

DESIGNING AN EFFICIENT ALGORITHM FOR RECOGNITION OF HUMAN EMOTIONS THROUGH SPEECH

A thesis submitted in partial fulfilment of the requirements for the Degree of
DOCTOR OF PHILOSOPHY

By

Youddha Beer Singh
Enrolment No: E16SOE806

Under the supervision of
Dr. Shivani Goel



School of Computer Science Engineering and Technology,
BENNETT UNIVERSITY
(Established under UP Act No 24, 2016)
Plot Nos 8-11, Tech Zone II,
Greater Noida-201310, Uttar Pradesh, India.
September 2022

© Youddha Beer Singh, (2022)

Bennett University has the royalty-free permission to reproduce and distribute copies of this Thesis for teaching and research as well as for dissemination of Knowledge.

Thesis Evaluation

This is to certify that **Youddha Beer Singh** (Enrolment Number: **E16SOE806**) has successfully defended the Thesis entitled **Designing an efficient algorithm for recognition of human emotions through speech** on --/-- /2022.

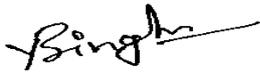
The committee recommends the candidate for the award of the degree of
Doctor of Philosophy.

Signature of External Examiner

Signature of Internal Examiner

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled “**Designing an efficient algorithm for recognition of human emotions through speech**” submitted at **School of Computer Science Engineering and Technology, Bennett University, Greater Noida, India**, is an authentic record of my work carried out under the supervision of **Dr. Shivani Goel**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. Theses.



Signature of the Scholar

Youddha Beer Singh
(E16SOE806)

School of Computer Science Engineering and Technology
Bennett University, Greater Noida, India

Date



CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled “**Designing an efficient algorithm for recognition of human emotions through speech**”, submitted by **Youddha Beer Singh** at School of Computer Science Engineering and Technology, **Bennett University, Greater Noida, India**, is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree or diploma.

(Signature of Supervisor)

Dr. Shivani Goel

Professor,

School of Computer Science Engineering and Technology

Bennett University, Greater Noida

Place: **Bennett University**

(Head of SCSET)

Date:

(Office Seal)

PREFACE AND ACKNOWLEDGEMENT

I will be eternally grateful to Dr. Shivani Goel for her tireless efforts in guiding and supporting me throughout my studies, for helping me in completing my research objectives, for showing me the right path when no path seemed possible, and for standing beside me in pivotal moments of my life.

I would like to express my gratitude to Dr. Deepak Garg, Professor, Dean School of Computer Science Engineering and Technology, Bennett University, Greater Noida, for his assistance and helpful ideas, and for answering my questions and providing required research facilities. I am also thankful to RAC members Dr. Anurag Goswami, and Dr. Neelam Choudhary for their helpful recommendations, assistance and helpful ideas.

There are no enough words to adequately convey my thanks to my parents (Father: Late Indra Jeet Singh; Mother: Basmati Singh) and family (Elder Brother: Dr. Ashish Singh; Wife: Shashi Singh) for their support, love, and encouragement that allowed me to complete my research work. Thank you to everyone who helped with my research, whether directly or indirectly. Finally, I want to thank my son Yojit and daughter Yati, whose time I stolen to finish my thesis work.



Youddha Beer Singh
(E16SOE806)

ABSTRACT

In today's world, human emotion recognition from speech is being utilized for many of the real-life applications such as healthcare, behaviour assessment, human and robot interaction, robot-robot interaction, and many more. The recognition of emotion from speech is still a challenging task which requires for the SER system such as the availability of suitable emotional databases, identification of the relevant feature vector, and suitable classifiers. The first challenge is the availability of a speech database of high quality which is critical for the performance of the machine learning algorithms to recognize emotions from a particular language. There is a lack of an emotional speech database in the Indian ascent. The second challenge is to identify the relevant feature vector to correctly classify the emotions with a low computational cost. The third challenge is to identify/modify/create new classifier for identification of emotions. Research says that in the recent years, there is a growing interest of researchers to use deep learning approaches for SER and get improvement in recognition rate. In this field, researchers focused on either hand-crafted classifiers or deep learning approaches to increase the recognition rate. The major challenge for hand-crafted classifiers is to identify the suitable feature vector. In this thesis, our contributions are: i) reduce the computation cost of the SER model and ii) improve the average accuracy of the SER model than the state-of-the-art, and iii) created Indian emotional speech database to overcome the above challenges. This research has critically analysed the literature on SER in terms of speech database, speech features, and deep learning approaches to investigate the current research work and is focused on designing an efficient algorithm for recognition of human emotions through speech to overcome the challenges and limitations in SER system. For that, a novel emotional speech database IESC (Indian Emotional Speech Corpora) is created, and efficient CNN based architectures have been proposed for SER. The IESC database created from eight north Indian speakers 5 males and 3 females in the English language with 600 samples. The created database IESC has been validated from more than twenty people by conducting a subjective test. The proposed models have been evaluated on IESC dataset and publicly available benchmark datasets namely the Italian emotional speech database (EMOVO), Berlin database of emotional speech (EMODB), Surrey audio-visual expressed emotion database (SAVEE), and Ryerson audio-visual database of emotion (RAVDESS) databases. The results of the experimentations show that the proposed model has out-performed the state-of-the-art SER models. The latest

SER model CNN- assisted is also implemented on IESC database. The average accuracy of the proposed model is found to be 95% for IESC which is better than CNN- Assisted model (88%). The average accuracy of the proposed model is found to be even better than the state-of-the-art SER approaches and the baseline CNN-based architectures ResNet-18 and ResNet-34.

TABLE OF CONTENTS

	Page Number
DECLARATION BY THE SCHOLAR	iv
CERTIFICATE	v
PREFACE AND ACKNOWLEDGEMENT	vi
ABSTRACT	vii
TABLE OF CONTENTS	ix
LIST OF FIGURES	xii
LIST OF TABLES	xiv
ACRONYMS/ABBREVIATIONS	xv
1. CHAPTER 1: PROBLEM DEFINITION	1-8
1.1 Introduction	1
1.2 Background and Motivation	3
1.3 Challenges and Research Gaps in Present Study	4
1.4 Problem Statement	5
1.5 Research objectives and work Flow	6
1.6 Organization of the thesis	8
2. CHAPTER 2: SPEECH DATABASES AND SPEECH FEATURES	9-21
2.1 Research Methodology for Literature Review	9
2.1.1 Search idea	9
2.1.2 Inclusion Criteria	10
2.1.3 Exclusion Criteria	10
2.1.4 Quality Assessment	10
2.2 Emotional Speech Database	10
2.2.1 Actor Based Emotional Speech Database	11
2.2.2 Induced/Elicited Emotional Speech Database	11
2.2.3 Natural Emotional Speech Database	11
2.3 Speech Features	16
2.3.1 Spectral Features	16
2.3.2 Prosodic Features	17

2.3.4	Teager Energy Operator (TEO) Features	17
2.3.5	Voice quality features	18
2.4	The Review's Findings	18
2.5	Conclusion	21
3	CHAPTER 3: SER CLASSIFIERS	22-36
3.1	Related Work for SER classifiers	22
3.1.1	Traditional Approaches	22
3.1.2	Deep Learning Approaches	24
3.2	Findings of the Review	34
3.3	Conclusion	36
4	CHAPTER 4: DEEP CNN BASED ALGORITHM FOR RECOGNITION OF EMOTION	37-48
4.1	Overview	37
4.2	Proposed algorithm	38
4.2.1	Pseudocode of the proposed algorithm	38
4.2.2	Details of DCNN architecture	39
4.2.3	Converting speech signal into spectrograms	41
4.3	Experimental details and results	42
4.3.1	Data sets	43
4.3.2	Experimental results	43
4.3.3	Performance comparison of the proposed algorithm with state-of-the-art	47
4.4	Conclusion	48
5	CHAPTER 5: LIGHTWEIGHT CNN BASED ALGORITHM WITH NEW INDIAN EMOTIONAL SPEECH CORPORA	49-62
5.1	Overview	49
5.2	Proposed Model	51
5.3	Experimental Details	54
5.3.1	Indian Emotional Speech Corpora (IESC)	54
5.3.2	Publicly Available Datasets	56
5.3.3	Implementation details and results	56
5.3.4	Performance comparison of the proposed algorithm	61
5.4	Conclusion	62

6	CHAPTER 6: 1D CNN BASED ALGORITHM USING MFCC FEATURES	63-67
6.1	Overview	63
6.2	Proposed Algorithm	63
6.3	Experimental Details	65
6.3.1	Data Set	65
6.3.2	Experimental Results	65
6.3.3	Performance Comparison of the Proposed Algorithm	67
6.4	Conclusion	67
7	CHAPTER 7: CONCLUSION AND FUTURE SCOPE	68-71
7.1	Conclusion	68
7.2	Future Scope	69
8	LIST OF PUBLICATIONS	72
	REFERENCES	73-89

LIST OF FIGURES

Figure Number	Caption of Figures	Page Number
1.1	Basic SER Architecture	2
1.2	Work Flow	7
2.1	Speech features categories	16
2.2	Distribution of paper according to the databases	19
3.1	The number of publications and their percentages based on the available resources	34
3.2	Distribution of the collected paper in terms of year-wise number of papers and resources	35
4.1	Steps of the proposed algorithm	39
4.2	DCNN architecture details	40
4.3	Spectrograms of each type of emotion for EMOVB, SAVEE, and EMOVO corpus	42
4.4	Learning rate variation for the number of iterations and variation in a loss concerning learning rate	45
5.1	Conversion of the speech signal to spectrogram	51
5.2	Proposed CNN-based model to recognize emotions from speech	53
5.3	Validation result of database IESC	56
5.4	Sample spectrograms of each emotion for EMOVO, EMOVB, SAVEE, and IESC speech corpus	57
5.5	Proposed Model and CNN-Assisted model training and testing accuracy plot concerning the number of epochs on IESC database	57
5.6	Performance comparison of the proposed model with CNN-Assisted on IESC database	58

5.7	Proposed Model training and testing accuracy variation concerning the number of epochs on EMOVO, EMOVB, and SAVEE databases	59
5.8 a)	Confusion matrix of ResNet-18 and ResNet-34 on database EMOVO	60
5.8 b)	Confusion matrix of ResNet-18 and ResNet-34 on database EMOVB	60
5.8 c)	Confusion matrix of ResNet-18 and ResNet-34 on database SAVEE	61
5.8 d)	Confusion matrix of ResNet-18 and ResNet-34 on database IESC	61
6.1	1D CNN Based Algorithm for SER	64
6.2	Training loss and validation loss variation of the model	66

LIST OF TABLES

Table Number	Caption of Tables	Page Number
2.1	The distribution according to the resource and the number of papers	9
2.2	An overview of Actor based, Induced, and Natural emotional speech database	12
2.3	Database of emotive speech literature review	13
2.4	The distribution of the number of papers and percentage according to the type of databases	18
2.5	The distribution of the databases according to the languages	19
2.6	Speech Features types their purpose and approaches	20
3.1	Literature Summary of the traditional approaches for SER	23
3.2	The deep learning algorithms utilised for SER are briefly described, including significant features and limitations.	24
3.3	A survey of the literature on deep learning techniques for SER	28
3.4	Deep learning techniques for SER motivations and limits	36
4.1	Pseudo code to implement the proposed algorithm	39
4.2	Main Parameters and their value	41
4.3	Steps to get spectrograms	41
4.4	Number of samples in speech corpora used	43
4.5 a)	Confusion matrix for emotions prediction on EMODB at stage-1	44
4.5 b)	Confusion matrix for emotions prediction on SAVEE at stage-1	44
4.5 c)	Confusion matrix for emotions prediction on EMOVO at stage-1	44
4.6 a)	Confusion matrix for emotions prediction on EMODB at stage-2	46
4.6 b)	Confusion matrix for emotions prediction on SAVEE at stage-2	46
4.6 c)	Confusion matrix for emotions prediction on EMOVO at stage-2	46
4.7	Accuracies at stage-1 and stage-2	47
4.8	Performance comparison of the proposed algorithm with state-of-the-art methods	47

5.1	Speaker details and statistics of IESC database with nationality as Indian	55
5.2	Statistics of Publicly available database EMOVO, EMODB, and SAVEE	56
5.3	Classification report of the proposed model and SER CNN-Assisted Model on IESC	58
5.4	The Experimental Results of the proposed Model	59
5.5	Performance of the proposed model on EMOVO, EMODB, and SAVEE databases	59
5.6	Comparison of the proposed model with the state-of-the-art SER approaches	61
6.1	Steps to implement the proposed model	64
6.2	Number of samples in each emotion	65
6.3	Experimental Results	66
6.4	Classification Report of the proposed model on RAVDESS data sets	66
6.5	Confusion matrix for emotion recognition on RAVDESS	66
6.6	Comparison of the proposed algorithm with state-of-the-art methods in term of average Accuracy	67

LIST OF ACRONYMS/ABBREVIATIONS

ABC	Airplane Behaviour Corpus
ACRNN	Attention based Convolutional Recurrent Neural Network
AFEW	Acted Facial Expressions in the Wild
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
AVEC	Audio/Visual Emotion Challenge
BPNN	Back Propagation Neural Network
CASEC	Chinese Academy of Science Emotional Corpus
CASIA	Chinese Academy of Sciences Institute of Automation
CHEAVD	Chinese Natural Emotional Audio-Visual Database
CNN	Convolutional Neural Network
DBN	Deep Belief Networks
DCNN	Deep Convolutional Neural Network
DECN	Deep Ensemble Capsule Network
DEMoS	Database of Elicited Mood in Speech
DFFN	Deep Feed Forward Network
DNN	Deep Neural Network
DSCNN	Dependency Sensitive Convolution Neural Network
DWT	Discrete Wavelet Transform
EDFLM	Emotion Discriminative and domain-invariant Feature Learning Method
eGEMAPS	Extended Geneva minimalistic acoustic feature set
EHSD	Emotional Hindi Speech Database
ELM	Extreme Learning Machine
EMODB	Berlin Database of Emotional Speech
GFCC	Gammatone frequency cepstral coefficient
GFNN	Generalized Feedforward Neural Network
GMM	Gaussian Mixture Model
HHTC	Hilbert–Huang Transform Combination

HMM	Hidden Markov Model
IEMOCAP	Interactive Emotional Dyadic Motion Capture Database
IIIT-H TEMD	International Institute of Information Technology-Hyderabad Telugu Emotional Database
IITKGP-SESC	Indian Institute of Technology Kharagpur Simulated Emotional Speech Corpus
JAVED	Japanese Audio-Visual Emotion Database
KNN	K-Nearest Neighbor
KVDERW	Korean Video Dataset for Emotion Recognition in the Wild
KSUE	King Saud University Emotions
LDC	Linguistic Data Consortium
LFPC	Log Frequency Power coefficients
LPCC	Linear Prediction Cepstral Coefficients
LSTM	Long Short Term Memory
MELD	Multimodal Emotion Lines Dataset
MFB	Mel-frequency Filter Bank
MFCC	Mel-frequency Cepstral Coefficient
MLP	Multilayer Perceptron
PLP	Perceptual Linear Prediction
RAVDESS	The Ryerson Audio-Visual Database of Emotional Speech and Song
RBM	Restricted Boltzmann Machine
RECOLA	Remote COLlaborative and Affective
RNN	Recurrent Neural Network
QDA	Quadratic Discriminant Analysis
SAVEE	Surrey Audio-Visual Expressed Emotion
SLR	Systematic Literature Review
SUSAS	Speech Under Simulated and Actual Stress
SVM	Support Vector Machine
SER	Speech Emotion Recognition
TEO	Teager Energy operator features
TDNN	Time Delay Neural Network
TESS	Toronto Emotional Speech Set

TEVD
VAM

The Emotional Voices Database
Vera a Mittag

CHAPTER 1

PROBLEM DEFINITION

1.1. Introduction

Human emotion detection is a crucial aspect of computer-human interaction and one of the most recent challenges. Human emotions play a vital role in real-life communication. Gestures, facial expressions, bodily postures, and vocal communication may all be used to identify human emotions. Temperature, heart rate, blood pressure, muscular activity, and skin resistance are among the physical qualities utilised to recognise human emotions [1]. Speech communication may also be used to identify human emotion. A key research topic in human emotion detection from speech is to determine the speaker's emotional state from a given speech signal. In this study, human emotions recognised through the speech as speech data contains more information for the identification of emotions. In today's world, detecting human emotions from speech is a crucial topic of study. Many studies on human emotion identification concentrate on separating emotions from audio or visual communication. Speech Emotion Recognition (SER) is a topic that is gaining popularity. In recent years, this SER technology has benefitted a wide range of applications. A tutoring system used in distant learning that can identify bored or uninterested users and change the style and level of study material offered is an example of such a system. Human emotional states must be recognised, responded to, and analysed in systems such as auto driving systems, contact centre conversation analysis, robots, call analysis in emergency services such as fire and ambulance, speech to voice translation, and many other things.

There are three important factors for emotion recognition from speech: “what is said” i.e. contents, “how it is said” i.e. the way/style of speaking, and who is saying it (male or female). Without language information, a standard SER system recognises emotional states from voice signals [2]. The researchers' initial task is to find the necessary information (speech characteristics) that may be retrieved from speech signals to detect emotional states [3,4]. The second problem is to find appropriate classifiers to classify the emotions [5]. In the recent decade, deep learning has emerged as a novel and appealing machine learning technique that has been studied by a variety of academics on a variety of themes [6]. Deep learning methods are generally the result of numerous layers of linear and non-linear processes added to neural

networks. Deep learning methodologies are generally used to improve the ability of computers so that computers can grasp what humans can accomplish, which includes emotional identification from speech.

There are three important aspects in the critical examination of the literature review on SER: speech emotional database, speech characteristics, and classifiers. From 2000 through 2021, this study examines the literature on Speech emotional databases, speech characteristics, and classical and deep learning-based classifiers. Even though many publications have given a review of previous research in SER, it aims to provide analytical information on the existing literature and SER outcomes throughout time. Extracting hand-crafted features via traditional / deep learning methodologies, automatically learning the features of speech based on deep learning approaches, and extracting features from spectrograms and then applying deep learning approaches are all reported in the literature for SER. There are also some hybrid ways. In Figure 1.1, the basic processes in SER are shown.

