Wiesel in 1968, which introduces the visual neural cortex as a spatially specialized structure, in which every region responds to a specific characteristic of the input signal. One positive perspective of CNNs is the ability to learn features from high-dimensional input data.

However, it also learns features from small variations and distortion appearance that leads to the large storage requirement at the time of development. Hence, in CNNs, there usually exists a layer of convolution followed by a down sampling mechanism. The convolution layer has various filter banks, in which their weights will be tuned through often back propagation. A convolutional neural network that is capable of detecting angry, happy, and sad emotions with a 66.1% accuracy rate.

They have compared their blind trained network outcome with a baseline Feature based SVM. To be able to train and test their method, they have used a corpus of TED talks, labelled by students and crowd sourced. They have implemented their CNN using the Theano toolkit. For comparison, they have trained a linear SVM with a feature set from the INTERSPEECH 2009 emotion challenge.

They reported that their CNN network, with less than a couple hundred milliseconds response time, can detect three classes of angry, happy, and sad with a 66.1% accuracy rate. They showed that their neural activity is concentrated around the fundamental frequencies most correlated to the emotions. A one-dimensional convolutional neural network and report the accuracy of 96.60% in classifying negative emotions from Tai language datasets. In their study, the developed method has been applied to SAVEE, RAVDESS, TESS, and CREMA-D and Tai datasets.

**4.7 DEEP CONVOLUTIONAL NEURAL NETWORKS**

Deep convolutional networks usually consist of multiple layers of convolution nodes, followed by one or more fully connected layers to finish the classification task. In SER, there are many efforts on deep convolutional neural networks, which we will review some of the most recent ones in the following part. A method based on a deep neural network containing convolutional pulling and fully connected layers. They have implemented their system on the Berlin Database of Emotional Speech. To compare to previous research, they have limited their classes to angry, neutral, and sad.

In their system, they have removed silence from their signals and then divided the files into 20 ms chunks with no overlap. In their network, before any feature selection, they have six layers of convolution with succeeding dropout layers with p values of 0.1 and after that follows a lattice of two parallel feature selectors and then series of fully connected layers. The segment accuracy of their system was 77.51%, but the file level accuracy was 96.97%, with a 69.55% confidence rate. Although the system’s accuracy on file level was high, in real scenarios, there is no sign to point a chunk of speech, and the system needs to improve on independent detection.

An Emotion recognition system based on a deep convolutional neural network designed for the LSVRC-2010 contest. This net- work, Alex Net, is also pre-trained with a dataset of 1.2 million images, then fine-tuned using samples that they had from EMO-DB. Using this system, they can recognize three classes of emotions (angry, sad, and happy) plus a neutral category. Moreover, they have demonstrated that their system can have accuracies over 80% with EMO-DB, about 20% more than the baseline SVM standard.

They have also applied their method to 3 other databases (RML, eNterface05, and Baum-1s) and were able to get results higher than the baseline method. In their system, the focus was on how automatic feature selection in deep convolutional neural networks can outperform feature selection in shallow convolutional networks and Statistical model-based methods like GMM and HMM. One of the essential features used in this system was the use of the discriminant temporal pyramid matching (DTPM) strategy.

Which helps in concatenating the learned segment level feature to form an utterance level feature representation. Deep convolutional neural networks are potent in modelling the smallest transient attributes of the signal. However, this capability comes with the cost of exponentially more variables to tune, and this means more samples are needed to train the system. In the case of image applications, these networks get trained by millions of samples. However, in SER, usually, the numbers of the samples are limited to thousands.

**4.8 LSTM NETWORKS**

Recurrent neural networks can learn and react to the temporal event without changing the slowly shaped weights thanks to their feedback connection, forming short-term activations for recent events. This feature can be beneficial in case of applications that time is an essential feature, like Speech Processing, music composition, and video description. However, as they are trained using Back Propagation through Time, error signals flowing backward in time can either become bigger and bigger or vanish depending on the size of the weights. This will create either oscillating weights or makes the network to be slow to train and converge.

To be able to incorporate the short-term adaptation of RNNs and avoid the problems above, a new architecture called Long Short Term Memory in 1997. LSTM networks are capable of bridging time intervals bigger than 1000 steps, even if input sequences are incompressible and noisy. They are incorporating a gradient-based algorithm enforcing constant error flow through individual units, specially designed to handle the short-term; thus, they can truncate the gradient computations at a definable point without affecting the long-term activations.

In recent years, LSTM networks were becoming in the centre of attention for many applications involving time series events. Speech Processing and especially speech emotion recognition are two of these applications. An early proposal for using LSTM networks in 2013 was in the work. They have proposed a multimodal LSTM based classification network, exploiting acoustic, linguistic, and visual information. In their study, they have compared both unidirectional and bidirectional LSTM networks. They also have compared their proposed results with the AVEC 2011 Audio/Visual Emotion. In this research, they extract 1941 audio features composed of Prosodic, Spectral, and Voice quality features, linguistic word-level content, and all the video features extracted by applying the Viola-Jones method, segmented optical flow, and head tilt.

Then all the features are being fed to a unidirectional and a bidirectional LSTM network. A context-aware system for end to end recognition of emotions in speech using CNNs followed by LSTM networks. The big difference in their method versus other deep learning algorithms is that they do not preselect features before training the network. They introduce raw input to the system and let the black box choose the best representations for the features. In their system, they create a segment of 6 s of raw audio first and pre-process it at a 16 kHz sample rate.