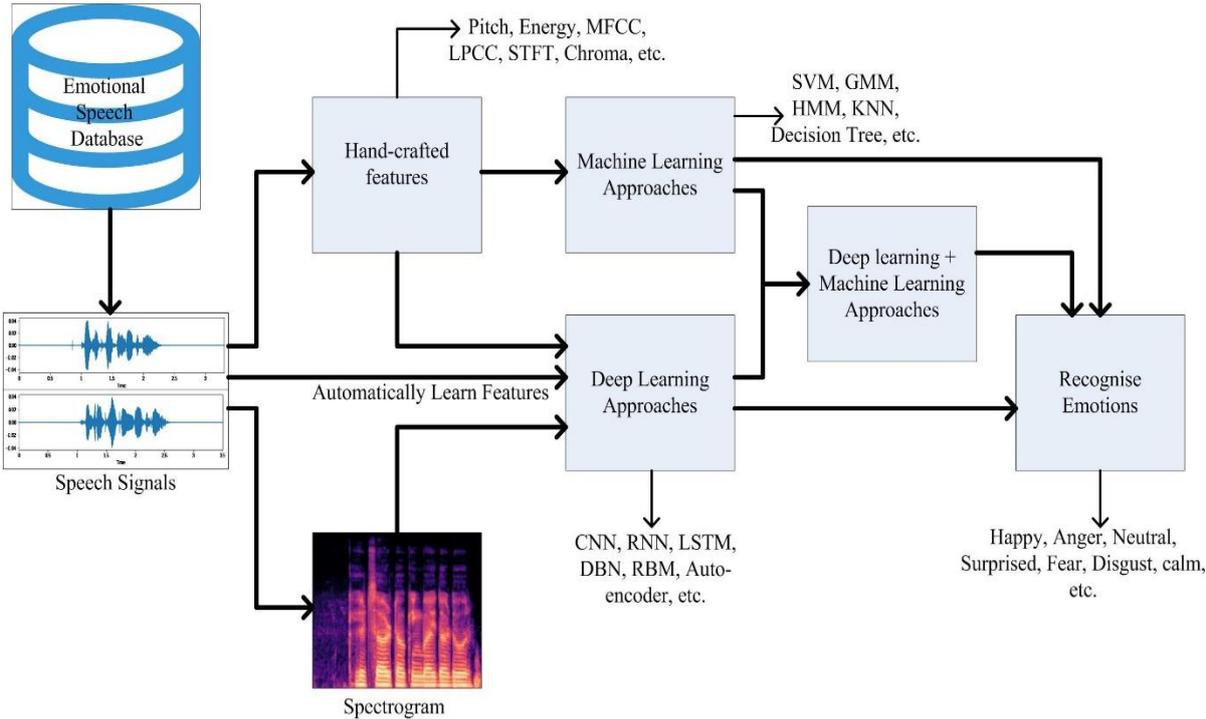


Figure 1.1. Basic processes in SER

The SER's basic processes shown in Figure 1.1 explain the three types of SER. The first category involves extracting hand-crafted characteristics from voice data before using machine/deep learning methods. Deep learning techniques in the second category learn features from voice signals at multiple levels of computations automatically. Speech signals are

transformed into spectrograms and deep learning algorithms are used in the third category. In this study third category has been considered in which speech signals are transformed into spectrograms and novel CNN based architectures are proposed to recognise the emotions.

1.2. Background and Motivation

Emotion recognition from speech is a widespread human instinct that has been investigated for more than 70 years by academics from several fields. Emotional intelligence is the capacity to identify, analyse, and express emotions of the speaker. It is the foundation of human communication. Emotions, behaviour, and ideas are continually interacting in such a manner that they impact each other. Emotions are tightly linked to decision making, according to psychologists and neuroscientists. So, emotions play a fundamental part in our logical behaviours. Emotion research, on the other hand, is highly difficult in various ways. In human social relationships, understanding emotions is critical. According to studies, only 10% of human existence is absolutely emotionless, while the rest is effectively colored by emotions. Despite the fact that emotional signals have been researched since the 1950s, they have made significant progress in recent years. This is mostly owing to new application innovations in the areas of human-machine interfaces, human-robot interfaces, and multimedia retrieval.

Because of the objective of automatically identifying and comprehending emotions, interface design between people and computers is gaining traction. Automatic emotion recognition will allow people and computers to connect in a more natural way. Because computers and computer-based applications are so firmly embedded in our daily lives, building computers and personal robots with emotion recognition skills is critical. Because of our ability to infer the emotional states of others based on their emotional states, the interaction process among humans ensures convergence and optimization. As a result, we may adapt our replies and behavioral patterns accordingly, making communication more meaningful and successful. When compared to human skills, a computer's performance and the emotional categories it can cover are severely constrained. One major challenge stem from a lack of comprehensive knowledge of emotions in human minds, as well as a lack of consensus among psychology researchers.

The following points served as motivation for us to work in emotion recognition from speech

- Emotion recognition from speech was created with the goal of increasing the flexibility of human-computer interaction and creating more intelligent robots and more responsible human-robot interactions are all benefits of this research.
- Our SER system should be able to distinguish various emotional patterns quickly and efficiently since it is a difficult challenge for humans.
- There is currently no perfect mathematical/statistical model that can accurately represent the enormous differences in emotional qualities that change from speaker to speaker and with different emotional utterances.
- Because of their computational complexity, most SER model are difficult to apply in real-time application systems. As a result, in this work, an effort is made to build a Deep learning based model for SER that overcomes the limitations.
- Today's SER systems mostly use classic spectral features, prosodic features and their combinations. Because of the failure in transitory processing of the emotional speech signals, the spectral and prosodic features are also losing information. The spectrograms contain more information than the speech signal. As a result, in this study, an effort is made to convert the speech signal to the spectrograms and then apply the Deep learning based SER models which automatically learn the relevant features.

1.3. Challenges and Research Gaps in Present Study

There is no universally accepted theory of emotion. Emotion recognition from speech is a challenging task because of the following reasons. To begin with, it is unclear than which speech characteristics are most useful in discriminating distinct emotions. Because these attributes directly impact most of the extracted speech characteristics, the acoustic variability generated by the existence of diverse sentences, speaking styles, speakers, and speaking speeds add hurdles. Another challenging issue is that the way a person expresses an emotion is mostly determined by the speaker, his or her culture, and the environment emotion categorization, assuming that there are no cultural differences among speakers. People, on the other hand, are aware of their emotions when they experience them. As a result, researchers were able to investigate and describe several areas of the study and analysis of speech emotion recognition utilising spectral features, prosodic features, and hybrid features.

Emotion is a unique mental state that occurs spontaneously rather than via deliberate effort, making it difficult to characterise objectively. As a result, there is no universally accepted

objective definition or consensus on the term emotion. There are no standard speech corpora available for comparing the performance of research methodology for recognising emotions across languages and cultures. Majority of the emotional speech databases, are English language databases followed by Chinese and German language databases. Very few databases are collected in Indian languages such as Hindi and Telugu. The majority of the databases contain emotions like happy, anger, sad and neutral. In contrast most of the databases rarely contain uncommon emotions such as approval, attention, antipathy, prohibition, etc. Some databases have collected emotional speech signals from TV shows, movies and then annotated the emotions by experts.

From the literature review of the SER approaches following research gaps are identified.

- The quality of speech database is one of the major concerns in case of human emotion recognition from speech. There are no measures defined for quality of speech database. The emotional speech database in the Indian ascent for the English language is lacking for public access.
- Gathering and maintaining large collections of speech data is one thing, but extracting useful feature and detecting emotions from them is even more challenging. There is no standard for identifying the useful features and feature vector size.
- Machine learning algorithms are based on statistical algorithms that are used for analysing big volume of data sources (image, sound, video, social network, structured database, etc.) in real time. Various classifiers used in machine learning till dates are not able to achieve near 100% accuracy in recognizing emotions from speech. There is a need to focus on the development of fast and more accurate algorithm to recognize emotions from the speech.

1.4. Problem Statement

The majority of SER techniques, whether traditional approaches utilising handcrafted or deep learning approaches, focus on the recognition rate. It is significantly more difficult to extract usable features and detect emotions from them, since there is no standard for selecting important features or feature vector size. The development of a rapid and more accurate algorithm to discern emotions from speech remains a research issue, and the establishment of an emotional speech database in India is missing for public access. As a result, developing a SER framework

for emotion identification from speech and creating emotional speech databases in Indian language is strongly encouraged. As a result, the research " **Designing an efficient algorithm for recognition of human emotions through speech** " has been proposed to realise the necessity and relevance of the SER. To solve this issue, we created the SER framework, which is based on a number of recent research papers and can accurately predict emotions from speech using CNN-based techniques.

1.5. Research objectives and work Flow

Several techniques and algorithms have been proposed for human emotion recognition. Each one of these has certain pros and cons and give different outputs with different databases and features. The main objective of this work is to design an efficient SER approach. The objectives of the proposed research are as follows:

RO 1: To study existing techniques and algorithms for recognition of emotions from speech.

RO 2: To propose an efficient algorithm for recognition of emotions from speech.

RO 3: To implement the proposed algorithm.

RO 4: Verification and validation of the proposed algorithm through case studies.

The work Flow process depicted below in Fig 1.2 has been set to design and develop the generic framework for the proposed study. The research design process has many phases. We do our best effort and in dept study to achieve the outcome of each phase and for the completion of this thesis. To achieve the above mentioned research objectives following work flow mentioned in the Figure 1.2 has been adopted.

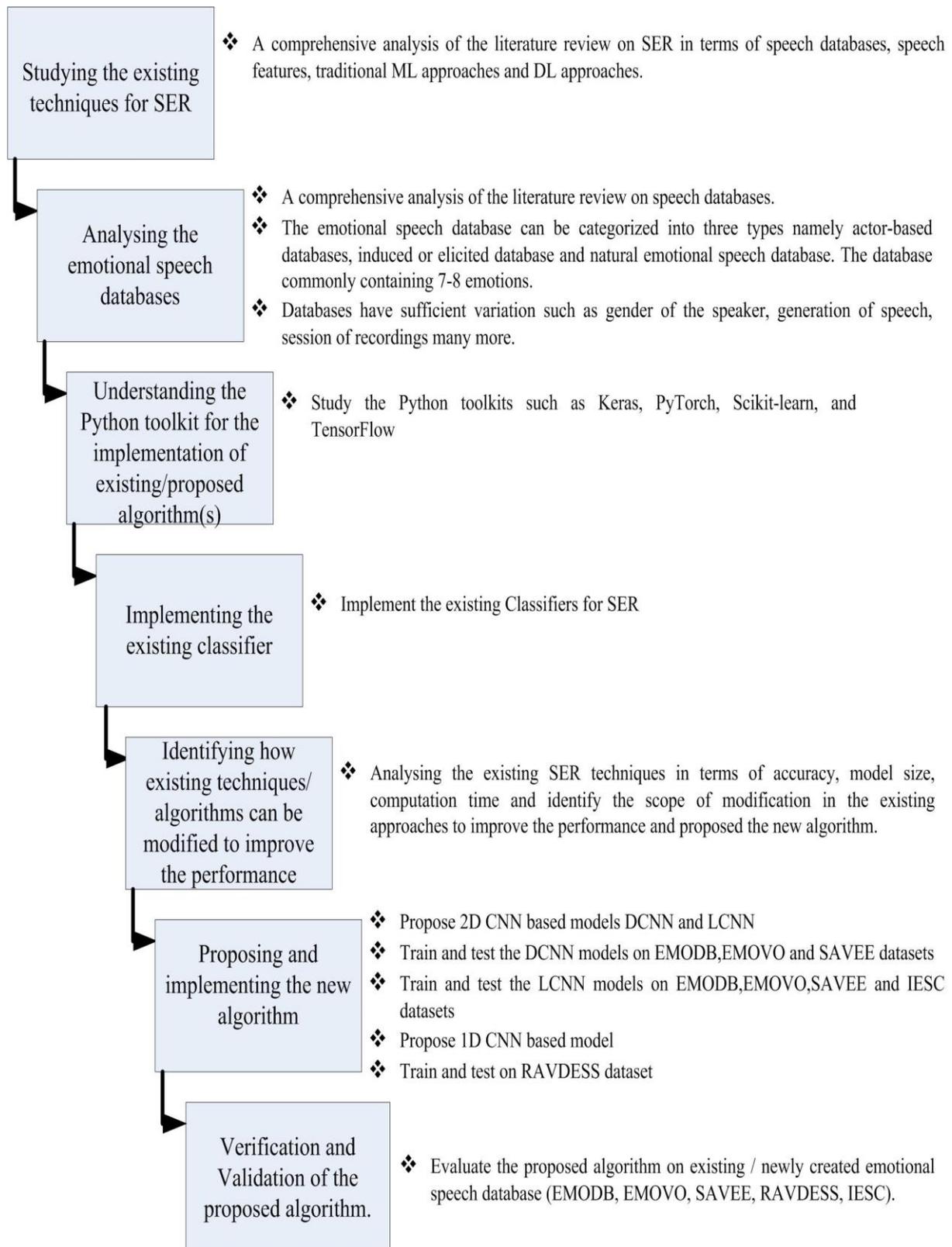


Figure 1.2 Work Flow

1.6. Organization of the thesis

Organization of the thesis are discussed in this section. In chapters 2 and 3 the detailed analysis of the related work on SER is discussed. Related work is discussed in terms of emotional speech database, speech features, and classifiers used for SER and the findings of the related work are summarized. Detailed study of emotional speech databases and speech features are presented in Chapter-2. In chapter 3 SER classifiers such as traditional approaches and DL approaches are critically reviewed. Chapter 4 includes the implementation, experimentation, and validation of the DCNN based proposed model. In Chapter 5, we discussed the new created database IESC and proposed model based on LCNN. 1D CNN Based Approach using MFCC Features is discussed in Chapter 6. In Chapter 7, we summarized results, contribution of the thesis, and future scope

CHAPTER 2

SPEECH DATABASES AND SPEECH FEATURES

2.1. Research Methodology for Literature Review

The main objective of this work is to identify the trending research work in the SER domain in terms of emotional speech databases, and speech features. The methodology is divided into different phases. In the first phase the search idea is described to find the related research papers. The search idea used in this paper is described in terms of search strings and resources from where research papers are collected.

2.1.1 Search idea

The search idea used for the literature review is explained in this phase in terms of search strings and resources. Search strings aim to capture all relevant research papers in the field of emotional speech database and speech features. Search strings were identified using the population, intervention, comparison, and outcome-based criteria. The selected search strings are documented below:

- Emotional speech database or Emotional speech corpus
- Speech features for emotion recognition from speech

The research papers were selected from different libraries such as Science Direct, Springer, IEEE, Google Scholar and others. The distribution of the selected research papers according to the resources is given in Table 2.1.

Table 2.1: The distribution according to the resource and the number of papers

Digital Database libraries	Selected Search String	
	Emotional database/Emotional speech corpus	speech Speech features for emotion recognition from speech
Elsevier	1	5
Springer	6	7
IEEE	6	4
Google Scholar	2	-
Others	10	4

The literature review presented here is inspired by the SLR presented by Kitchenham and Charters [7]. To be included in the study all article were examined using the criteria specified in inclusion, exclusion, and quality evaluation.

2.1.2. Inclusion Criteria

The inclusion criteria were developed to assess selected publications in a methodical manner, and only those papers that satisfied the following criteria were chosen:

IC 1: The information in the article is well-researched.

IC 2: The focus of the paper is on SER and the speech emotional database.

2.1.3. Exclusion Criteria

To exclude irrelevant items from the list, we set some exclusion criteria. The article was dismissed if it matched at least one of the following criteria:

EC 1: The article was written in a non-English language.

EC 3: Another database has already indexed the article.

2.1.4. Quality Assessment

All of the articles that were picked were judged on their quality. Each paper was evaluated using the following checklist to determine the research's reliability and soundness:

QA 1: Are the research goals well defined?

QA 2: Is there any evidence that the work has been referenced by other authors?

QA 3: Is the study's conclusion credible and evidence-based?

The article was included in the research because it fulfilled the inclusion, exclusion, and quality evaluation criteria.

2.2. Emotional Speech Database

The emotional databases quality is crucial to SER, and having a proper emotional database is a need [7]. There are several things to consider while evaluating the validity of an emotional database's architecture. General research topics such as the scope of the database, annotation of the emotions, database size, and type of speakers in terms of age and gender should all be addressed when recording the emotional speech database. Actor-based, elicited, and natural emotional speech databases are the three categories of emotional speech databases.

2.2.1. Actor Based Emotional Speech Database

Simulated emotional speech databases are another term for actor-based emotional speech databases. These databases are developed by skilled professional actors, such as radio and theatre performers, or by someone who can talk in a variety of moods. The speaker, who is reciting the same material in different emotions, is recorded. To account for the variation of the physical speech and expressiveness of the human individual, recording may have been done in many sessions such as in the morning, afternoon, evening, and night. This is one of the most reliable methods for capturing the emotional speech database throughout the whole emotional spectrum. A simulated database makes for more than 60% of the emotional speech database. In general, the simulated emotional database is more expressive than the real-life database. [8].

2.2.3. Induced/Elicited Emotional Speech Database

Because it is closer to the natural database, the induced emotional speech database is also known as a semi-natural emotional database. These databases are created without the actor's awareness in a manufactured emotional environment. Following the creation of the fictitious emotional circumstance, the actor engaged in emotional conversion with the speaker. This is a more natural database than the simulated database. However, if the speaker is aware that the recording is not expressive, this is a manufactured emotional scenario. These sorts of databases are sometimes captured via vocal interactions with the computer, which may be managed by anybody without the subject's awareness [9].

2.2.4. Natural Emotional Speech Database

Natural emotional databases contain genuine data and recognising emotions might be challenging at times. This sort of database may be created from a customer service chat, a television broadcast, a doctor-patient conversation, a courtroom, a cockpit recording in an unusual scenario, and many more sources. It's tough to feel a whole range of emotions in these settings, and there are also some copyright and security issues [9]. Table 2.2 summarises all sorts of emotional databases: actor-based, induced, and natural.

Table 2.2: An overview of Actor based, Induced, and Natural emotional speech database

Type of database	Description	Advantages	Disadvantages	Preferred Applications
Actor based/ Simulated	These datasets are gathered from trained actors who speak the same topic in different emotions.	<ul style="list-style-type: none"> • A large number of databases available • Commonly used database • All types of emotions are available • Standardized • Available in most the language • Results can be easily compared 	<ul style="list-style-type: none"> • Episodic in nature • Not true in real-world environments 	<ul style="list-style-type: none"> • SER • Expressive speech synthesis system
Elicited/ Induced	The database is produced under an artificial emotional setting without the actor's knowledge. It is also known as semi-natural.	<ul style="list-style-type: none"> • Contextual information is present but still artificial • Near to natural database • More emotions are available than the natural database 	<ul style="list-style-type: none"> • Fewer number of emotions are available than the actor based database • If actors know that they are being recorded they it will be like an actor-based database 	<ul style="list-style-type: none"> • SER • Categorical emotion analysis • Dimensional emotion analysis
Natural	These databases are compiled from real-world data such as contact centre discussions, patient-doctor interactions, and so on.	<ul style="list-style-type: none"> • Completely natural database • Useful for real-world modeling emotion system 	<ul style="list-style-type: none"> • Overlapping of utterances • All type of emotions is not available • Presence of noise in the background • Privacy and copyright issues arise • Difficult to model 	<ul style="list-style-type: none"> • SER • Categorical emotion analysis • Dimensional emotion analysis

The critical examination of the emotional database is offered in the literature review based on language, database type, number of emotions, kind of data (Video/Audio), number of speakers,

number of samples, and number of utterances, as well as a brief description. Table 2.3 summarises the findings of the emotional speech databases literature review.

Table 2.3: Database of emotive speech literature review

Ref.	Brief Description
Emotional Speech databases in Indian Languages	
EHSD [10]	This is an actor-based database with 28 (14 males and 14 female) speakers recording in Hindi. There are a total of 6048 samples and 12 utterances in this collection. The database's objective is to support speech recognition, speech synthesis, and speaker recognition, as well as emotions such as happiness, anger, sadness, neutrality, surprise, and sarcasm. For research reasons, it is open to the public.
IIT-H-TEMD [11]	38 people have contributed to this database in Telgu (20 males and 18 female). It's a semi-natural emotional database that has been gathered in audio files, with a total of 2450 samples. Trained and untrained actors gathered data from the natural environment to establish a database. Happy, anger, sadness, neutral, astonished, fear, contempt, sarcastic, annoyed, calm, frightened, bashful, excited, and yell are among the emotions mentioned. This information is available to the general public for the research.
IITKGP-SESC [12]	This is a database of actors that has been compiled by ten people (5 males and 5 female). There are a total of 12000 samples and 15 utterances in this study. The data was obtained from professional actors in Telugu and covered emotions such as happiness, anger, neutrality, surprise, fear, contempt, sarcasm, and compassion. For research reasons, it is open to the public.
Database of passionate speech in English	
RAVDESS [13]	There are 7256 samples and 2 utterances uttered by 24 people (12 Males and 12 Female). This database is audio-video captured in the American English language. The database is provided to the public in the form of speeches and songs. It is simply audible, visual, or audio-visual, and the emotions are joyful, angry, neutral, astonished, fear, disgust, and calm. It can also be purchased for business use.
SAVEE [14]	This database, which contains 480 samples and 15 utterances, is contributed to by four male speakers. This database is available in audio-visual format in English. The database is listed as a need for SER. It is free to use for research purposes. Anger, happiness, neutral, surprise, fear, disgust, and sadness are all included.
eNTERFA CE [15]	It is an audio-visual database. There are 1287 samples and 42 utterances. The purpose of the database is audio-visual speech recognition and SER. It is publically available for research purposes and emotions included are anger, happiness, surprised, fear, disgust, and sadness.

TESS [16]	There are 2800 audio samples in this database. It is a fictitious database created by two actresses, ages 26 and 64. Anger, happiness, surprise, fear, disgust, sadness, neutral, and pleasant emotions are all included.
MELD [17]	This database was recorded in audio and video format. Friends, TV shows with diverse speakers, and 1433 discussions were utilised to collect data, with good, negative, and neutral feelings indicated. It is accessible to the general public.
IEMOCA P [18]	This database is an audio-visual representation of a semi-natural emotional database. It is recorded by ten speakers (5 males and 5 females). Anger, happiness, sadness, neutral, and dimensions level activation, valence, and dominance are all annotated emotions.
Chinese Language Emotional Database	
HEU Emotion [19]	The initial portion of the database was compiled using Tumblr, Google, and gypsy, while the second half was compiled using movies, TV programmes, and variety shows. Anger, boredom, disappointment, confusion, fear, disgust, glad, neutral, sad, and surprise are among the 10 emotions. There are 1900 audio-visual examples in this collection, which is a semi-natural emotional database.
CHEAVD [20]	This collection comprises 140 minutes of emotive statements from movies, television shows, and talk shows. It gives you emotional levels as well as suppressed/fake emotional labels. There are 26 emotions and speakers in all, ranging in age from kid to older. This is an audio-visual representation of a natural emotional database.
CASIA [21]	A database comprising 12000 audio-visual samples. This collection includes 219 speakers from 25 films, 1 television programme, and 3 talk shows. And the total number of emotions is 24.
German Language Emotional Database	
VAM Corpus [22]	The database comprises 12 hours of audio-visual recordings from the conversation programme "Vera a Mittag." Activation and dominance are three emotional valences. This is a real-life emotional database. Which is accessible to the scientific community.
EMODB [23]	The database was built by ten persons using ten utterances in German, five short and five long. Anger, happiness, sorrow, neutral, disgust, boredom, and fear are examples of emotions. Which is made available to the general population. There are 800 audio samples in all.
Italian Language Emotional Database	
DEMoS [24]	DEMoS database encompassing 9365 samples of anger, sadness, happiness, fear, contempt, surprise, guilt, and natural emotion was also included. This database is a semi-natural emotional database captured in audio files by 68 speakers (45 males and 23 females).
EMOVO [25]	This collection of database contains 588 samples and 14 utterances in Italian recorded by 6 speakers (3 men and 3 females). The first Italian emotional corpus is now publicly available. There are seven emotions in this database: disgust, delight, fear, surprise, sad, joy, and neutral.
Japanese Language Emotional Database	

JAVED [26]	This database was captured in audio-visual format by 24 male Japanese speakers over the course of 100 minutes. The database was compiled using scripted speech and monologue and contains fine emotions such as happiness, rage, sadness, neutrality, and satisfaction.
Korean language Emotional Database	
KVDERW [27]	It comprises 1246 Korean language samples that were collected from movies. This collection contains emotive video clips from real-life situations. All video snippets were taken from Korean movies and captioned to represent seven emotions: happiness, anger, sadness, neutral, surprise, disgust, and terror.
Arabic Language Emotional Database	
KSUE corpus [28]	This is the first publicly accessible Arabic emotional speech corpus. There are 3280 samples in all, captured by 23 speakers in one audio file. It includes the moving speeches of 23 people from Saudi Arabia, Syria, and Yemen. Emotions such as neutral, joyful, sad, surprise, and rage are included.
Multilingual Emotional database	
TEVD [29]	The database was developed in audio form by 5 speakers in English and French. There are 7590 samples in all. The database is open source and intended for synthesis, creation, and SER. Neutral, Amused, Anger, Sleepy, and Disgust are the annotated emotions.
EmoFilm [30]	This collection was built by extracting lines from 43 films in English, Spanish, and Italian. Anger, sorrow, pleasure, fear, and contempt are the emotions annotated in a database constructed from 43 films in three languages. Access is restricted.

Databases are classified into three groups in the literature. There are three types of emotional speech databases: actor/simulated emotional speech databases, elicited/induced/semi-natural emotional speech databases, and natural emotional speech databases. Table 2 summarises the characteristics of these databases. The Table shows that there are several variations across the databases, including the quantity of emotions, language, technique, and database goal. Among all databases, actor-based emotional databases account for 66% of all databases. English language databases are the most numerous among the emotional speech databases shown in Table 2.2, followed by Chinese and German language databases. In Indian languages like Hindi and Telugu, there are very few databases. Emotions such as joyful, angry, sad, and neutral may be found in most datasets. In contrast, unusual emotions like acceptance, attentiveness, antagonism, prohibition, and so on are rarely found in most databases. Experts annotated the emotions in several datasets that collected emotional speech signals from TV episodes and movies. Typically, the database contains 7-8 emotions. Databases include enough variance, such as the speaker's gender, speech generation, recording session, and so on.

2.3. Speech Features

Speech characteristics are the most essential aspect of the SER. For SER, the researchers looked at and used a variety of characteristics. There is currently no standard approach for extracting speech characteristics and specific classifiers. Discriminant information was introduced in [31] to safeguard local information and improve outcomes. The speech characteristics retrieved for SER are determined by our requirements. The voice signal can be used to extract either local or global information, or both. [32] presented a unique feature set HHTC for video emotion identification. A novel feature selection technique based on fisher and correlation analysis was introduced in [33], and the results were improved. The temporal dynamics indicate the local characteristics. Short-term segmental characteristics is another name for it. Global characteristics, on the other hand, are represented by comprehensive statistics such as minimum and maximum values, mean, and standard deviation. It's also known as supra-segmental features or long-term characteristics. Spectral characteristics, prosodic features, Teager Energy operator, and voice quality features are examined in four categories for SER [34]. In Figure 2.1, the categories of local and global properties are displayed.

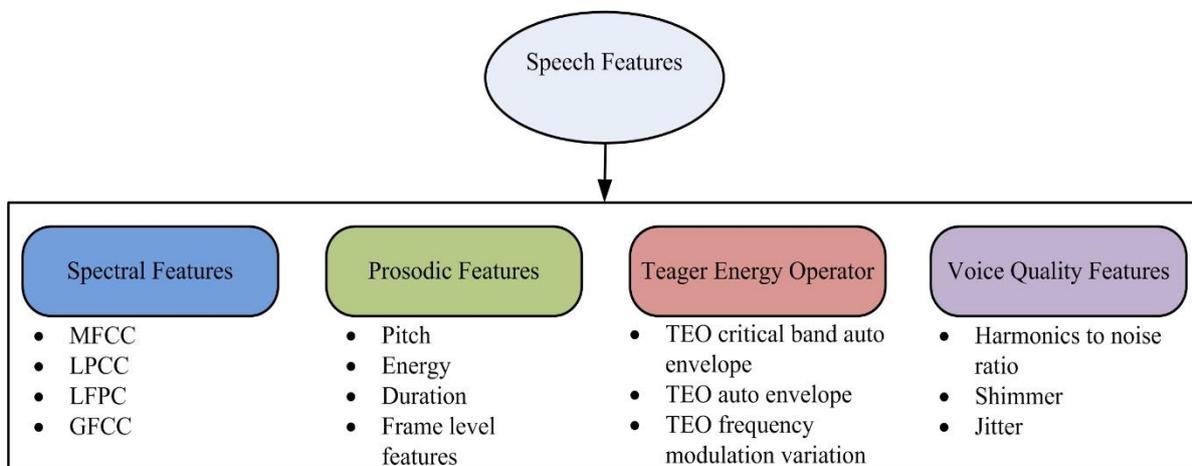


Figure 2.1: Speech features categories

2.3.1. Spectral Features

When any of the individual speakers produces a speech signal, it filters and regulates the vocal tract. The properties of the vocal tract are used to create spectral features, which are then worthily represented in the frequency domain [35]. The Fourier transform can be used to extract spectral characteristics. The frequency domain is transformed into the time domain using the Fourier transform. Mel frequency cepstral coefficient is the most advanced spectral

characteristic (MFCC). The supplied equation 2.1 is used to convert frequency to Mel frequency.

$$m = 2595 \log \left(\frac{f}{700} + 1 \right) \quad (2.1)$$

Before being converted into the frequency domain, the voice signals are separated into frames. The inverse Fourier transform is used to extract MFCC characteristics from voice samples [36]. The linear prediction cepstral coefficients are the second significant spectral characteristic (LPCC). The vocal tract, which represents crucial emotional information, includes it [37]. Gammatone frequency cepstral coefficient (GFCC) spectral characteristics were estimated in the same way as the MFCC [38]. In Mandarin [39], [40], a feature vector combination of LPCCs, MFCCs, PLP, and LFPCs was suggested for SER. LFPCs outperformed the LPCC standard features by a little margin [41].

2.3.2. Prosodic Features

Prosodic characteristics such as pitch, length, and vigour are thought to be effective predictors of mood [42-44]. Maximum, minimum, comparable pitch components, variance, range, mean, and standard deviation are all used as good prosodic information sources for emotion recognition [45]. Segment-level extraction prosodic elements were classified by Koolagudi and Rao [46] into three categories: pitch, intensity, and intonation. Prosodic characteristics are influenced by the vocal fold's air pressure. The statistical value of prosodic elements contains emotional-specific information that may be used to identify the emotion [47]. Meaning, minimum, maximum, range, standard deviation, skewness, median, slope, kurtosis, and many more statistical variables of pitch are available. Prosodic characteristics such as intensity, pitch, and length are critical for emotion perception [48]. Prosodic characteristics are a group of traits that people can detect, such as intonation and rhythm. These features may be found in words, phrases, syllables, and expressions [49]. Long-term characteristics collected from speech are prosodic features. Prosodic characteristics are based on fundamental frequency, duration, and energy aspects [34].

2.3.3. Teager Energy Operator (TEO) Features

The TEO function is introduced by Teager and Kaiser [43, 44]. TEO has included confirmation that the hearing process is in charge of energy detection. It has been observed that under a stressful scenario, the crucial band and frequency shift according to the distribution of

harmonics without changing the airflow during speech creation under stressful conditions. TEO is generated from a voice signal using the non-linear equation Eq (2.2).

$$\varphi[f(n)] = f^2(n) - f(n - 1)f(n + 1) \quad (2.2)$$

TEO is the time of arrival, and $f(n)$ is the speech signal. The three TEO-based novel features are TEO critical band auto envelope, normalised TEO auto envelope, and decomposed TEO frequency modulation variation [45].

2.3.4. Voice quality features

Individual vocal characteristics have been identified as voice quality attributes. Many speech processing tasks, such as speaker identification, emotion detection, and others, rely on voice quality attributes. Apart from spectrum characteristics, voice quality characteristics characterise the glottal source's quality. Voice quality elements include things like format frequency, bandwidth, glottal parameter, harmonic to noise ratio, jitter, and shimmer. Interrelationships between voice quality, discourse, and emotional content differ [50].

2.4. The Review's Findings

The purpose of this literature review is to examine particular research topics and current SER investigations. This section discusses all of the research findings. In terms of stated research topics, the findings of particular research concerns and current investigations are addressed below.

(a) Different types of databases used for SER: According to the research on speech databases, the databases are classified into three types: actor-based, semi-natural, and natural emotions based. Over 66.67 percent of all databases are actor-based emotional databases, 19.05 percent are semi-natural emotional databases, and 14.29 percent are natural databases. Table 2.4 shows the distribution of the paper and the percentage of each of the three types of databases.

Table 2.4: The distribution of the number of papers and percentage according to the type of databases

Type of database	Number of papers	Percentage
Actor/Simulated emotional database	[10], [12], [13], [14], [15], [16], [17], [21], [23], [25], [26], [27], [28], [29], [30]	66.67%
Induced/Semi-natural emotional database.	[11], [18], [19], [24]	19.05%
Natural emotional database.	[20], [22], [27]	14.29%

(b) Different languages in which emotional databases are available: The goal of this research is to examine the databases that are available in various languages. The emotional speech datasets gathered in Table 2.5 are in English, followed by Chinese and German language databases. There are very few databases available in Japanese, Korean, Indian, and multilingual languages. Only Hindi and Telugu are available in Indian language databases. Table 2.5 shows how the databases are organised according to language.

Table 2.5: The distribution of the databases according to the languages

Languages	Number of papers	Percentage
Indian Languages emotional speech database	[10], [11], [12]	14.28%
English Languages emotional speech database	[13], [14], [15], [16], [17], [18]	28.57%
Chinese Language Emotional Database	[19], [20], [21]	14.28%
German Language Emotional Database	[22], [23]	09.51%
Italian Language Emotional Database	[24], [25]	09.51%
Japanese Language Emotional Database	[26]	04.76%
Korean language Emotional Database	[27]	04.76%
Arabic Language Emotional Database	[28]	04.76%
Multilingual Emotional database	[29], [30]	09.51%

(c) Frequently used databases by the researchers for SER: According to the findings of the study, the researchers used a variety of databases. The IEMOCAP database was used to analyse the bulk of the SER, followed by EMODB and RAVDESS. We regarded IEMOCAP, EMODB, and RAVDESS to be widely used datasets because they were used by more than 15% of the SER models. Figure 2.2 depicts the distribution of paper according to databases and its proportion.

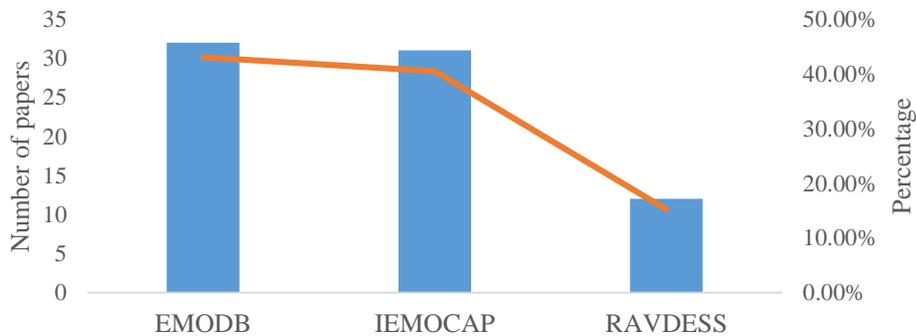


Figure 2.2: Distribution of paper according to the databases

(d) Different features frequently used by the researchers for SER: According to the literature, scholars have used and studied a variety of aspects. The spectral features, prosodic features, Teager Energy operator, and voice quality features extracted from speech for SER are classified as local features, global features, or both, and these local and global speech features are investigated in four types: spectral features, prosodic features, Teager Energy operator, and voice quality features. Table 2.6 lists the speech characteristics, as well as their purposes and techniques.