They then pass the signal through the first convolution layer of 40 filters with a kernel of size 2 to bold the temporal features. The output of the first convolution layer is fed to pooling of size 2; the results are then fed to another convolution layer with a kernel of size 10 to smooth the temporal and extract the spectral features. Then it is fed to a pooling layer of size 20, reducing the dimensions of the data drastically. The reduced data for 6 s segments are then fed to a recurrent layer divided into 40 ms blocks; the LSTM layer is made of 128 cells.

A framework based on two class of network blocks, a single layer convolutional block creating a local feature learning block (LFLB) and an LSTM block to learn global features. In their research, they have tried both 1-dimensional convolutions with the raw audio signal and 2-dimensional convolutions with Log Mel Spectrogram (LMS) features. The LFLB block in their research is constructed by a convolution layer followed by a batch normalization (BN).

An exponential linear unit (ELU), and a max-pooling at the end. Adding batch normalization by keeping mean and variance fixed helps every layer to be immune from large fluctuations from the previous layer, thus will improve the stability of the network, also by keeping input values in a limited range can help the model to converge faster and shorten training time. The next block speeding up the learning in LFLB is ELU. To rectified linear units, tries to solve the vanishing gradient problem.

Still, to improve the learning characteristics, ELU can have negative values that helps to push the mean activations closer to zero. Thus, increasing stability and speeding up the calculations by having smaller numbers in equations. The last part in LFLB is the max pooling, which is widely used in almost every convolutional network. After extracting the low-level local features, the LSTM part has the job of extracting long term contextual dependencies.

LSTM has four components that can influence the state of the block: Input gate, Output gate, forget gate, and a cell with a recurrent connection to itself. The next stage in this proposed method is the fully connected layer for classification. Their network in both cases of 1D and 2D is structured as follows, two layers of LFLB with 64 filters followed by two layers of LFLB with 128 filters followed by an LSTM layer with a kernel size of 256 and lastly, a fully connected layer.

The convolution kernel in both implementations has a size of 3 and stride of one. To get the best results, they have run their experiments in both speaker-dependent and speaker-independent setting. In speaker dependent settings, they have reached up to 92% accuracy for 1D and 95% accuracy for 2D networks. Additionally, in a speaker-independent setting, they have reached up to 62% accuracy for 1D and 82% accuracy for 2D networks. In another research late in 2019, a system based on two layers of modified LSTMs with 512 and 256 hidden units, followed by a layer of attention weighting on both time dimension and feature dimension and two fully connected layers at the end.

In their research, they have stated that humans’ attention on the whole stimuli is not balanced, and it has been shown incorporating this concept creates excellent results in image processing. Therefore, they have proposed a self-attention mechanism to the forgetting gate of an LSTM layer, which results in the same performance while reducing the computations. Their change in coupled LSTM structure reduces the forgetting gate required four training parameters to two parameters while experimentally does not affect the performance of the final LSTM mode.

The input of the fully connected layer or any other block needs a predetermined length of features. In contrast, the LSTM output length varies based on the variable length of the input data. To solve this problem, they have proposed an attention weighting method for the output of all time steps and feature dimensions and then combining the results as a final output of the LSTM block. They have experimented with five combinations of their proposed methods, LSTM with Time attention, LSTM with feature attention, LSTM with both time and feature attention, LSTM with modified forget gate, and LSTM with modified forget gate and time and feature attention.

Additionally, as the results on their English speech dataset eNTERFACE, they have reached to 89.6% UAR accuracy in which they claim is the best result on that dataset. LSTM networks have shown to be very effective in time series like data due to their pattern history memorizing capability. One of the default applications of such a system is speech emotion recognition. LSTM based systems are very well capable of learning the spectral features of the signal. When coupled with CNNs to learn the temporal characteristics of the signal, they can form a competent system to model and learn the samples. All the mentioned exciting capabilities of LSTMs come with the cost of more process and exponential memory requirements. They also, similar to CNNs, need a vast number of training samples to tune their large number of variables.

**CHAPTER-5**

**RESULTS AND DISCUSSION**

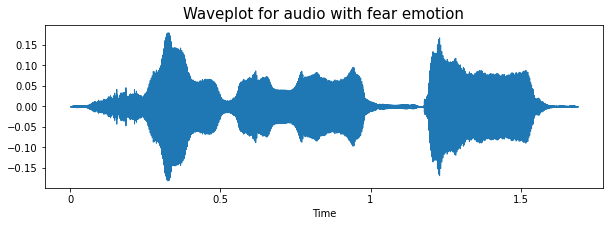
**5.1 EXPERIMENTAL APPLICATIONS**

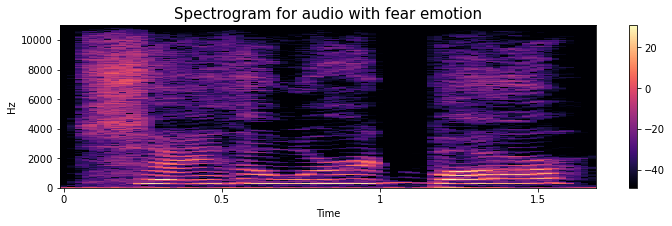
In this study, three different voice datasets are utilized, which are widely used by researchers in emotion recognition. This dataset contains audio and visual recordings of 12 male and 12 female actors pronouncing English sentences with eight different emotional expressions. Only speech samples are used for this study. Emotion tags in the dataset are: sad, happy, angry, calm, fearful, surprised, and neutral and disgust. Besides, the total number of records in the dataset is 1440. Which is widely used by researchers in the field of speech-based emotion recognition and allows us to make more extensive comparisons with previous studies.