Table 2.6: Speech Features types their purpose and approaches

Speech			
Feature types	Features	Purpose and approach	Ref.
Spectral features	MFCC, LPCC, LFPC, GFCC	<ul style="list-style-type: none"> • Spectral features are created using vocal tract characteristics • Spectral features are created using vocal tract characteristics • Spectral features are acquired using the Fourier transform • The smaller the number of MFCCs, the more speech information is included Stress-related emotional states in humans are being modelled. 	[36], [37], [38], [39], [40], [41]
Prosodic feature	Fundamental frequency, Pitch, Energy, Duration, Frame level features	<ul style="list-style-type: none"> • Prosodic traits are thought to be good predictors of human emotions. • Prosodic characteristics at the global and local levels derived from sentences • Long-term characteristics collected from speech are prosodic features. • Prosodic characteristics are often used on fundamental frequency, duration, and energy features. 	[42], [43], [44], [45],[46], [47], [48]
Teager Energy operator	TEO critical band auto envelope, TEO auto envelope, TEO frequency modulation variation	<ul style="list-style-type: none"> • It has been shown that in a stressful condition, there is a shift in crucial band and frequency owing to the distribution of harmonics • Without stressful settings, altering the airflow during speech creation 	[43], [44], [45]

Voice quality features	Bandwidth, Glottal parameter, Harmonics to noise ratio, Shimmer, Jitter	<ul style="list-style-type: none"> • The voice quality speech and emotional content have conflicting interrelationships • The voice quality speech and emotional content are at the centre of speech processing and speaker identification 	[50]
------------------------	---	--	------

2.5. Conclusion

This chapter provides literature review on SER in terms of speech databases, speech characteristics and different types of emotional speech databases. In this chapter different types of emotional speech databases like actor-based, induced/semi-natural, and natural emotional databases are analysed and reviewed. The review of existing speech features such as local and global features are presented briefly. Some of the most commonly used speech datasets are highlighted. Finally, the reviews' results are discussed.

CHAPTER 3

SER CLASSIFIERS

3.1. Related Work for SER classifiers

This chapter presents a thorough assessment of the systematic literature review of SER classifiers. The use of traditional and deep learning approaches to SER classifiers is contrasted and compared. The DL approaches used for SER are studied in depth. Finally, the findings of the reviews are discussed. Classifiers are equally as important as speech features in SER. Classical classifiers and deep learning classifiers are the two types of classifiers. Hybrid approaches, which blend conventional and deep learning classifiers, have been used by certain researchers. A large number of classifiers have been evaluated for SER, but it is still difficult to determine which ones are the most effective. This chapter contains a critical analysis of traditional/deep learning classifiers, which is based on the type of database used, recognised emotions, features used, classifier type, and average accuracy.

3.1.1. Traditional Approaches

After extracting the necessary characteristics from the voice data, traditional SER techniques are used. Many classifiers have been examined by the researchers for SER to attain improved accuracy. SVM, HMM, GMM, ANN, and k-NN are the most often used classical classifiers [5]. Using ELM Extreme Learning Machine, SVM, BPNN, and KNN classifiers, a novel feature selection technique based on fisher and correlation analysis was suggested and assessed on the CASIA datasets in [33]. The average accuracy rates were 89.9%, 87.20 percent, 82.30 percent, and 80.70 percent, respectively. For SER, SVM and HMM were utilised and assessed on the SUSAS database [51]. The author employed a KNN classifier to recognise emotions and achieved an average accuracy of 66.4 percent with four emotions [52]. The features were extracted using LPA and MFCC in [54, 55], and the SER system was tested using GFNN and SVM classifiers on the EMODB database, with accuracy of 98 percent and 82 percent, respectively. The author utilised a hybrid technique and tested it on the three datasets [56]. SVM classifiers were used to distinguish four emotions: anger, neutral, sad, and happy, with an average accuracy of 73 percent [57].

In [58], neural networks and SVM were deployed and assessed on the eNTERFACE and FAU datasets. The author developed SER based on Fourier parameters, which he tested on three datasets and found to be the most accurate at 71% [59]. The MFCC features extraction approach was utilised for SER in [60] and [61], and the results were assessed on EMODB using SVM and ANFIS MLP classifiers, respectively. The conventional techniques' literature is examined, including the databases utilised, the number of emotions identified, the feature set used, the type of classifiers employed, and the average accuracy attained. The literature on classic SER techniques is summarised in Table 3.1 below.

Table 3.1: Literature Summary of the traditional approaches for SER

Ref.	Database Used	Recognized Emotions	Features used	Classifiers	Accuracy
[33]	CASIA	Happy, Sadness, Surprise, Angry, Fear, Neutral	A Feature selection method used	ELM Decision Tree, SVM, BPNN, KNN	89.90%, 87.20%, 82.30%, 80.70%
[51]	SUSAS	Neutral, Angry, Happy, Sad, and Bored	Pitch, MFCC	QDA, SVM, HMM, LDA	70%
[52]	Linguistic Data Consortium	Anger, Neutral, Sad, and Happy	MFCC	KNN	66.4%
[53]	Malayalam	Anger, Neutral, Sad, and Happy	MFCC, LPC	DWT, ANN	68%
[54]	EMODB	Anger, fear, joy, sad, neutral, and disgust, and boredom	LPC	GFFNN	98%
[55]	EMODB	Anger, fear, joy, sad, neutral, and disgust, and boredom	MFCC, PPCMCC	SVM	82%
[56]	Polish, EMODB, LDC	Neutral, Fear, and Anger	Mutual Information, dimensional correlation, entropy, Lempel-Ziv	NN	72.28%, 75.40%, 80.75%
[57]	IEMOCAP	Anger, Neutral, Sad, and Happy	Pitch, Energy, (MFBs)	DBN, SVM	73%
[58]	eNTERFACE, FAU	Anger, Disgust, Fear, Happy, Sad, Surprise	Pitch, utterance level	NN, SVM	70%, 60%
[59]	EMODB, ESSDB, CASIA	Anger, fear, joy, sad, neutral, and disgust, and boredom	MFCC	Fourier parameters	71%

[60]	EMO-DB	Anger, fear, joy, sad, neutral, and disgust, and boredom	MFCC	NN, SVM	73-85.5%
[61]	EMODB	Anger, fear, joy, sad, neutral, and disgust, and boredom	MFCC	ANFIS MLP	72.5%

Table 3.1 summarises the literature and shows that SVM was used as a classifier in 54.45% of the articles, with the greatest accuracy of 98 percent utilising MFCC features on the EMODB database. For emotion recognition, the bulk of publications use either MFCC or MFCC with a mix of additional characteristics as speech features. The EMODB was used to analyse the bulk of the SER model, followed by eNTERFACE.

3.1.2. Deep Learning Approaches

Deep learning is a subtype of machine learning that learns from a variety of layers. In 2015, SER experienced a surge in deep learning. Deep learning algorithms have recently yielded promising results in the field of SER. Because of its flexibility, automated learning of a wide variety of characteristics, scalability, and identification rate, deep learning algorithms are regarded the best classifiers for emotion recognition when compared to classical approaches. Traditional training methods, on the other hand, needed less data. Emotion recognition has been studied using a number of deep learning algorithms. Table 3.2 summarises the important characteristics and limitations of the deep learning algorithms employed for SER.

Table 3.2: Deep learning algorithms utilised for SER

Deep Learning Approaches	Key Feature	Limitations	Brief Description
CNN	<ul style="list-style-type: none"> • More suitable for 2D data. • Only a few connections are required concerning NN 	<ul style="list-style-type: none"> • Required large, labelled data. 	The first CNN model was created by LeCun et al. [62]. CNN is mostly used for image processing and handwriting recognition. In the first layer of the CNN architecture, the input is followed by a convolution operation to extract low-level features, then an activation function, and finally a pooling layer to reduce the output dimension (extracted feature). Next, a convolution operation is performed, followed by an activation function, and finally a pooling layer, which is supplied to a connected multilayer perceptron. The activation function (soft-max), which classifies the

			output, comes after the fully linked layer. In the application of image processing, CNN performs better than other deep learning algorithms.
RNN	<ul style="list-style-type: none"> • In NLP area give better result • Widely use where output depends on previous data 	<ul style="list-style-type: none"> • Learning issue is there due to gradient problem 	The RNN has a complex architecture. The RNN has a feedback link to the previous layer, which is one of its distinguishing features. RNN may be improved and spent. RNN is based on the work of Williams [63].
LSTM	<ul style="list-style-type: none"> • Similar to RNN and it has memory units. 	<ul style="list-style-type: none"> • Take a long time to train and require more memory. • Dropout is harder to implement in LSTM 	Many applications employ Long Short Term Memory (LSTM), and IBM chose it for speech recognition. LSTM is similar to RNN in that it contains a memory unit that stores data for a given period of time and may be used as an input function. The last computed value may be remembered with the aid of this memory unit. LSTM is employed in various applications and provides a makeable outcome. For spectrum prediction, Yu et al. [64] developed LSTM. Lopez et al. [65] describe a model based on LSTM for greater predicting accuracy.
Auto-encoder	<ul style="list-style-type: none"> • No need for labelled data • Many auto encoder versions are designed according to the need 	<ul style="list-style-type: none"> • Training can suffer from error 	The unsupervised learning notion underpins the auto-encoder deep learning architecture. The auto-design encoder's is made up of three layers: the first is an input layer, the second is a hidden layer called the encoder layer, and the third is an output layer called the decoding layer. Principal Component Analysis is linked to auto-encoder (PCP). Azar and Hamey [66] describe a solution based on an unsupervised auto-encoder to decrease training and deployment costs.
DBN	<ul style="list-style-type: none"> • It allows both supervised and unsupervised learning 	<ul style="list-style-type: none"> • Training can be expensive 	To generate output, a deep belief network (DBN) employs unsupervised learning and probability. It has levels that are both directed and undirected. Each layer in DBN learns the full input and functions on a global scale. The DBN network is similar to a constrained Boltzmann machine, in which each layer is connected to all nodes in the layer before it, and so on. Three alternatives were offered by Maghera [66]. CNN, DNN, and DBN are some of the most well-known news organisations in the world.

RBM	<ul style="list-style-type: none"> • Better than an auto-encoder it ignores the noise during training • Concerning DBN higher time complexity • For large dataset optimization of the parameter is tedious 	<p>The visible layer is connected to the hidden layer in the Restricted Boltzmann Machine (RBM) model, and there is no contact between the hidden layer and the input. A network categorization is provided by the output layer. The training is divided into two parts: one is a surprise fine training session, and the other is an unsupervised pre-training session. When one component of the training is done, the next section of the training begins. A stack of RBM is another name for RBM.</p>
-----	---	---

The present researchers are applying deep learning algorithms for emotion identification and boosting accuracy, according to the SER literature. Utilizing techniques that automatically learn important features, employing hand-crafted features, and using spectrograms are three major kinds of SER models that use deep learning methodologies. The author suggested a CNN-based unique spatiotemporal and frequently cascaded network for emotion identification and tested it against the IEMOCAP, EMODB, eNTERFACE, and SAVEE, with average accuracy of 71.98 percent, 82.10 percent, 75.60 percent, and 54.75 percent [67]. The DECN SER model offers to enhance the result by correcting mistakes generated by emotion recognition algorithms [68]. The authors presented a SER method based on an LSTM classifier, which they tested on the RAVDESS database and found to be more accurate [69], [74]. To recognise emotions, a one-dimensional CNN has been suggested [70], [71]. In [72], [73], [76], [77], [78], [79], [84], speech signals were transformed to spectrograms, and emotions were detected using CNN. The suggested model was tested on several databases, with the greatest accuracy for the SAVEE database being 81.70 percent [75]. MFCC and prosodic characteristics were retrieved, and a deep neural network was used. In [82], [83], [84], CNN was combined with other classifiers to enhance the results in terms of average accuracy. The authors of [86] picked GoogLeNet for emotion identification and tested it against a variety of databases. The authors of [87], [88], [89], [90], [92], [101], [104] developed and tested SER techniques based on CNN and LSTM classifiers on various databases. To recognise speech emotions, the authors in [93], [101] employed 3D CNN-based SER models. To recognise emotions, the authors in [95] employed a hybrid technique of GMM+DNN. In [102], RBM-based DBN classifiers incorporating eGEMAPS features were suggested and shown to be more accurate. For emotion recognition, RNN classifiers with LSTM and CNN classifiers were employed to improve the results [105], [107]. CNN classifier was utilised by authors with different architectures for SER system and assessed on numerous databases in [108], [112], [114], [116]. The authors of [109] used Bi-LSTM classifiers and achieved an average accuracy of 93.97 percent. For emotion recognition,

the authors of [110], [115] employed an auto-encoder classifier. Hand-crafted characteristics are taken into account when utilising the DBN classifier to distinguish emotion and increase recognition accuracy [111]. To increase the recognition rate, features retrieved using openSMILE in [117] and RNN with LSTM classifiers were utilised. CNN layers have automatically learnt characteristics to discern emotions [119]. [120] collected MFCC features and used RNN classifiers to recognise emotions, and the proposed model was tested on EMODB [120]. The authors developed a SER model utilising DBN classifiers and tested it on the EMODB and IEMOCAP datasets in [121], [123]. The authors employed an RNN model to improve the results [125]. To increase accuracy, the authors suggested a SER model employing DNN classifiers and features collected from spectrograms in [126]. The CNN classifier has been tested on the IEMOCAP and CAM3D datasets for emotion identification [127, 128]. The CNN model is used to identify the emotion when speech inputs are transformed into spectrograms [130,131]. In [129, 132], the authors suggested the SER model based on DNN classifiers. DBN classifiers are utilised for emotion identification with hand-crafted features in [133] and [135]. To improve accuracy, MFCC features taken from [134, 136] for SER and LSTM and DNN classifiers are employed. In [137], an ANN classifier was employed for emotion identification and tested on the EMODB and Mandarin datasets.

The literature on deep learning methodologies is examined in depth, including the databases utilised, the number of emotions identified, the feature set used, the type of deep learning classifiers employed, and the average accuracy reached. Table 3.3 discusses the critical examination of the literature review of deep learning algorithms for emotion recognition.

Table 3.3: A survey of the literature on deep learning techniques for SER

Ref.	Database Used	Recognised Emotions	Features used	Classifiers	Accuracy
2021					
[69]	RAVDESS	Happy, Sad, Fear, Anger, Disgust, and Surprised	MFCC and spectrograms	LSTM	73.50%
[70]	IEMOCAP, EMODB	Happy, Sad, Fear, Anger, Disgust, Boredom and Neutral	Automatically learning	ID dilated CNN	73.01%, 90.01%
[71]	RAVDESS	Happy, Sad, Anger, Calm, Fear, and Nervous	MFCC	1D CNN	82.30%
[72]	IEMOCAP, EMO-DB, and RAVDESS	Happy, Sad, Fear, Anger, Disgust, Boredom and Neutral	Spectrograms	dilated CNN	78.01% 93.00%, 80.00%
[73]	EMODB, SAVEE, EMOVO	Happy, Sad, Fear, Anger, Disgust, Boredom and Neutral	Spectrograms	DCNN	84.62%, 75.00%, 69.65%
[74]	IEMOCAP RAVDESS	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Automatically learning	LSTM	67.20%, 84.30%
[75]	RAVDESS, SAVEE, IEMOCAP	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	13 MFCC, + 19 prosodic features	DNN	81.20%, 81.70%, 74.50%
2020					
[76]	RAVDESS EMODB IEMOCAP	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	MFCC, spectrogram, Chroma, Contrst, Tonnetz,	Deep CNN	71.61%, 86.10%, 64.30%
[77]	IEMOCAP RAVDESS	Anger Disgust, Fear, Happy, Sad, Neutral, and Surprise	spectrogram	CNN	81.75%, 79.50%
[78]	IEMOCAP EMODB RAVDESS	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	spectrogram	CNN BiLSTM	72.25%, 85.57%, 77.02%
[79]	EMODB RAVDESS TESS IEMOCAP	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Automatically learning	CNN	70.97%, 55.85%, 53.05%, 55.01%

[80]	IEMOCAP RAVDESS	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Spectrograms	DSCNN	84.00%, 81.00%
[81]	eINTERFACE	Anger, Disgust, Fear, Happy, Sad, Surprise	OpenSMILE, MFCC, Fundamentl frequency, jitter, shimmer	CNN	81.36%
[82]	IEMOCAP	Anger, Disgust, Happy, and Neutral	Automatically learning	DNN+CN N+RNN	58.3 %
[83]	IEMOCAP	Anger, Disgust, Happy, and Neutral	spectrograms	Time- frequency CNN+EL M	70.78%
[84]	IEMOCAP EMODB	Anger, Disgust, Fear, Happy, Sad, Neutral, and Boredom	spectrograms	CNN	83.00%, 95.00%
[85]	IEMOCAP	Anger, Disgust, Happy, and Neutral	MFCC	CNN+LST M	73.30%
[86]	RAVDESS, EMODB, IEMOCAP	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Spectrograms	GoogLeNe t	67.10%, 72.5%, 67.20%
2019					
[87]	IEMOCAP,E MOVO,SAVE E, EMODB, EPST, RAVDESS, TESS	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	GeMAPS feature set	CNN, LSTM	Max (69%) among all
[88]	AFEW	Happy, Surprise, Anger, Disgust, Fear, Sad and Neutral	Automatically learning using CNN	LSTM	60.59%
[89]	IEMOCAP	Anger, Disgust, Happy, Neutral	Automatically learning	CNN+LST M	61.20%
[90]	EMODB, IEMOCAP	Happy, Sad, Neutral, Surprised, Disgust, Fear, and Anger	Spectrograms	2D CNN LSTM	95.89%, 89.16%
[91]	CASEC	Happy, Sad, Neutral, Surprised, Fear, and Anger	Prosodic and spectral features	DNN- Decision Tree	75.83%

[92]	FAO-Aibo RAVDESS	Anger, Calm, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Automatically learning	CNN+BiL STM	77.60%, 56.20%
[93]	SAVEE, RML, eNTERFACE	Anger, Disgust, Fear, Happy, Sad, and Surprise	spectrograms	3D CNN	81.05%, 77.00%, 72.33%
[94]	EMODB, Spanish database	Joy, Anger, Disgust, Fear, Neutral, Sad, and Surprise	MS, MFF, SMFCC, ECC, EFCC	RNN	85.94%, 91.16%
[95]	Emirati speech database	Neutral, Happy, Fear, Sad, Disgust, Anger	MFCC	GMM- DNN	83.97%
[96]	IEMOCAP	Anger, Disgust, Happy, and Neutral	Energy, Entropy, Flux, MFCC, Chroma	BiLSTM	70.34%

2018

[97]	IEMOCAP, EMODB	Anger, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Mel- spectrograph	ACRNN	82.82%
[98]	EMODB, IEMOCAP	Anger, Disgust, Fear, Happy, Sad, Neutral, and Surprise	Spectrograms	DCNN	92.71%, 86.36%
[99]	IEMOCAP	Anger, Frustration, Sad, and Neutral	spectrograms	CNN	47%
[100]	IEMOCAP	Happy, Sad, Neutral, Surprise, Disgust, Fear, and Anger	MFCC	TDNN- LSTM	70.60%
[101]	AVEC	Anger, Happy, Sad, Neutral, Bored, Disgust, and Fear	eGEMAPS	3-D CNN + LSTM	66.40%
[102]	FAU-AIBO, IEMOCAP, EMODB, SAVEE, EMOVO	Touchy, Rest, Anger, Happy/joy, Sad, Neutral, Bored, Disgust, and Fear	eGEMAPS	RBM based DBN	74.11%, 54.77%, 72.38, 56.76%, 76.22%
[103]	IEMOCAP	Anger, Disgust, Happy, and Neutral	Extracted features using openSMILE toolkit	Auto- encoder	57.88%
[104]	CMU-MOSEI	Sadness, Happy, Anger, Disgust, Surprised, and Fear	Spectrograms	LSTM based CNN	83.11
[105]	IEMOCAP	Anger, Happy, Sad, and Neutral	MFCC	LSTM based RNN	71.04%

[106]	IEMOCAP	Anger, Happy, Sad, and Neutral	Spectrograms	2D CNN	68.5%
[107]	IEMOCAP	Anger, Happy, Sad, and Neutral	eGEMAPS	RNN + CNN	83.20%
<hr/>					
2017					
[108]	IEMOCAP	Anger, Happy, Sad, and Neutral	Automatically learning	CNN	64.78%
[109]	SEED, DEAP	Positive, Neutral, Negative, Arousal, Valence, Like/Dislike, Dominance, and Familiarity	Differential Entropy	Bi-LSTM	93.97%, 83.83%
[110]	DEAP	Valence, Like/Dislike, Dominance, and Familiarity	Automatically learning	Auto-encoder	80.78%
[111]	CASEC	Fear, Anger, Neutral, Joy, Surprise, and Sad	Pitch, energy, zero-crossing, MFCC	DBN	94.60%
[112]	RECOLA	Anger, Happy, Sad, and Neutral	Automatically learning	CNN+ LSTM	78.70%
[113]	ABC, and EMODB	Happy, Sad, Neutral, Surprise, Disgust, Fear, and Anger	ZCR, energy, F0, HNR, MFCC	EDFLM	65.62%, 61.63%
[114]	MODB, eINTERFACE 05, RML and BAUM	Anger, Fear, Happy, Sad, Neutral, and Surprise, and Average	Spectrograms	DCNN	87.31%, 76.56%, 69.70%, 44.61%
[115]	ABC, EMODB, SUSAS	Anger, Fear, Happy, Sad, Neutral, and Disgust, and Boredom	Extracted features using openSMILE toolkit	Auto-encoder	63.30%, 62.00%, 62.80%
[116]	EMODB	Anger, Fear, Joy, Sad, Neutral, and Disgust, and Boredom	Spectrograms	CNN	56.37
[117]	RECOLA	Anger, Happy, Sad, and Neutral	Extracted features using openSMILE toolkit	RNN with LSTM	74.40%
[118]	Mandarin Data set	Anger, Fear, Joy, Sad, Neutral, and Disgust, and Boredom	Automatically learning	DNN based ELM	82.18%

2016						
[119]	AFEW4	Anger, Disgust, Fear, Happy, Sad, Surprise, and Neutral	Automatically learning	CNN		49.49%
[120]	EMODB	Anger, Fear, Happy, Sad, Neutral, Disgust, and Boredom	MFCC	RNN		60%
[121]	EMODB	Anger, Fear, Happy, Sad, Neutral, and Disgust, and Boredom	Automatically learning	DBN		65%
[122]	CHEAVD	Anger, Happy, Sad, Neutral, Worried, Anxious, Disgust, and Surprised	Extracted features using openSMILE toolkit	DFFN		50.01%
[123]	IEMOCAP	Anger, Happy, Sad, and Neutral	MFCC	DBN		59.60%
[124]	ABC, EMODB	Happy, Sad, Neutral, Surprised, Disgust, Fear, and Anger	ZCR, energy, f0, HNR, MFCC	Two-layer DNN		61.54%, 57.58%
2015						
[125]	RECOLA	Arousal, and Valence	Loudness, f0, Jitter, shimmer, MFCC	RNN		81%, 55%
[126]	eINTERFACE, SAVEE	Anger, Boredom, Disgust, Joy, Sad, and Neutral	Spectrograms	DNN		60.53%, 59.7%
[127]	IEMOCAP	Neutral, Anger, Fear, Sad, Joy, and Happy	STFT	CNN		40.02%
[128]	CAM3D	Anger, Boredom, Disgust, Joy, Sad, and Neutral	Automatically learning	CNN		74.20%
2014						
[129]	IEMOCAP	Neutral, Excitement, Frustration, Surprised, and Happy	Segment level features	DNN based ELM		54.30%
[130]	SAVEE, EMODB, DES, MES	Anger, Boredom, Disgust, Happy, Sad, and Neutral	Spectrograms	CNN		73.60%, 85.20%, 79.90%, 78.30%
[131]	SAVEE, EMODB, DES, MES	Anger, Boredom, Disgust, Happy, Sad, and Neutral	Spectrograms	CNN		71.80%, 57.20%, 60.40%, 57.80%

[132]	9595 Emotion sentences	Anger, Fear, Happy, Surprised, Sad, and Neutral	MFCC	DNN	92.10%
<hr/>					
Till 2013					
[133]	IEMOCAP	Anger, Sad, Happy, and Neutral	Pitch, energy, MFCC	DBN	72.77%
[134]	SEMAINE	Sensible, Happy, Sad, and Anger	MFCC	BiLSTM	69.8%
[135]	NVIE	Disgust, Happy, and Fear	Facial Image	DBM	51.3%
[136]	ABC, AVIC, DES, EMODB, EnTER, SAL, Smart, SUSAS, VAM	Arousal, and Valence	Energy, pitch, MFCC	DNN	61.5%, 79.50%, 56.60%, 81.90%, 61.10%, 34.30%, 59.50%, 53.60%, and 68.00%
[137]	EMODB, Mandarin	Anger, Happy, Sad, Boredom, Disgust, Fear, and Neutral	Pitch, energy, f0, duration, formant	ANN	61.40%, 63.30%

Table 3.3 shows that before 2013, relatively few researchers employed DL methods for SER, with the majority using the DBN classifier [133], [135]. The majority of studies utilised CNN with spectrograms in 2014 [130], [131]. In 2015, an RNN classifier assessed on the RECOLA database [125] got the best recognition accuracy of 81 percent. RNN, DBN, CNN, and DNN classifiers were utilised by researchers in 2016. The DBN classifier was assessed on the CASEC database in 2017 and achieved the best accuracy of 94.60 percent [111]. On the EMOBD datasets [98], greater accuracy of 92.71 percent was attained using DCNN in 2018. All two dimensional techniques have employed quite different and hybrid approaches in 2019. On the EMOBD dataset, CNN + LSTM techniques achieved the highest accuracy of 95.89 percent and 89.16 percent on the IEMOCAP datasets [90]. The greatest accuracy attained using CNN on EMOBD datasets in 2020 will be 95% [84]. In the year 2021, the majority of research will be conducted utilising CNN and hybrid techniques.

Table 3.3 further shows that the majority of the researchers tested their suggested SER model on the IEMOCAP database (48 percent), followed by the EMOBD database (38 percent), RAVDESS database (17 percent), and SAVEE database (10 percent). It's also evident that CNN, LSTM, and RNN deep learning algorithms were employed in the majority of the work. In a three-way experiment, researchers employed deep learning algorithms for SER. I the first method, in which researchers extract speech features and then use deep learning to identify

emotions; ii) the second method, in which researchers convert speech signal into spectrogram and then use deep learning to identify emotions; and iii) the third method, in which deep learning approaches automatically learn the relevant speech features to identify emotions. Authors [90] used the CNN +LSTM technique to extract features from spectrograms and got the greatest accuracy of 89.16 percent for IEMOCAP and 95.89 percent for the EMODB database. Using the LSTM method, which automatically learns speech characteristics, the greatest accuracy for the RAVDESS database was 84.30 percent [74]. Using the CNN technique to extract features from spectrograms, the greatest accuracy for the SAVEE database was found to be 81.05 percent [93]. Auto-encoder, DBN, and DBM deep learning techniques have only been employed by a few researchers.

3.2. Findings of the Review

This section discusses all of the research findings.

More than 65 percent of the articles collected in this study are connected to SER classifiers, 18 percent are for emotional speech databases, more than 14 percent are for speech characteristics, and the remaining papers are for the SER system's background. Papers are gathered from a variety of sources, including Elsevier, Springer, IEEE, Research Gate, and others. Figure 3.1 shows the distribution of the number of papers and the percentage according to the resources.

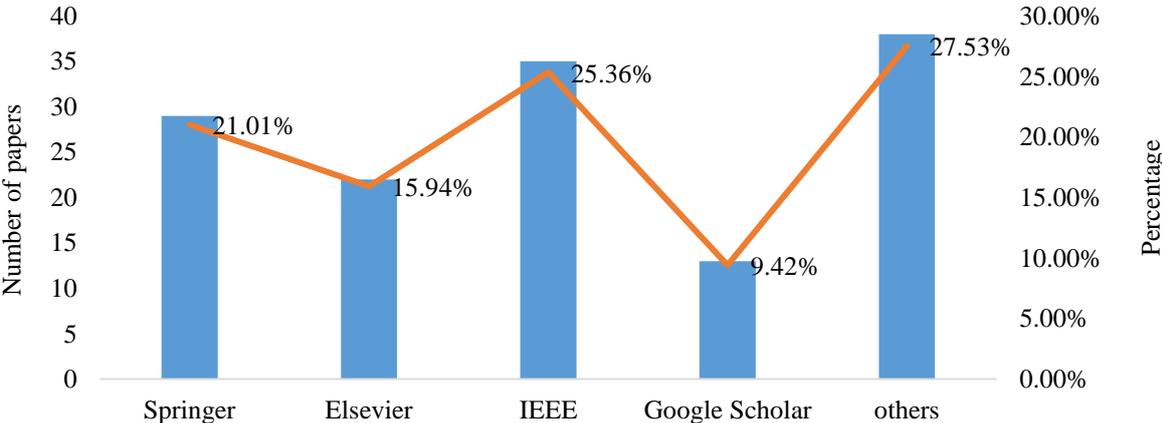


Figure 3.1: The number of publications and their percentages based on the available resources

The distribution of all gathered articles in terms of year-by-year number of papers and resources is illustrated in Figure 3.2 in this literature.

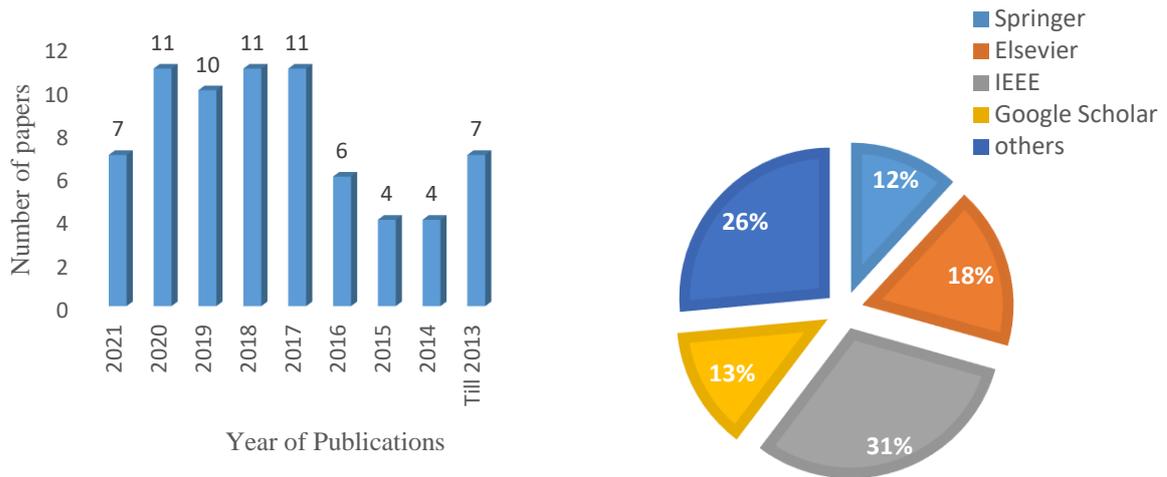


Figure 3.2: Distribution of the collected paper in terms of year-wise number of papers and resources

Figure 3.2 shows that before 2013, relatively few researchers employed DL techniques for emotion recognition, and after that, the number of researchers using DL approaches for emotion recognition has progressively increased. DL techniques are now used in the majority of SER models provided by researchers, and they have reported superior outcomes in terms of average accuracy and computation cost.

We looked at 68 SER articles that used DL techniques for this research. Table 2.2 shows that IEMOCAP is the most often used database, followed by EMODB and RAVDESS. CNN, RNN, LSTM, Auto-encoder, RBN, RMB, and a mixture of these deep learning algorithms were employed by the authors for SER. The CNN+LSTM model with spectrogram treated as features attained the best accuracy (89.16 percent) on the IEMOCAP database and (95.89 percent) on the EMODB database [90]. The LSTM model, which automatically learns features, was used to attain the greatest accuracy (84.30 percent) for the RAVDESS database [74].

There are several other reasons why Deep Learning techniques to recognising emotions from speech are used. Although, in terms of recognition rate and calculation cost, DL techniques give a better option for SER. However, there are other obstacles that DL methods must overcome. Table 3.4 contains a summary of incentives and limits.

Table 3.4: Deep learning techniques for SER motivations and limits

Deep learning algorithms for SER are being used since there is a strong desire to do so.	Deep Learning's Limitations
<ul style="list-style-type: none">• Deep learning based on NN with the addition of more than two-layer.• It has a strong learning ability.• It can solve the complex computational task, and it solves the task end to an end basis.• DL approaches are more scientific and correct when the data is large.• DL extract new and relevant features.	<ul style="list-style-type: none">• It is crucial to catch the training parameter and topology for DL.• It requires a large amount of data.• It requires resources like high-performance GPUs and memory requirements.• DL approaches are required to continuously manage input data.• It is difficult to manage hyper parameters and complex design.• It requires high computational power.• It may be expensive for a complex problem.

3.3. Conclusion

This chapter evaluated and analysed the existing literature in terms of conventional ML and DL approaches utilised for SER from 2000 to 2021. The finding of the literature reviews such as accuracy achieved by different deep learning approaches of the frequently used databases are summarised. Existing research focuses on a number of issues, such as classifiers, cross-lingual emotion detection, and real-time emotion identification, which are all yet unsolved. Finally, there are reviews of the deep learning methodologies for SER's limitations and rationale.

CHAPTER 4

DEEP CNN BASED ALGORITHM FOR RECOGNITION OF EMOTION

4.1. Overview

Emotions are crucial to a successful human-computer relationship. Affective computing is a branch of computer science concerned with the study of emotions. Sentiment analysis seeks to discover emotions that may be utilised to improve client connections in a variety of systems, such as recommender systems. This chapter focuses on speech-based emotion recognition. Developing systems that can communicate with people via voice is difficult [138]. One of the key research concerns in emotion identification is the extraction of acoustic information from speech signals to improve accuracy. Several traits have been classified as Prosodic features, Spectral features, or a combination of both in the literature [5]. Intone, energy, speaking pace, fundamental frequency, length, intensity, and spectrum characteristics are all acoustic qualities. Several Machine Learning Algorithms (MLAs) have been used to discern emotions based on the spectral and prosodic aspects of voice signals for emotion recognition. The Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), linear discriminant classifiers, nearest neighbour classifiers, Support Vector Machine (SVM), Artificial Neural Networks (ANNs), and Convolutional Neural Networks (CNN) are some of the classifiers that are widely used to recognise emotions based on acoustic features of utterances [139, 140]. Recently, researchers have looked into deep learning techniques for emotion detection to enhance recognition rates, while others have employed hand-crafted low-level features to train CNN, RNN, and DNN models to improve emotion recognition accuracy.

Many studies have employed CNN from deep learning methodologies for emotion identification from voice, according to the literature. Because of the huge pre-trained architecture, using deep learning methodologies increases the model's accuracy but also increases its processing cost. Using spectrograms as input, a few researchers have created algorithms for recognising emotion from speech. The accuracy of these classifiers' emotion recognition is determined by the spectral and prosodic features of speech, as well as the feature extraction techniques used. The features of speech signals are influenced by language, culture, speaker, and gender. As a result, there are many "hand-crafted" components of deep learning

that can be handled automatically. The German Emotional Speech Database (EMODB), British English Surrey audio-visual Expressed Emotion database (SAVEE), and Italian Emotional Speech Corpus are used to test the DCNN architecture's ability to recognise emotion from actor-based corpora in three languages (EMOVO). The proposed study aims to improve emotion recognition accuracy for all languages and speakers. The following are the study key contributions:

- Using the DCNN architecture, an algorithm for recognising emotions that is independent of language and actor is suggested.
- Rather than audio files, RGB spectrograms (pictures) of actor-based speech corpora are produced and normalised before training. For fine-tuning to DCNN, all spectrograms are kept to the same size of 224x224x3.
- The DCNN architecture then automatically learns the best features from labelled examples. Seven emotions are detected with more accuracy than previous experiments using three databases of various languages.
- When comparing the optimal learning rate to the random learning rate, the improvement in accuracy is provided.