The dataset contains 535 audio outputs divided into 7 emotion classes in German. Emotion classes in the dataset are; anger, sadness, fear/anxiety, neutral, happiness, disgust, and boredom. This dataset consists of audio, video and facial movement samples collected from five pairs of male and female actors. The audio files of the data series are divided into ten emotion classes: angry, happy, sad, neutral, frustrated, excited, fearful, surprised, disgusted and other. The proposed frame on improvised data. In addition, we consider only 4 emotion classes (angry, happy, neutral, and sad) dataset in this study. The number of audio files in four classes is 889.

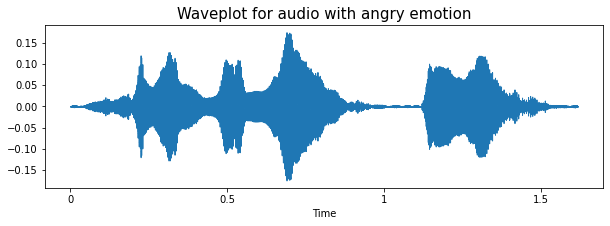
****

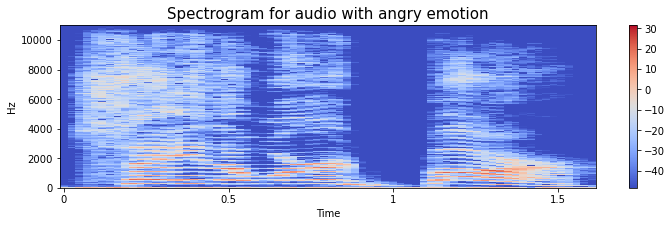
**Figure: 5.1** Data Set

****

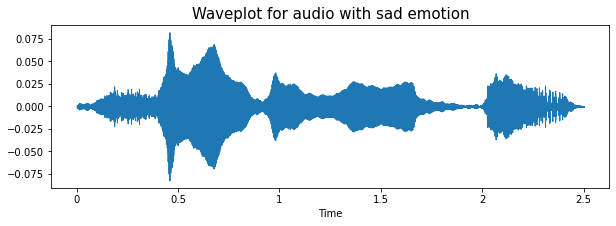
****

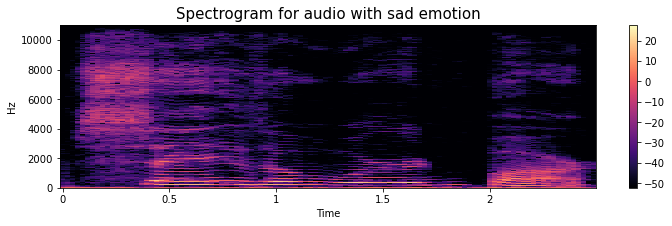
**Figure: 5.2** Waveplot And Spectrogram for Audio with Fear

****

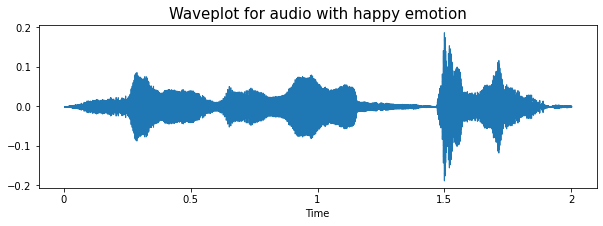
****

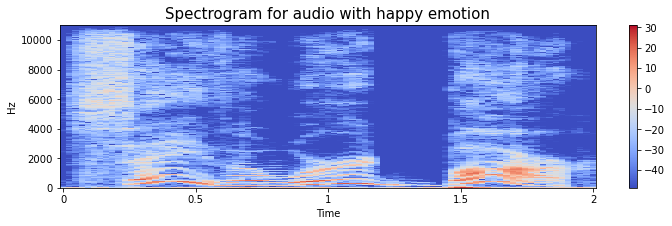
**Figure: 5.3** Waveplot And Spectrogram for Audio with Angry Emotion