The rest of the chapter is organised as follows: In section 4.2, the proposed algorithm is described in depth. The experimental details and findings are given in section 4.3. Finally, in section 4.4, conclusion is discussed.

4.2. Proposed algorithm

An algorithm for emotion recognition based on DCNN architecture is suggested in this study. Stride as well as pooling layers are employed in convolution for down - sampling the output feature map. It lowered the computational cost of the proposed DCNN-based model while simultaneously improving emotion identification accuracy. It is assessed using the EMOVB, EMOVO, and SAVEE speech corpora.

4.2.1. Pseudocode of the proposed algorithm

The implementation of the suggested algorithm for emotion identification from speech using DCNN architecture is detailed in this part. Table 4.1 shows the pseudo-code for implementing the suggested method.

Table 4.1: Pseudo code to implement the proposed algorithm

Input: Audio files (Labelled actor-based speech corpora)

Output: Recognize the emotion as output (like happiness, sadness, anger, fear, disgust, boredom/surprise, and neutral)

Step 1: Read .wav files from the speech corpora.

Step 2: Get spectrograms of each .wav file by a Short-Time Fourier Transform(STFT) of the wave signal.

Step 3: Convert all spectrograms to size 224x224x3.

Step 4: Divide the data sets into

- Train=80% (of all data sets)
- Test =10% (of all data sets)
- Validation=10% (of all data sets)

Step 5: Train the deep learning model and save as stage-1 (i.e. freeze)

Step 6: Using the Learning Rate (LR) range test, find an optimal learning rate.

Step 7: Unfreeze stage-1 and train the deep learning model with an optimal learning rate and save as stage-2 (i.e. freeze).

First and foremost, the voice samples are taken from the audio files of the speech corpora EMODB, SAVEE, and EMOVO. The .wav files from the speech corpus are read and transformed into spectrograms. Figure 4.1 depicts the entire procedure in full.

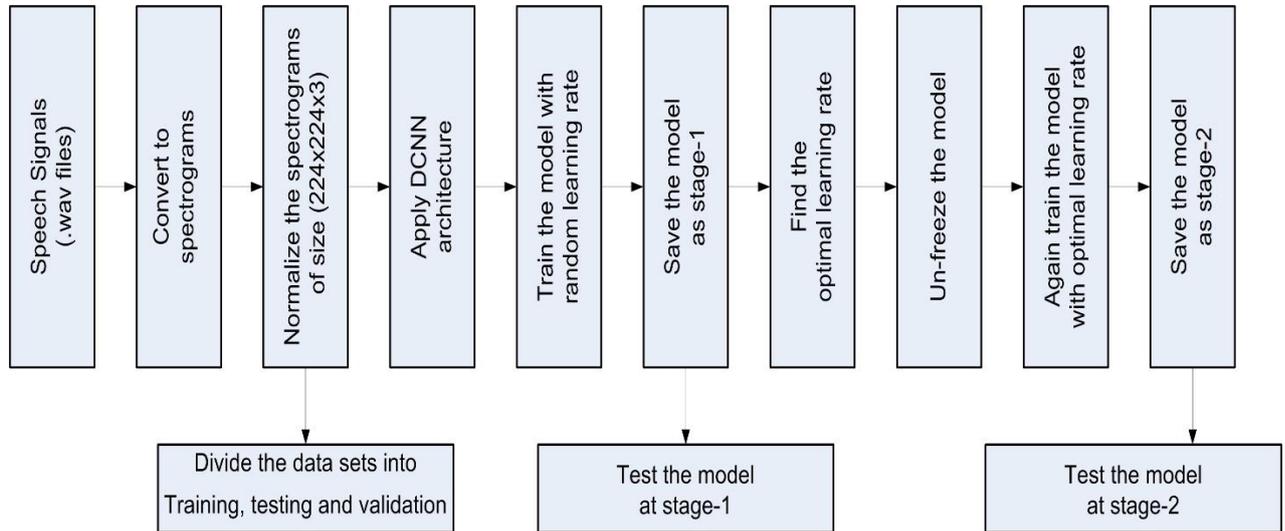


Figure 4.1: Steps of the proposed algorithm

4.2.2. Details of DCNN architecture

In the early phases of this technique, audio recordings from speech corpora are transformed into spectrograms and resized to 224x224x3. The accuracy of emotion detection improves because spectrograms include more information that cannot be retrieved from voice signals. From RGB spectrograms, the DCNN architecture learns high-level characteristics.

Convolutional layers, a max-pooling layer, and fully linked layers make up the DCNN design. Fully connected layers are supplied to a softmax classifier to compute the likelihood of each emotion. Instead of using the pooling layers, a special stride is employed to down sample the output features. By using a convolution layer, labelled samples learn optimum characteristics Figure 4.2 depicts the DCNN architecture in further detail.

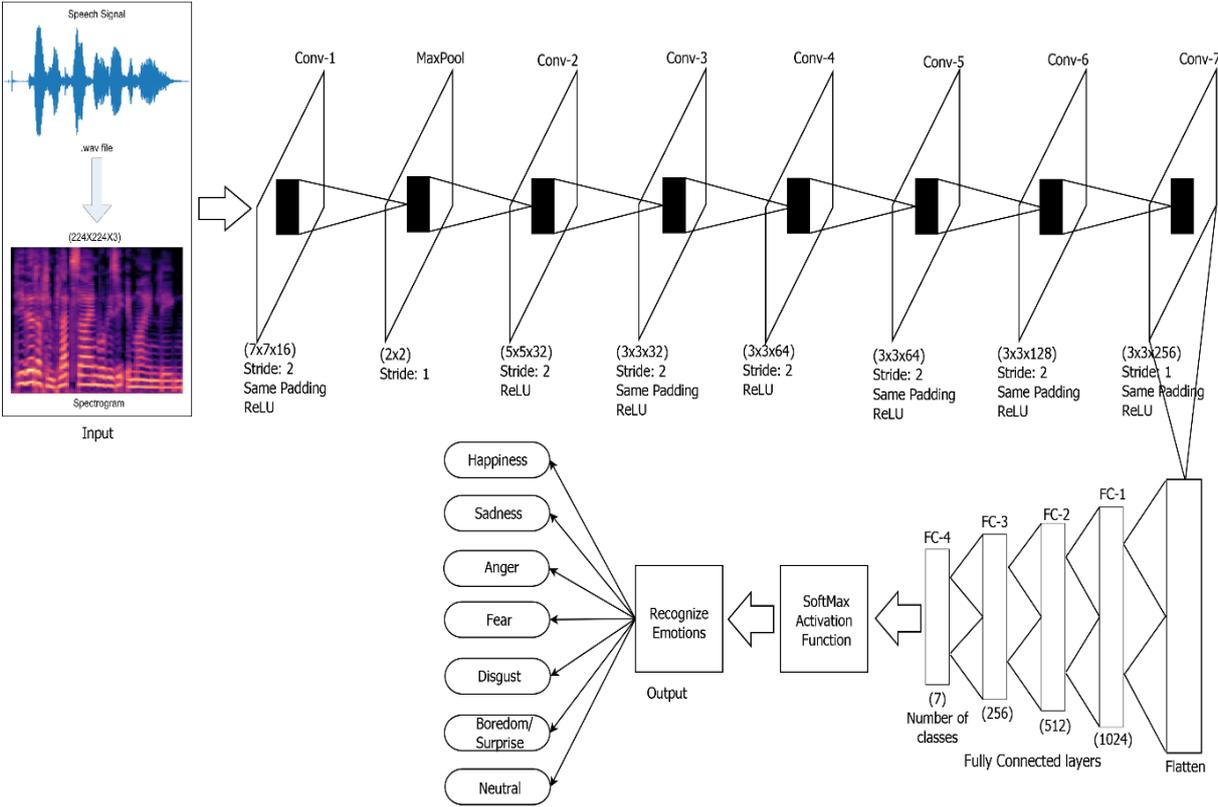


Figure 4.2: DCNN architecture details

The initial convolutional layers (Conv-1) in this DCNN architecture contain 16 filters of size (7x7) that are applied to the RGB spectrograms of size (224x224x3) with the stride (2x2) and same padding. Conv-1's output feature map is subjected to max-pooling (2x2) with stride (1x1). The second convolutional layer (Conv-2) comprises 32 filters of the same size (5x5) and stride (2x2). Conv-3 has the same number of filters of size (3x3), the same stride (2x2), and the same padding as Conv-2. Conv-4 and Conv-5 both include 64 filters of size (3x3) and stride (2x2), with the same padding in Conv-5. Conv-6 features 128 filters, while Conv-7 has 256 filters of the same size (3x3), stride (2x2), and same padding. ReLU activation function is utilised after each convolutional layer in this DCNN architecture to correct the output features map specified as: $f(z) = 0$, if $z \leq 0$ and $f(z) = z$, if $z \geq 0$ i.e. $f(z) = \max(0, z)$.

The corrected output layers are followed by batch normalisation with momentum 0.9 to regularise the DCNN model. A flatten layer follows the last convolutional layer, Conv-7. The completely linked layers get input from the flattening layer. The number of classes in the first fully connected layer (FC-1) is 1024, and the number of classes in the last fully connected layer (FC-4) is 7. Following FC-1, there is a 20% dropout rate. Finally, a softmax classifier is used to determine the likelihood of each emotion. Table 4.2 lists the major parameters utilised in the experiment during training the model.

Table 4.2: Main Parameters and their value

Parameter Name	Parameter value
Learning rate	Find optimal value during training
Momentum	0.9
Decay	1e-6
Eps	1e-5
solver type	SGD

4.2.3. Converting speech signal into spectrograms

To begin, acquire speech samples in the form of audio files from the speech corpora EMODB, SAVEE, and EMOVO. Read the .wav files from the speech corpus, then convert all of the files to spectrograms in step 2. Table 4.3 shows the steps for extracting spectrograms from a .wav file.

Table 4.3: Steps to get spectrograms

Input: .wav file from speech corpora
Output: spectrograms (image file)
Step-1: Divide the speech signal into 25 milliseconds. (Framing)
Step-2: Apply Hamming windowing
Step-3: Short Time Fourier Transform (STFT) of .wav files
Step-4: scale frequency axis logarithmically
Step-5: plot spectrograms

Frame size (25ms), overlapFac=0.5 (50 percent overlap), and window=np.hanning are the STFT settings (window type like Hamming, Kaiser, etc.). It is determined mathematically as shown in Equation (4.1).

$$Xm(w) = \sum_{n=-\infty}^{\infty} x(n)w(n - mR)e^{-jwn} \quad (4.1)$$

where $X_m(w)$ denotes the Discrete Time Fourier Transform (DTFT), $x(n)$ is the input signal at time n , and $w(n)$ denotes the window function. Figure 4.3 shows several example spectrograms of each speech corpus for each emotion, with the vertical axis representing frequency and the horizontal axis representing time.

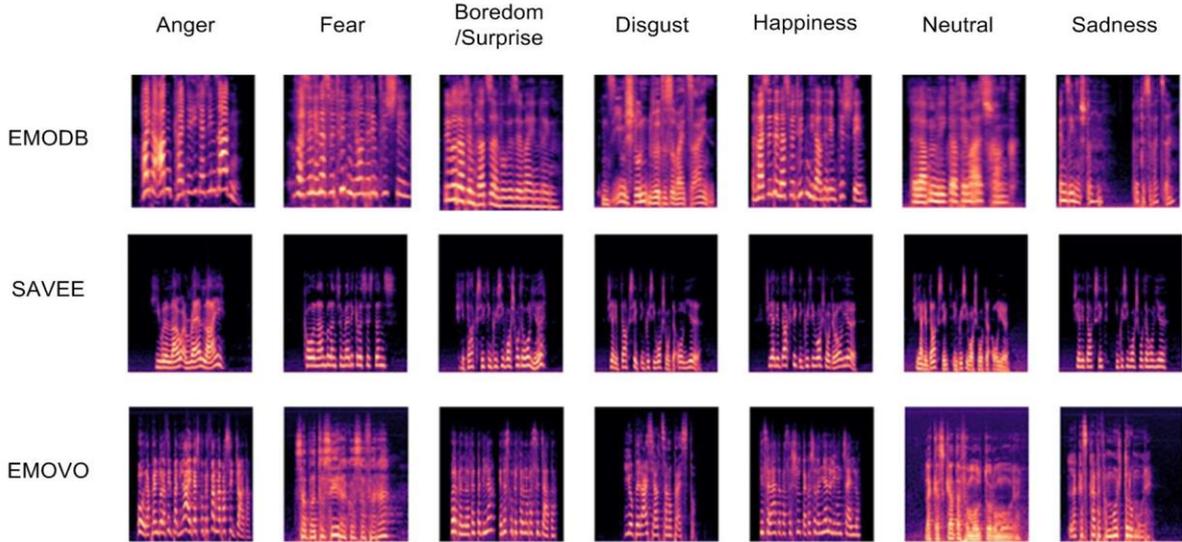


Figure 4.3: Spectrograms of each type of emotion for EMOVB, SAVEE, and EMOVO corpus.

All spectrograms are standardised to $224 \times 224 \times 3$ in step 3. Step 4 divides spectrograms into training, testing, and validation groups in the proportions of 80%, 10%, and 10%, respectively. The DCNN model is then trained and validated using the proposed technique before being stored (i.e. frozen) as stage-1 in step 5. The optimal learning rate is discovered in step-6 using the stage-1 model and the LR range Test. In step 7, the stage-1 model is unfrozen and trained with an ideal learning rate before being stored as a stage-2 model. Section 4.3 delves into the specifics of the trials and their outcomes.

4.3. Experimental details and results

In this part, the suggested method is tested on the EMOVB, EMOVO, and SAVEE datasets for emotion identification from speech. All of the tests were carried out on a normal Windows 10 laptop with an Intel(R) Core™ i5-7200U CPU running at 2.70 GHz and 8 GB of RAM, as well as an x64-based processor. The suggested algorithm's performance is compared to that of classifiers that use handmade features and modern deep learning techniques. In the next sections, we'll go through the specifics of the experiments.

4.3.1. Data sets

Experiments are carried out on three publicly accessible labelled, actor-based emotional data sets in this study: EMODB, EMOVO, and SAVEE. The following is a list of them:

- Anger, boredom, disgust, fear, happiness, sorrow, and neutral are the main seven emotions represented in the EMODB speech corpus. This corpus has 535 items in total, and the language of the corpus is German. Five male and five female actors record their utterances. In this piece, all ten utterances are used.
- SAVEE, a popular public emotional speech corpus recorded by four male performers, is another popular public emotional speech corpus. There are seven emotions represented in these data sets: happiness, sorrow, anger, fear, disgust, surprise, and neutral. There are 15 utterances and 480 samples in all.
- EMOVO is a freely accessible corpus of emotional speech. It divides emotions into seven categories: happiness, sorrow, anger, fear, disgust, surprise, and neutral. It comprises of 588 samples recorded by six performers, three men and three females. This corpus is written in Italian.

Table 4.4 summarises the statistics of various emotion speech corpora.

Table 4.4 Number of samples in speech corpora used

Database	Happy	Sad	Anger	Scared	Neutral	Disgust	Surprised	Bored	Total
EMODB	71	62	127	69	79	46	-	81	535
SAVEE	60	60	60	60	120	60	60	-	480
EMOVO	84	84	84	84	84	84	84	-	588

4.3.2. Experimental results

All considered speech corpora (EMODB, SAVEE, and EMOVO) are preserved as stage-1 once the proposed model has been trained. To demonstrate the efficiency of the suggested approach, the proposed model's prediction performance at stage-1 is assessed using EMODB, SAVEE, and EMOVO datasets. On the EMODB, SAVEE, and EMOVO datasets, Table 4.5a), 4.5b), and 4.5c) exhibit the prediction performance of the proposed model at stage-1 in terms of confusion matrix.

Table 4.5a) Confusion matrix for emotions prediction on EMODB at stage-1

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.82	0.08	0.00	0.00	0.00	0.10	0.00	
<i>Bored/ Surprise</i>	0.10	0.60	0.12	0.00	0.18	0.00	0.00	
<i>Disgust</i>	0.10	0.17	0.63	0.00	0.10	0.00	0.00	
<i>Happy</i>	0.00	0.08	0.02	0.74	0.16	0.00	0.00	
<i>Neutral</i>	0.01	0.00	0.00	0.00	0.75	0.15	0.09	
<i>Sad</i>	0.00	0.00	0.00	0.00	0.10	0.79	0.11	
<i>Scared</i>	0.09	0.00	0.00	0.00	0.01	0.12	0.78	
<i>Overall Accuracy</i>							73.08%	

Table 4.5b) Confusion matrix for emotions prediction on SAVEE at stage-1

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.60	0.10	0.10	0.00	0.10	0.10	0.00	
<i>Bored/ Surprise</i>	0.10	0.40	0.20	0.00	0.10	0.10	0.10	
<i>Disgust</i>	0.00	0.20	0.45	0.01	0.20	0.10	0.04	
<i>Happy</i>	0.00	0.18	0.02	0.50	0.20	0.00	0.10	
<i>Neutral</i>	0.10	0.10	0.10	0.10	0.49	0.02	0.09	
<i>Sad</i>	0.10	0.10	0.10	0.00	0.10	0.54	0.06	
<i>Scared</i>	0.20	0.10	0.10	0.00	0.02	0.06	0.52	
<i>Overall Accuracy</i>							50.00%	

Table 4.5c) Confusion matrix for emotions prediction on EMOVO at stage-1

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.67	0.00	0.03	0.00	0.10	0.20	0.00	
<i>ored/ Surprise</i>	0.01	0.49	0.10	0.00	0.30	0.10	0.00	
<i>Disgust</i>	0.10	0.09	0.51	0.00	0.20	0.10	0.00	
<i>Happy</i>	0.00	0.20	0.02	0.58	0.10	0.00	0.10	
<i>Neutral</i>	0.10	0.10	0.00	0.03	0.57	0.10	0.10	
<i>Sad</i>	0.20	0.00	0.00	0.00	0.10	0.61	0.09	
<i>Scared</i>	0.10	0.20	0.00	0.02	0.01	0.10	0.57	
<i>Overall Accuracy</i>							57.15%	

The overall accuracy of the proposed model at stage-1 for EMODB, SAVEE, and EMOVO is 73.08 percent, 50 percent, and 57.15 percent, respectively, as shown in Tables 4.5a), 4.5b), and

4.5c). The LR range test is used to find an appropriate learning rate. Figure 4.4 depicts learning rate variation in terms of the number of iterations and learning rate variation in terms of loss.