****

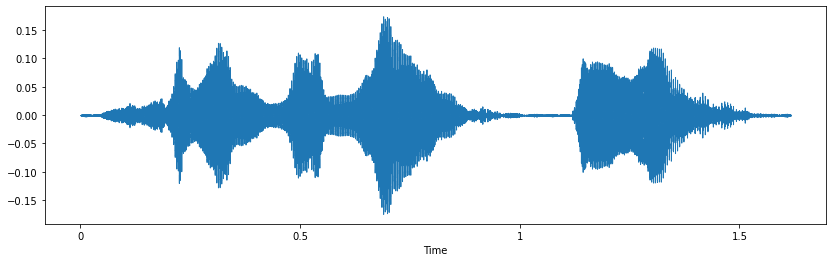
****

**Figure: 5.4** Waveplot And Spectrogram for Audio with Sad Emotion

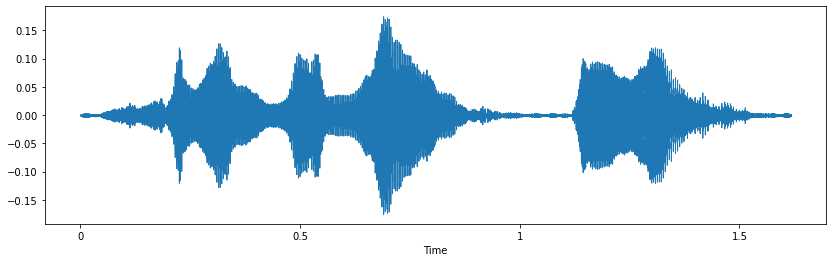
****

****

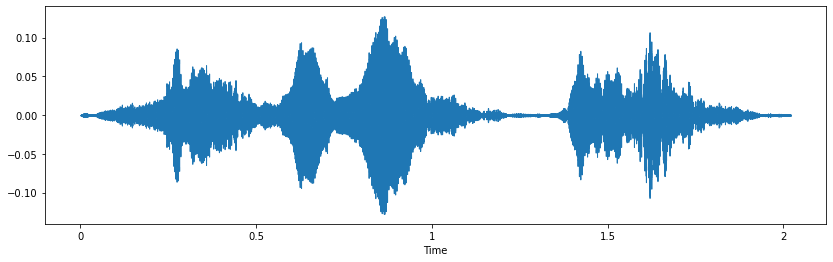
**Figure: 5.5** Waveplot and Spectrogram for Audio with Happy Emotion