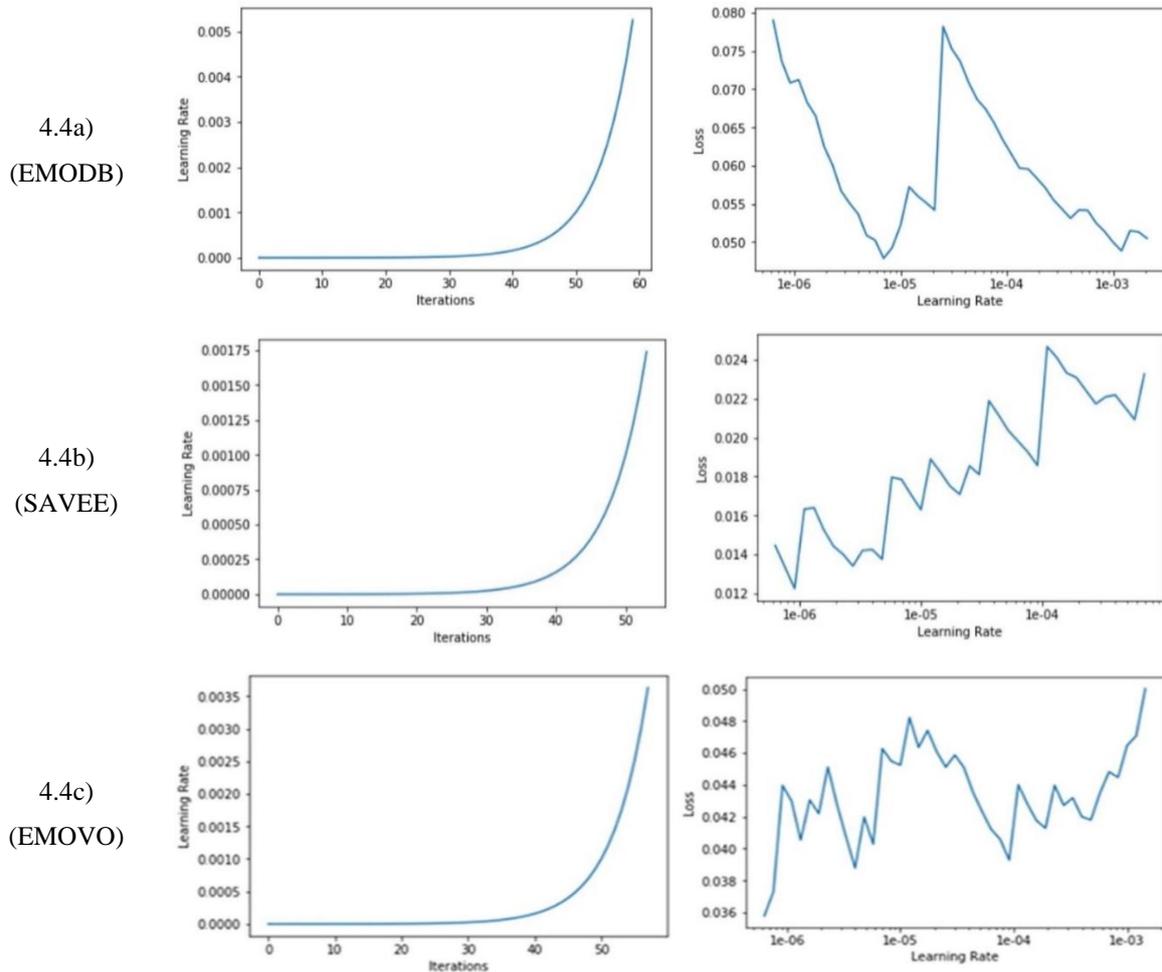


Figure 4.4: Learning rate variation for the number of iterations and variation in a loss concerning learning rate

The update of parameters is described by the learning rate. In this graph, the X-axis depicts what happens as learning rate is raised, while the Y-axis depicts the loss. Figure 4.4a) shows that after the learning rate exceeds 10^{-5} , fine-tuning is completed, which causes the loss to worsen. It is selected to pass an ideal learning rate of 10^{-5} using the learning rate finder (for EMOVB). According to Figure 4.4b), the loss is at a minimum at 10^{-6} and then increases. Therefore, the learning rate has above 10^{-6} . (for SAVEE). It can be shown in Figure 4.4c) that there was no range where the loss either rapidly dropped or thereafter grew worse. Thus, we have reached the 10^{-6} optimum learning rate, when loss is at its lowest (for EMOVO). The stage-1 model was unfrozen after determining the optimum learning rate. Additionally, the model was stored as stage-2 after being trained once again at the best learning rate. To demonstrate the effectiveness of the suggested approach, the prediction performance of the

proposed model at stag- 2 is assessed using the EMODB, SAVEE, and EMOVO datasets. The proposed model's stage-2 prediction performance on the EMODB, SAVEE, and EMOVO datasets is shown in Tables 4.6a), 4.6b), and 4.6c).

Table 4.6a): Confusion matrix for emotions prediction on EMODB at stage-2

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.93	0.00	0.00	0.00	0.03	0.04	0.00	
<i>Bored/ Surprise</i>	0.00	0.71	0.10	0.00	0.10	0.09	0.00	
<i>Disgust</i>	0.00	0.09	0.70	0.01	0.10	0.10	0.00	
<i>Happy</i>	0.00	0.00	0.02	0.88	0.10	0.00	0.00	
<i>Neutral</i>	0.00	0.00	0.00	0.00	0.91	0.00	0.09	
<i>Sad</i>	0.00	0.00	0.00	0.00	0.10	0.90	0.00	
<i>Scared</i>	0.00	0.00	0.00	0.00	0.01	0.10	0.89	
Overall Accuracy							84.62%	

Table 4.6b): Confusion matrix for emotions prediction on SAVEE at stage-2

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.85	0.15	0.00	0.00	0.00	0.00	0.00	
<i>Bored/ Surprise</i>	0.20	0.66	0.14	0.00	0.00	0.00	0.00	
<i>Disgust</i>	0.00	0.12	0.68	0.01	0.19	0.00	0.00	
<i>Happy</i>	0.00	0.00	0.10	0.75	0.15	0.00	0.00	
<i>Neutral</i>	0.00	0.13	0.10	0.00	0.77	0.00	0.00	
<i>Sad</i>	0.10	0.00	0.00	0.00	0.00	0.78	0.12	
<i>Scared</i>	0.10	0.00	0.00	0.00	0.00	0.14	0.76	
Overall Accuracy							75.00%	

Table 4.6c): Confusion matrix for emotions prediction on EMOVO at stage-2

<i>Emotional Class</i>	<i>Anger</i>	<i>Bored / Surprise</i>	<i>Disgust</i>	<i>Happy</i>	<i>Neutral</i>	<i>Sad</i>	<i>Scared</i>	
<i>Anger</i>	0.79	0.00	0.00	0.00	0.10	0.11	0.00	
<i>Bored/ Surprise</i>	0.10	0.61	0.19	0.00	0.10	0.00	0.00	
<i>Disgust</i>	0.00	0.12	0.62	0.16	0.00	0.10	0.00	
<i>Happy</i>	0.00	0.08	0.12	0.70	0.10	0.00	0.00	
<i>Neutral</i>	0.10	0.18	0.00	0.00	0.72	0.00	0.00	
<i>Sad</i>	0.10	0.00	0.00	0.00	0.16	0.74	0.00	
<i>Scared</i>	0.19	0.00	0.00	0.00	0.01	0.11	0.69	
Overall Accuracy							69.65%	

The overall accuracy of the proposed model at stage 2 for EMOVB, SAVEE, and EMOVO is shown in Tables 4.6a), 4.6b), and 4.6c) to be 84.62 percent, 75 percent, and 69.65 percent, respectively. This also demonstrates that improved accuracy has been attained in stage-2 as compared to the stage-1 model. The summary results of the suggested Model at stage-1 and stage-2 (Mean + Standard deviation) is presented in Table 4.7.

Table 4.7: Accuracies at stage-1 and stage-2

Speech Corpus	Average Accuracy (%)		Accuracy (%) gain at stage-2 as compared to stage-1
	Stage-1	Stage-2	
EMODB	73.08± 6.23	84.62± 3.32	11.56
SAVEE	50.00± 4.86	75.00± 2.01	25.00
EMOVO	57.15± 5.23	69.65± 2.54	12.50

The findings indicate that stage 2 accuracy is on average 16.35 percent higher than stage 1 accuracy. The greatest improvement is for SAVEE corpus (25 percent). The findings of the suggested method lead to the conclusion that, when compared to a random learning rate, choosing an ideal learning rate is crucial.

4.3.3. Performance comparison of the proposed algorithm with state-of-the-art

The suggested algorithm outperforms the previous result over the EMOVB, SAVEE, and EMOVO datasets utilising spectrograms as input, as shown in Table 4.8

Table 4.8: Performance comparison of the proposed algorithm with state-of-the-art methods

	Methods	Input	Average Accuracy (%)		
			EMODB	SAVEE	EMOVO
Özseven T [141]	SVM	Features	71.12	72.39	60.40
	K-NN	extracted with	63.74	53.37	39.05
	MLP	openSMILE	81.32	71.17	58.58
Badshah et al. [142]	Random Forest	MFCC	77.18	-	-
	Alexnet		81.33	-	-
	Decision Tree		72.82	-	-
Chen et al. [98]	3D CRNN	Mel-spectrograph	82.82	-	-
Huang et al. [131]	CNN	TEO	84.50	69.00	-
Fayek et al. [127]	DNN	MFCC	-	59.7	-

Parry [88]	CNN-LSTM	GeMAPS feature set	69.72	72.66	53.24
Lee et al. [87]	GoogLeNet	Spectrogram	72.55	-	-
Daneshfar et al. [143]	pQPSO(GMM)	MFCC+LPCC+ PMVDR+Pitch	82.82	60.79	-
Proposed Algorithm	Stage-1	Spectrogram	73.08	50.00	57.15
	Stage-2	Spectrogram	84.62	75.00	69.65

Table 4.8 shows that the proposed algorithm stage-2 performs better on the EMODB corpus (84.62 percent accuracy) than SVM (71.12 percent), K-NN (63.74 percent), and MLP (81.32 percent) [141], Random forest (77.18%), Alexnet (81.33%), and Decision Tree (72.82%); [142], 3D CRNN (82.82%); [98], CNN (84.50%); [131], CNN-LSTM (69.72%); [87], GoogLeNet (72.55%); and pQPSO (82.82%) [143]. For SAVEE corpus, the proposed algorithm stage-2 gives better results (75%) accuracy as compared to SVM (72.39%), K-NN (53.37%), and MLP (71.17%) [141], CNN (69.00%) [131], DNN (59.70%) [127], CNN-LSTM (72.66%) [88], and pQPSO (60.79%) [143]. And for EMOVO corpus, the proposed algorithm stage-2 gives better results (69.65%) accuracy as compared to SVM (60.40%), k-NN (39.05.37%), and MLP (58.58%) [141], CNN-LSTM (53.24%) [88].

4.4. Conclusion

This chapter proposes an algorithm that can extract high-level characteristics from raw data with better accuracy, regardless of language or the gender of speakers in voice corpora. DCNN model is trained in two steps. The Learning Rate(LR) range test is used in stage 1 to determine the optimal learning rate, and in stage 2 the model has been retrained using the optimal learning rate. The proposed model is tested on three major public speech corpora EMODB (German), EMOVO (Italian), and SAVEE (British English) with diverse languages. In comparison to previous research for speakers of all languages, the stated accuracy is improved.

CHAPTER 5

LIGHTWEIGHT CNN BASED ALGORITHM WITH NEW INDIAN EMOTIONAL SPEECH CORPORA

5.1. Overview

Many applications, including human-computer interaction, online teaching, healthcare, determining the emotional state of the consumer in contact centres, and many more, depend on the ability to recognise human emotion. To create intelligent devices that can recognise a speaker's real emotional state, many researchers are engaged in this field of study. The ascension of the speaker is one of the key elements that represents the strength and volume of the speech signal for emotion identification from speech [144]. Numerous languages, including German, Italian, Hindi, Telugu, American English, and many more, are supported by the current databases. There are just a few English-language databases, including eNTERFACE and RAVDESS. There is no public access to the English-language database in the Indian Ascent. In this study, 600 samples of emotional speech from 8 Indian speakers were used to generate the Indian Emotional Speech Corpora (IESC) database (5 males and 3 females).

The extraction of crucial low-level and high-level characteristics for Speech Emotion Recognition (SER) is a difficult challenge for investigators. In the fields of emotion detection and human behaviour recognition, deep learning and machine learning are more popular.

In order to increase the precision of emotion identification, researchers are attempting to identify robust characteristics using deep learning techniques [145]. For recognising emotions in voice, a variety of deep learning algorithms are utilised, including Convolutional Neural Networks (CNN), Long Short Term Memory (LSTM), Recurrent Neural Networks (RNN), Radial Basis Function (RBF), Deep Neural Networks (DNN), Deep Belief Networks (DBN), and others [5], [146], [147]. The extraction of more information from speech signals to improve accuracy is one of the primary research issues in emotion identification. Furthermore, it is unclear whether characteristics are more helpful in identifying emotional states. The second difficulty is to lower model computing costs since deep learning models' computation costs rise as the number of layers grows.

According to the literature, numerous researchers have employed CNN from deep learning techniques to identify emotions in speech. Due to the huge pre-trained architecture, using deep learning algorithms improves accuracy but also raises the computing cost of the model. To address the issues, this chapter proposes a Lightweight CNN-based deep learning model for emotion identification. As the spectrogram representation of the speech signal includes more information than the audio speech signal, the proposed CNN-based model leverages spectrograms of the speech signals to get more features with fewer layers. Stride is used to down sample the feature map and many features, which lowers the computational cost compared to pooling layers. When the suggested model has been utilised with publicly accessible databases in various accents and languages, such as EMOVO, EMODB, and SAVEE, improved accuracy of emotion identification has been recorded. The suggested model's performance is contrasted with that of the current SER CNN-assisted model [78], ResNet-18, and ResNet-34. The primary contributions to this chapter are:

- a new emotional corpus IESC is created using 600 samples of Indian people speaking English with five different emotions (5 males and 3 females). Using Google forms, a subjective test of emotion recognition is used to verify IESC. After hearing the speech samples for each sample, the user is prompted to select one of the potential emotions (such as anger, fear, happiness, neutrality, or sadness). More than 20 individuals have verified this database.
- For recognising emotions from voice signal spectrograms, a unique CNN-based deep learning model is also proposed. Instead of pooling layers, this model down samples the feature maps using five CNN layers and stride. According to the experiment, it lowers the computational cost and raises average accuracy. IESC, as well as the openly accessible datasets EMOVO, EMODB, and SAVEE, are used to assess the proposed model.
- The suggested model's average accuracy is also contrasted with that of the reference ResNet-18 and ResNet-34 CNN-based architectures. In order to compare performance with the suggested model, the CNN-Assisted SER model is also implemented and assessed on IESC.

The following sections make up the remaining chapter. The proposed model is discussed in section 5.2. The specifics of the experiment and its findings are explained in section 5.3. Section 5.4 has covers the conclusion.

5.2. Proposed Model

This model's initial phase involves reading audio files and dividing speech signals into frames known as windowed segments. Equation 5.1 provides the duration of each frame.

$$df = \frac{1}{Sr} \cdot K \quad (5.1)$$

where df denotes the length of a frame, Sr is the length of a single sample, and K denotes a sample (i.e. frame size). Then, each frame from equation 5.2 is subjected to the window function $w(k)$.

$$w(k) = 0.5 \left(1 - \cos\left(\frac{2\pi k}{K-1}\right) \right), k = 1, 2, \dots, K \quad (5.2)$$

Here, the 20ms of each frame and 50% of the overlapping frame are taken into account to lessen the loss signal at the frame's termination. Then, using equation 5.3, the Hann window is applied to the initial signals.

$$Sw(k) = S(k) \cdot w(k), k = 1, 2, \dots, K \quad (5.3)$$

where $S(k)$ is the speech signal and $Sw(k)$ is the Hann window function. Then, using equation 5.4, discrete Fourier transform (DFT) is applied to windowed data segments, and the coefficients are shown as a function of frequency and time.

$$X[n, f] = \sum_{m=0}^n x[n + m]w[m]e^{-jfm} \quad (5.4)$$

Here, n stands for time, f for frequency, and $X[n, f]$ for the DFT function. The square magnitude of the DFT of the speech signal is next calculated in order to obtain the spectrogram, as illustrated in equation 5.5.

$$spectrogram(n, f) = |X[n, f]|^2 \quad (5.5)$$

Speech signals are transformed into spectrograms, which are then normalised to 224x224 pixels and used as input for the first layers of the CNN. Figure 5.1 depicts the conversion of the spoken stream to a spectrogram.