****

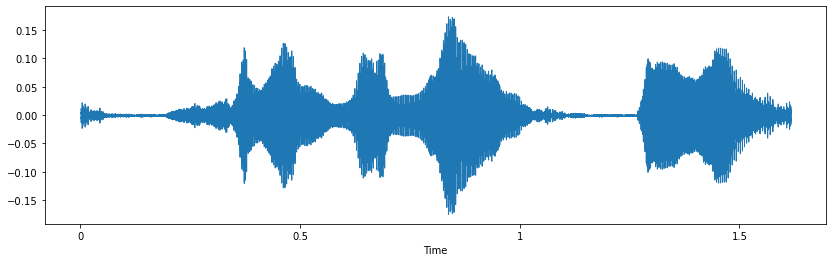
**Figure: 5.6** Simple Audio

****

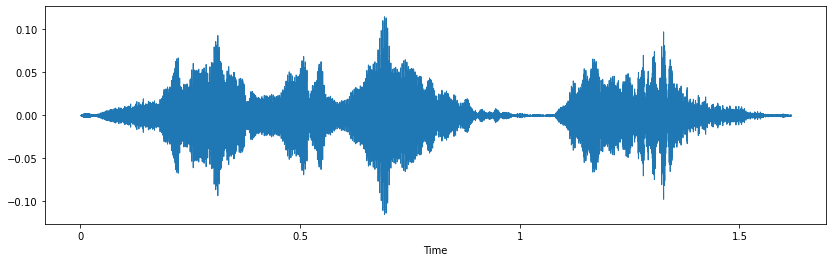
**Figure: 5.7** Noise Injection

****

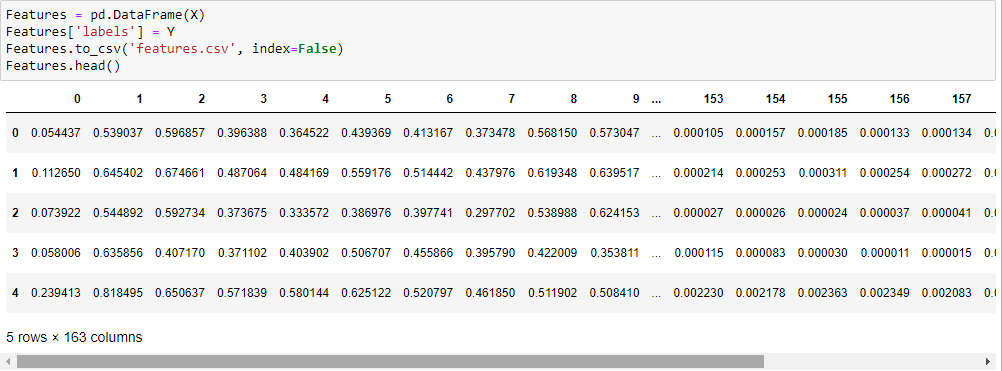
**Figure: 5.8** Stretching

****

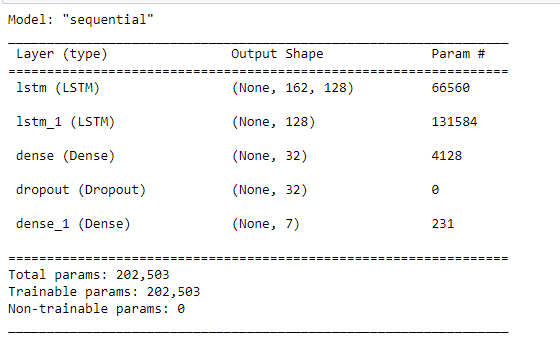
**Figure: 5.9** Shifting

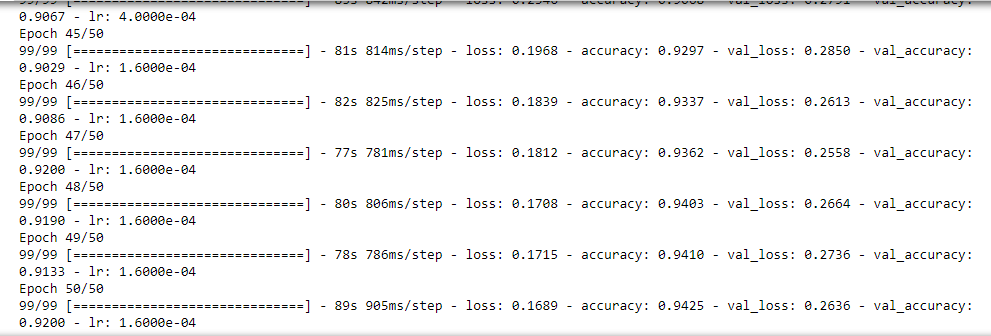
****

**Figure: 5.10** Pitch

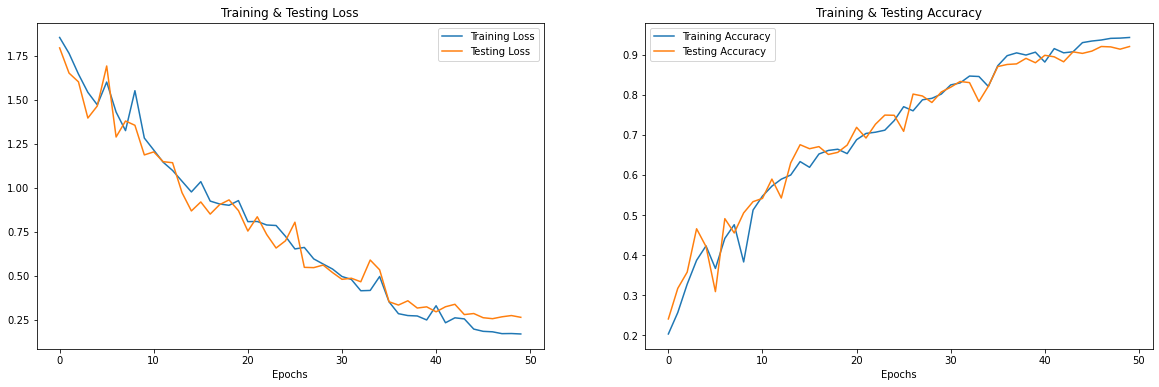
****

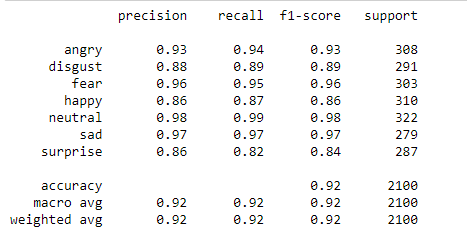
**Figure: 5.11** Feature Extraction

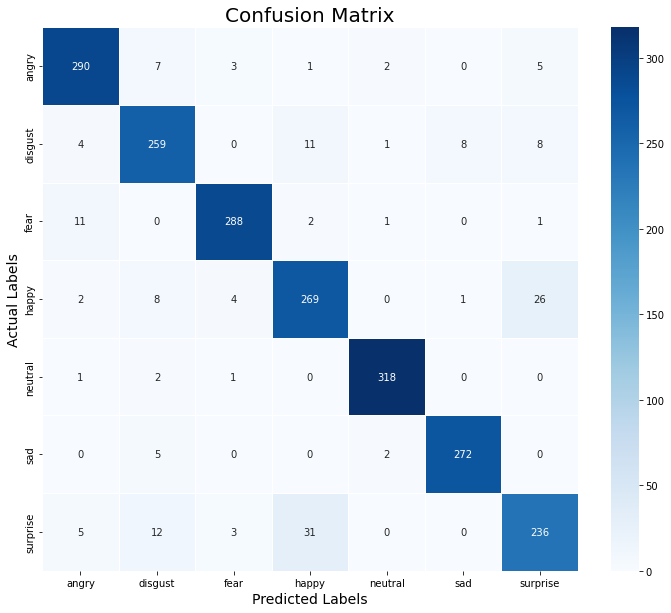
****

****

**Figure: 5.12** Build Model

****

****

****

**Figure: 5.13** Model Loss And Accuracy

**CHAPTER-6**

**SOURCE CODE**

**SAMPLE CODE**

import pandas as pd

import numpy as np

import os

import sys

# librosa is a Python library for analyzing audio and music. It can be used to extract the data from the audio files we will see it later.

import librosa

import librosa.display

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn.preprocessing import StandardScaler, OneHotEncoder

from sklearn.metrics import confusion\_matrix, classification\_report

from sklearn.model\_selection import train\_test\_split

# to play the audio files

from IPython.display import Audio

import keras

from keras.callbacks import ReduceLROnPlateau

from keras.models import Sequential

from keras.layers import Dense, Conv1D, MaxPooling1D, Flatten, Dropout, BatchNormalization

from tensorflow.keras.utils import to\_categorical

#from keras.utils import np\_utils, to\_categorical

from keras.callbacks import ModelCheckpoint

import warnings

if not sys.warnoptions:

warnings.simplefilter("ignore")

warnings.filterwarnings("ignore", category=DeprecationWarning)

Tess = "D:\\C:\Users\kolar\Downloads\SPEECH REGONATION USING LSTM\SPEECH REGONATION USING LSTM\\"

tess\_directory\_list = os.listdir(Tess)

file\_emotion = []

file\_path = []

for dir in tess\_directory\_list:

directories = os.listdir(Tess + dir)

for file in directories:

part = file.split('.')[0]

part = part.split('\_')[2]

if part=='ps':

file\_emotion.append('surprise')

else:

file\_emotion.append(part)

file\_path.append(Tess + dir + '/' + file)

# dataframe for emotion of files

emotion\_df = pd.DataFrame(file\_emotion, columns=['Emotions'])

# dataframe for path of files.

path\_df = pd.DataFrame(file\_path, columns=['Path'])

Tess\_df = pd.concat([emotion\_df, path\_df], axis=1)

Tess\_df.head()

# creating Dataframe using all the 4 dataframes we created so far.

data\_path = pd.concat([ Tess\_df], axis = 0)

data\_path.to\_csv("data\_path.csv",index=False)

data\_path.head()

plt.title('Count of Emotions', size=16)

sns.countplot(data\_path.Emotions)

plt.ylabel('Count', size=12)

plt.xlabel('Emotions', size=12)

sns.despine(top=True, right=True, left=False, bottom=False)

plt.show()

def create\_waveplot(data, sr, e):

plt.figure(figsize=(10, 3))

plt.title('Waveplot for audio with {} emotion'.format(e), size=15)

librosa.display.waveshow(data, sr=sampling\_rate)

plt.show()

def create\_spectrogram(data, sr, e):

# stft function converts the data into short term fourier transform

X = librosa.stft(data)

Xdb = librosa.amplitude\_to\_db(abs(X))

plt.figure(figsize=(12, 3))

plt.title('Spectrogram for audio with {} emotion'.format(e), size=15)

librosa.display.specshow(Xdb, sr=sr, x\_axis='time', y\_axis='hz')

#librosa.display.specshow(Xdb, sr=sr, x\_axis='time', y\_axis='log')

plt.colorbar()

emotion='fear'

path = np.array(data\_path.Path[data\_path.Emotions==emotion])[1]

data, sampling\_rate = librosa.load(path)

create\_waveplot(data, sampling\_rate, emotion)

create\_spectrogram(data, sampling\_rate, emotion)

Audio(path)

emotion='angry'

path = np.array(data\_path.Path[data\_path.Emotions==emotion])[1]

data, sampling\_rate = librosa.load(path)

create\_waveplot(data, sampling\_rate, emotion)

create\_spectrogram(data, sampling\_rate, emotion)

Audio(path)

emotion='sad'

path = np.array(data\_path.Path[data\_path.Emotions==emotion])[1]

data, sampling\_rate = librosa.load(path)

create\_waveplot(data, sampling\_rate, emotion)

create\_spectrogram(data, sampling\_rate, emotion)

Audio(path)

emotion='happy'

path = np.array(data\_path.Path[data\_path.Emotions==emotion])[1]

data, sampling\_rate = librosa.load(path)

create\_waveplot(data, sampling\_rate, emotion)

create\_spectrogram(data, sampling\_rate, emotion)

Audio(path)

def noise(data):

noise\_amp = 0.035\*np.random.uniform()\*np.amax(data)

data = data + noise\_amp\*np.random.normal(size=data.shape[0])

return data

def stretch(data, rate=0.8):

return librosa.effects.time\_stretch(data, rate)

def shift(data):

shift\_range = int(np.random.uniform(low=-5, high = 5)\*1000)

return np.roll(data, shift\_range)

def pitch(data, sampling\_rate, pitch\_factor=0.7):

return librosa.effects.pitch\_shift(data, sampling\_rate, pitch\_factor)

path = np.array(data\_path.Path)[1]

data, sample\_rate = librosa.load(path)

#simple audio

plt.figure(figsize=(14,4))

librosa.display.waveshow(y=data, sr=sample\_rate)

Audio(path)

#noise injection

x = noise(data)

plt.figure(figsize=(14,4))

librosa.display.waveshow(y=x, sr=sample\_rate)

Audio(x, rate=sample\_rate)

#stretching

x = stretch(data)

plt.figure(figsize=(14,4))

librosa.display.waveshow(y=x, sr=sample\_rate)

Audio(x, rate=sample\_rate)

#shifting

x = shift(data)

plt.figure(figsize=(14,4))

librosa.display.waveshow(y=x, sr=sample\_rate)

Audio(x, rate=sample\_rate)

#pitch

x = pitch(data, sample\_rate)

plt.figure(figsize=(14,4))

librosa.display.waveshow(y=x, sr=sample\_rate)

Audio(x, rate=sample\_rate)

def extract\_features(data):

# ZCR

result = np.array([])

zcr = np.mean(librosa.feature.zero\_crossing\_rate(y=data).T, axis=0)

result=np.hstack((result, zcr)) # stacking horizontally

# Chroma\_stft

stft = np.abs(librosa.stft(data))

chroma\_stft = np.mean(librosa.feature.chroma\_stft(S=stft, sr=sample\_rate).T, axis=0)

result = np.hstack((result, chroma\_stft)) # stacking horizontally

# MFCC

mfcc = np.mean(librosa.feature.mfcc(y=data, sr=sample\_rate).T, axis=0)

result = np.hstack((result, mfcc)) # stacking horizontally

# Root Mean Square Value

rms = np.mean(librosa.feature.rms(y=data).T, axis=0)

result = np.hstack((result, rms)) # stacking horizontally

# MelSpectogram

mel = np.mean(librosa.feature.melspectrogram(y=data, sr=sample\_rate).T, axis=0)

result = np.hstack((result, mel)) # stacking horizontally

return result

def get\_features(path):

# duration and offset are used to take care of the no audio in start and the ending of each audio files as seen above.

data, sample\_rate = librosa.load(path, duration=2.5, offset=0.6)

# without augmentation

res1 = extract\_features(data)

result = np.array(res1)

# data with noise

noise\_data = noise(data)

res2 = extract\_features(noise\_data)

result = np.vstack((result, res2)) # stacking vertically

# data with stretching and pitching

new\_data = stretch(data)

data\_stretch\_pitch = pitch(new\_data, sample\_rate)

res3 = extract\_features(data\_stretch\_pitch)

result = np.vstack((result, res3)) # stacking vertically

return result

X, Y = [], []

for path, emotion in zip(data\_path.Path, data\_path.Emotions):

feature = get\_features(path)

for ele in feature:

X.append(ele)

# appending emotion 3 times as we have made 3 augmentation techniques on each audio file.

Y.append(emotion)

len(X), len(Y), data\_path.Path.shape

Features = pd.DataFrame(X)

Features['labels'] = Y

Features.to\_csv('features.csv', index=False)

Features.head()

#Data preparation

X = Features.iloc[: ,:-1].values

Y = Features['labels'].values

# As this is a multiclass classification problem onehotencoding our Y.

encoder = OneHotEncoder()

Y = encoder.fit\_transform(np.array(Y).reshape(-1,1)).toarray()

# splitting data

x\_train, x\_test, y\_train, y\_test = train\_test\_split(X, Y, random\_state=0, shuffle=True)

x\_train.shape, y\_train.shape, x\_test.shape, y\_test.shape

# scaling our data with sklearn's Standard scaler

scaler = StandardScaler()

x\_train = scaler.fit\_transform(x\_train)

x\_test = scaler.transform(x\_test)

x\_train.shape, y\_train.shape, x\_test.shape, y\_test.shape

# making our data compatible to model.

x\_train = np.expand\_dims(x\_train, axis=2)

x\_test = np.expand\_dims(x\_test, axis=2)

x\_train.shape, y\_train.shape, x\_test.shape, y\_test.shape

x\_train.shape

model=Sequential()

model.add(keras.layers.LSTM(128, input\_shape=(x\_train.shape[1], x\_train.shape[2]), return\_sequences=True))

model.add(keras.layers.LSTM(128))

model.add(Dense(units=32, activation='relu'))

model.add(Dropout(0.3))

model.add(Dense(units=7, activation='softmax'))

model.compile(optimizer = 'adam' , loss = 'categorical\_crossentropy' , metrics = ['accuracy'])

model.summary()

rlrp = ReduceLROnPlateau(monitor='loss', factor=0.4, verbose=0, patience=2, min\_lr=0.0000001)

history=model.fit(x\_train, y\_train, batch\_size=64, epochs=50, validation\_data=(x\_test, y\_test), callbacks=[rlrp])

print("Accuracy of our model on test data : " , model.evaluate(x\_test,y\_test)[1]\*100 , "%")

epochs = [i for i in range(50)]

fig , ax = plt.subplots(1,2)

train\_acc = history.history['accuracy']

train\_loss = history.history['loss']

test\_acc = history.history['val\_accuracy']

test\_loss = history.history['val\_loss']

fig.set\_size\_inches(20,6)

ax[0].plot(epochs , train\_loss , label = 'Training Loss')

ax[0].plot(epochs , test\_loss , label = 'Testing Loss')

ax[0].set\_title('Training & Testing Loss')

ax[0].legend()

ax[0].set\_xlabel("Epochs")

ax[1].plot(epochs , train\_acc , label = 'Training Accuracy')

ax[1].plot(epochs , test\_acc , label = 'Testing Accuracy')

ax[1].set\_title('Training & Testing Accuracy')

ax[1].legend()

ax[1].set\_xlabel("Epochs")

plt.show()

# predicting on test data.

pred\_test = model.predict(x\_test)

y\_pred = encoder.inverse\_transform(pred\_test)

y\_test = encoder.inverse\_transform(y\_test)

df = pd.DataFrame(columns=['Predicted Labels', 'Actual Labels'])

df['Predicted Labels'] = y\_pred.flatten()

df['Actual Labels'] = y\_test.flatten()

df.head(10)

cm = confusion\_matrix(y\_test, y\_pred)

plt.figure(figsize = (12, 10))

cm = pd.DataFrame(cm , index = [i for i in encoder.categories\_] , columns = [i for i in encoder.categories\_])

sns.heatmap(cm, linecolor='white', cmap='Blues', linewidth=1, annot=True, fmt='')

plt.title('Confusion Matrix', size=20)

plt.xlabel('Predicted Labels', size=14)

plt.ylabel('Actual Labels', size=14)

plt.show()

print(classification\_report(y\_test, y\_pred))

model.save("sound\_model.h5")

**CHAPTER 7**

**CONCLUSION**

In this project, we have reviewed various emotional SER methodologies and the associated speech databases and compared them from different aspects. This model directly learns the relevant representation from the speech signal in a data-driven manner and calculates the conditional probability for each phoneme class. In this, LSTMs as an acoustic model consists of a feature stage and classifier stage. Both the stages are trained jointly. Raw speech is supplied as input to first convolutional layer, and it is further processed by several convolutional layers. We tried to cover all the major deep learning techniques used for the task of SER, from LSTMs to attention mechanisms. One widespread limitation in almost all the related works examined was the fact that they were only reporting the accuracy of the recognition as their performance measure, but statistically, accuracy by itself is not a comprehensive measure of the performance of a system. The surge of new research on convolutional neural networks shows that they are capable of better solving the problem of emotion recognition by having higher low-level and short-term discriminative capabilities. The incorporation of LSTM networks and the introduction of deep convolutional LSTM structures has helped to take the solution to a new level and to give the network long-term memory to be able to identify long-term paralinguistic patterns. They have also shown higher capabilities of speaker-independent emotion recognition. Lastly, by the introduction of the attention methods, a new level of nonlinearity has been added to the classifiers that can, in turn, help in creating a more efficient system with fewer components.

**REFERENCES**

1. Hengshun Zhou, Jun Du, Yuanyuan Zhang, Qing Wang, Qing-Feng Liu, Chin-Hui Lee, 2021, “Information Fusion in Attention Networks Using Adaptive and Multi-Level Factorized Bilinear Pooling for Audio-Visual Emotion Recognition”, IEEE/ACM Transaction, Vol: 29, PP: 2617-1629.
2. Ying Zhou, Xuefeng Liang, Yu Gu, Yifei Yin, Longshan Yao, 2022, “Multi-Classifier Interactive Learning for Ambiguous Speech Emotion Recognition”, IEEE/ACM Transactions, Vol: 30, PP: 695-705.
3. Reem Hamed Aljuhani, Areej Alshutayri, Shahd Alahdal, 2021, “Arabic Speech Emotion Recognition from Saudi Dialect Corpus”, IEEE Access, Vol: 09, PP: 127081-127085.
4. Chenghao Zhang, Lei Xue, 2021, “Auto encoder With Emotion Embedding for Speech Emotion Recognition”, IEEE Access, Vol: 09, PP: 51231-51241.
5. Jia-Hao Hsu, Ming-Hsiang Su, Chung-Hsien Wu, Yi-Hsuan Chen, 2021, “Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations”, IEEE/ACM Transaction, Vol: 29, PP: 1675-1686.
6. Na Liu, Baofeng Zhang, Bin Liu, Jingang Shi, Lei Yang, Zhiwei Li, Junchao Zhu, 2021, “Transfer Subspace Learning for Unsupervised Cross-Corpus Speech Emotion Recognition”, IEEE Access, Vol: 09, PP: 95925-95937.
7. Mixiao Hou, Zheng Zhang, Qi Cao, David Zhang, Guangming Lu, 2021, “Multi-View Speech Emotion Recognition via Collective Relation Construction”, IEEE/ACM Transaction, Vol: 30, PP: 218-229.
8. Ting-Wei Sun, 2020, “End-to-End Speech Emotion Recognition with Gender Information”, IEEE Access, Vol: 08, PP: 152423-152438.
9. Yelin Kim, Emily Mower Provost, 2017, “ISLA: Temporal Segmentation and Labeling for Audio-Visual Emotion Recognition”, IEEE Transactions, Vol: 10, No: 10, PP: 196-208.
10. Srinivas Parthasarathy, Carlos Busso, 2020, “Semi-Supervised Speech Emotion Recognition with Ladder Networks”, IEEE/ACM Transactions, Vol: 28, PP: 2697-2709.
11. Reza Lotfian, Carlos Busso, 2019, “Curriculum Learning for Speech Emotion Recognition from Crowd sourced Labels”, IEEE/ACM Transactions, Vol: 27, No: 04, PP: 815-826.
12. Yi Lei, Shan Yang, Xinsheng Wang, 2022, “MsEmoTTS: Multi-Scale Emotion Transfer, Prediction, and Control for Emotional Speech Synthesis”, IEEE/ACM Transactions, Vol: 30, PP: 853-864.
13. Yue Xie, Ruiyu Liang, Zhenlin Liang, Chengwei Huang, Cairong Zou, Bjorn Schuller, 2019, “Speech Emotion Classification Using Attention-Based LSTM”, Vol: 27, No: 11, PP: 1675-1685.
14. Lu Yi, Man-Wai Mak, 2020, “Improving Speech Emotion Recognition With Adversarial Data Augmentation Network”, IEEE Transactions, Vol: 33, No: 01, PP: 172-184.
15. Norbert Braunschweiler, Rama Doddipatla, Simon Keizer, Svetlana Stoyanchev, 2022, “Factors in Emotion Recognition with Deep Learning Models Using Speech and Text on Multiple Corpora”, Vol: 29, PP: 722-726.