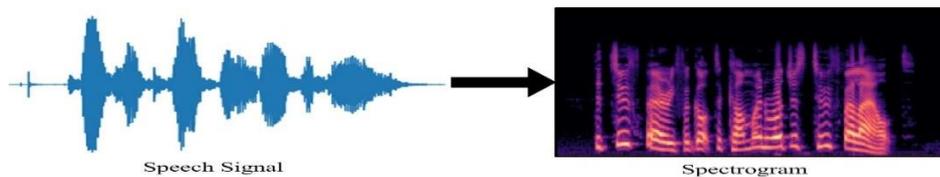


Figure 5.1: Conversion of the speech signal to spectrogram

Consider the spectrograms as an input with a size of 224x224, and 3 filters with a size of fxf are used to extract the features. Each filter will be applied to the three colours as spectrograms are RGB. To extract the pertinent feature map, the filter is convolved with the spectrograms. Equation 5.6 is used to calculate the convolution operation.

$$conv(I^l, K)_{x,y} = ReLU^{l-1} \left(\sum_{\substack{1 \leq i, j \leq n \\ 1 \leq k \leq 3}} K_{i,j,k} I_{x+i-1, y+j-1, k}^l \right) \quad (5.6)$$

where K is the filter of size fxf, I^l is the image at layer l, i and j are the spectrogram's width and height, respectively, and k is the channel count. After the convolution procedure, the ReLU activation function is employed to correct the feature map. Equation 5.7 provides a mathematical description of the ReLU activation function.

$$ReLU(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } x > 0 \end{cases} \quad (5.7)$$

Equation 5.8 provides a mathematical description of the dimension of the feature maps after the convolution process has been applied. Zero paddings are taken into consideration at all layers, features are employed in place of pooling layers, and stride is used for down sampling.

$$dim(conv(I^l, K)_{x,y}) = \left(\frac{n^l - f^l}{s^l} + 1, \frac{n^l - f^l}{s^l} + 1 \right) \quad (5.8)$$

where f^l represents the layer l filter size, s^l represents the layer l stride size, and n^l represents the layer l spectrogram size. $l=1,2,\dots,5$ do the same process for each CNN layer. The next step is to use a sigmoid function to determine the emotion from the output of the final convolutional layer, which has been flattened to generate a 1-dimensional array as input. Equation 5.9 gives a mathematical description of the sigmoid function.

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \quad (5.9)$$

From the spectrograms, the suggested model automatically picks out useful information. The identification rate increases when spectrograms contain more essential information. Convolutional, stride, and fully linked layers make up the suggested model. Fully linked layers are supplied to a soft-max function to determine the likelihood of each emotion. Figure 5.2 details the suggested CNN-based model in its entirety.

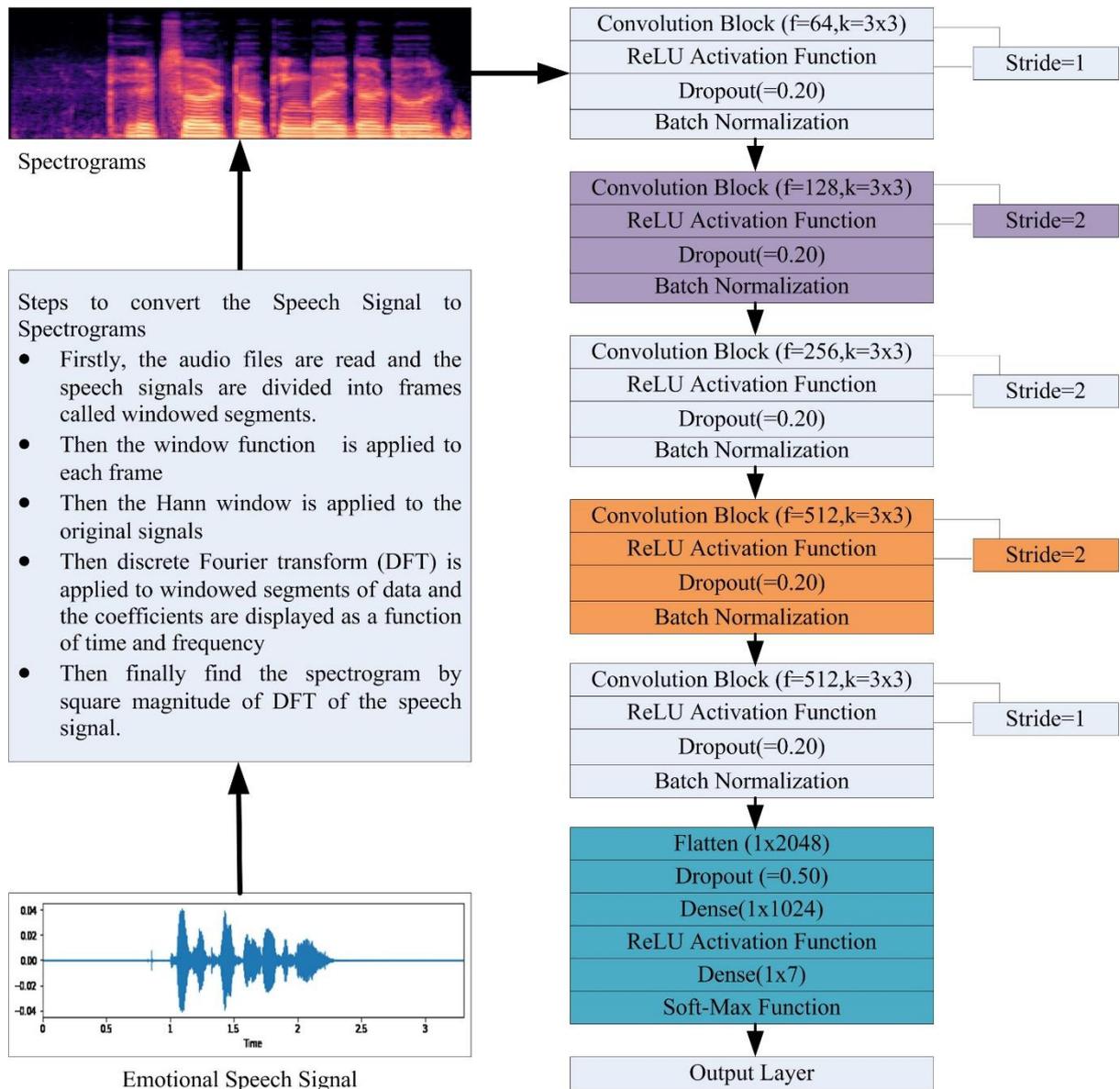


Figure 5.2: Proposed CNN-based model to recognize emotions from speech.

The first convolutional layer (CNN-1) in the proposed model, which has 64 filters of size (3x3) with stride (1x1), is assumed to be the input layer using normalised RGB spectrograms. The second convolutional layer (CNN-2) of CNN-1, which comprises 128 filters of size (3x3) with stride (2x2), receives the feature map obtained from CNN-1. The third convolutional layer (CNN-3) of CNN-2, which includes 256 filters of size (3x3) with stride (2x2), receives the output feature map from CNN-2. Similar to the third convolutional layer, CNN-3, CNN-4 includes 512 filters of size (3x3) with stride (2x2) taken into account as input of feature map. The fifth convolutional layer (CNN-5), which includes filters of size (3x3) with stride (1x1), receives the feature map obtained from CNN-4. Following each convolutional layer in the suggested model, the output features map is corrected using the ReLU activation function. The

output is regularised using batch normalisation after each CNN layer. The output of the final convolutional layer CNN-5 is flattened to generate a 1-dimensional array, which is then supplied as input to a fully connected layer, followed by a 50% dropout ratio to reduce validation loss. Once more, the validation loss is minimised by using a dropout ratio of 50%. The likelihood of each emotion is then determined using the soft-max function.

5.3. Experimental Details

This section discusses the experimental specifics of the suggested model as well as information on the built emotional database IESC. Following a brief discussion of the three databases utilised, EMOVO, EMODB, and SAVEE, the implementation details for the suggested model are described. With regard to metrics like accuracy, precision, recall, and F1-score, the experimental findings of the proposed model assessed on the IESC and the three databases are reviewed. On a typical Windows 10 laptop with an x64-based Intel(R) Core™ i5-7200U CPU running at 2.70 GHz and 8 GB of RAM, all experiments are conducted.

5.3.1. Indian Emotional Speech Corpora (IESC)

Indian Emotional Speech Corpora (IESC) is the name given to the emotional speech database that eight regular individuals (5 men and 3 women) generated. It comprises of 600 emotional audio recordings. The five emotions in the IESC database's audio recordings are neutral, happy, angry, sad, and terrified. To eliminate background noise, each audio file is captured using a speech recorder app on a smartphone in a quiet environment. For sound leakage and noise cancellation while recording, headphones are also utilised in conjunction with a microphone. The .wav file extension is used to store all audio recordings. For the purpose of capturing various emotions, two English lines extracted from the RAVDESS database are utilised. Sentences 1 and 2 are "Kids are talking by the door" and "Dogs are sitting by the door," respectively. In this database, four speakers each recorded ten and four speakers each recorded twenty samples of each emotion. Therefore, there are 600 audio files overall, or $(10 \times 5 \times 4 + 20 \times 5 \times 4)$. Table 5.1 includes information on the speakers as well as database statistics.

Table 5.1: Speaker details and statistics of IESC database with nationality as Indian

Speaker ID	Speaker information		Samples Distribution					
	Age	Gender	Neutral	Happy	Anger	Sad	Fear	Total
S1	24	Male	10	10	10	10	10	50
S2	24	Male	10	10	10	10	10	50
S3	24	Male	10	10	10	10	10	50
S4	24	Female	10	10	10	10	10	50
S5	34	Female	20	20	20	20	20	100
S6	35	Male	20	20	20	20	20	100
S7	23	Male	20	20	20	20	20	100
S8	23	Female	20	20	20	20	20	100
Total Samples								600

where the individual file name is used to save each audio file. Each audio file has a file name made up of four alphanumeric components for unique identification, such as "H-4-5-1.wav," where each component is specified as follows:

- First part represents the emotions (A= angry, F = fear, H = happy, N = neutral, S = sad).
- Second part shows the repetition (1 = 1st repetition, 2 = 2nd repetition and so on)
- The third part represents the speaker (1=1st Speaker, 2=2nd Speaker, and so on)
- And the last part represents the sentence (1 = "Kids are talking by the door", 2 = "Dogs are sitting by the door").

IESC database is made publicly accessible by being published to the Kaggle website. A subjective test of emotion recognition using Google forms is used to assess the validity of the samples in IESC. In this test, a total of 60 speech samples—or 10% of the total samples—are taken into account for validation. The user is given five options for each sample, including anger, fear, happiness, neutrality, and sadness. Each user is expected to identify one of the potential emotions after hearing the voice samples. More than 20 persons participated in the database validation. Equation 5.10 contains a formula that may be used to calculate each emotion's accuracy.

$$Accuracy_{emotions} = \frac{1}{NS} \sum_{i=1}^{NS} \left(\frac{NR-CA}{NR} \right)_i \quad (5.10)$$

where NR is the total number of replies, CA is the number of right answers for each emotion, and NS is the number of samples taken into account for each emotion. Equation 5.11 mathematical statement is used to compute the average accuracy.

$$\text{Average Accuracy} = \frac{1}{N} \sum_{k=1}^N (\text{Accuracy}_{\text{emotions}})_k \quad (5.11)$$

where N represents all possible emotions. The average degree of accuracy discovered is 89.96%. Figure 5.3 displays the IESC database's validation outcome.

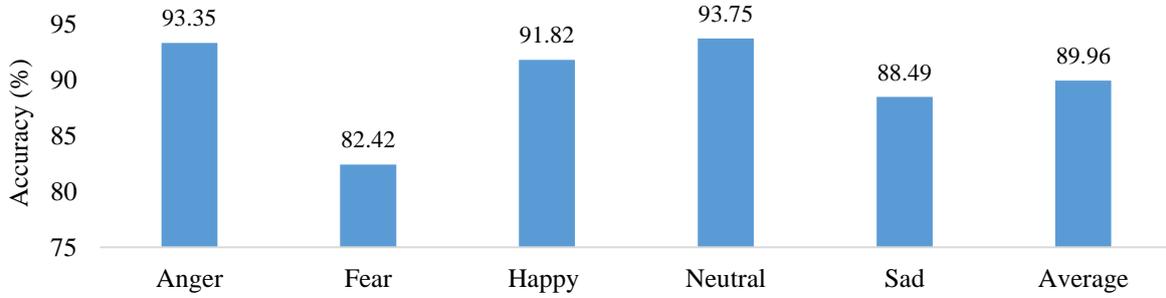


Figure 5.3: Validation result of database IESC

5.3.2. Publicly Available Datasets

Seven kinds of emotions such as anger, fear, happiness, neutral, sorrow, boredom/surprise, and disgust—are preserved as audio recordings in the publicly accessible labelled datasets EMOVO, EMODB, and SAVEE. Table 5.2 provides a summary of these databases specifics as well as the overall amount of samples for each emotion.

Table 5.2: Statistics of Publicly available database EMOVO, EMODB, and SAVEE

Database	Statistics of speech corpora				Samples in each emotion								
	Language	Total	Actor	Total Utterances	Anger	Fear	Happy	Neutral	Sad	Bored	Surprised	Disgusted	Total
EMOVO	Italian	6 (Male-3 Female-3)		14	84	84	84	84	84	-	84	84	588
EMODB	German	10 (Male-5 Female-5)		10	127	69	71	79	62	81	-	46	535
SAVEE	British English	4 (Male)		15	60	60	60	120	60	-	60	60	480

5.3.3. Implementation details and results

The implementation's specifics are covered in this section. First, the speech corpora IESC, EMODB, SAVEE, and EMOVO are used to gather the emotional speech data set. Next, the audio files are transformed into 224x224x3 spectrograms. Figure 5.4 displays sample spectrograms of each emotion for each speech corpus, with the horizontal axis denoting the emotion and the vertical axis denoting the speech corpus.

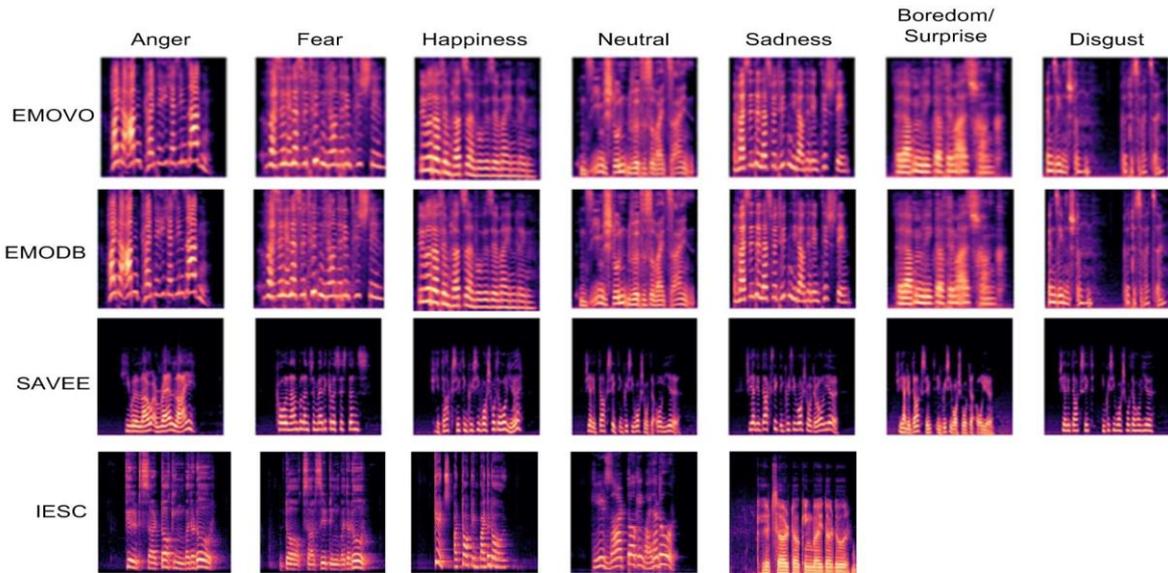


Figure 5.4: Sample spectrograms of each emotion for EMOVO, EMODB, SAVEE, and IESC speech corpus.

The voice samples are transformed into spectrograms and then provided as input to the proposed CNN-based model. Here, the percentages for training, testing, and validation are, respectively, 80%, 10%, and 10%. All of the speech corpora, including EMODB, SAVEE, EMOVO, and IESC, have been trained for the proposed model. Learning rate 10^{-5} , momentum 0.9, decay 10^{-6} , and $\text{eps } 1e^{-5}$ are the key parameters and their values are taken into consideration. 500 training epochs are used to perfect the suggested model. A speaker-independent evaluation of the proposed model's performance is done for the EMODB, SAVEE, EMOVO, and IESC databases. For the purpose of contrasting the performance of the suggested model, a CNN-Assisted state-of-the-art SER model is also assessed on the IESC. In Figure 5.5, training and testing accuracy charts for the proposed model and CNN-Assisted model are shown in relation to the number of epochs on the IESC database.

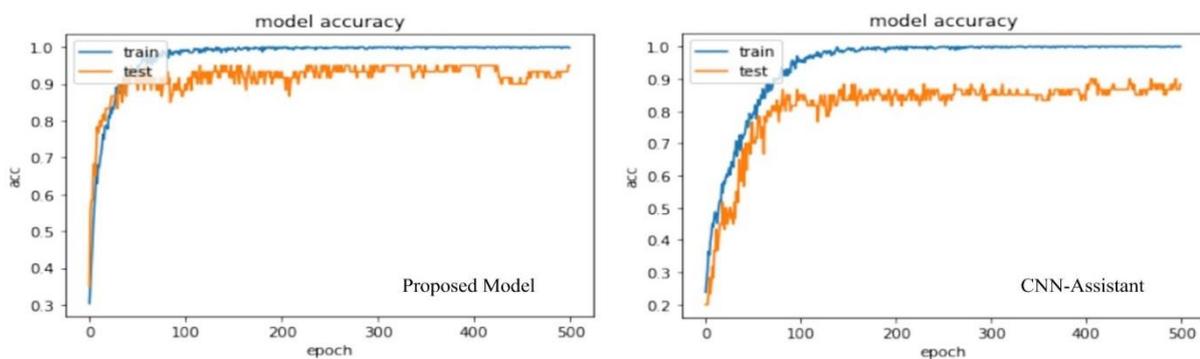


Figure 5.5: Proposed Model and CNN-Assisted model training and testing accuracy plot concerning the number of epochs on IESC database

In terms of class level precision, recall, and F1-score, the prediction performance of the proposed model and the CNN-Assisted model assessed on the IESC database is presented in Table 5.3, and accuracy is displayed in Figure 5.6.

Table 5.3: Classification report of the proposed model and SER CNN-Assisted Model [78] on IESC

Emotion	Proposed Model			CNN-Assisted Model		
	Precision	Recall	f1-score	Precision	Recall	f1-score
Anger	1.00	1.00	1.00	0.90	0.90	0.90
Fear	0.93	0.93	0.93	0.92	0.86	0.89
Happy	1.00	1.00	1.00	0.92	0.92	0.92
Neutral	0.92	0.92	0.92	0.92	0.92	0.92
Sad	0.91	0.91	0.91	0.75	0.82	0.78
Average			0.95			0.88

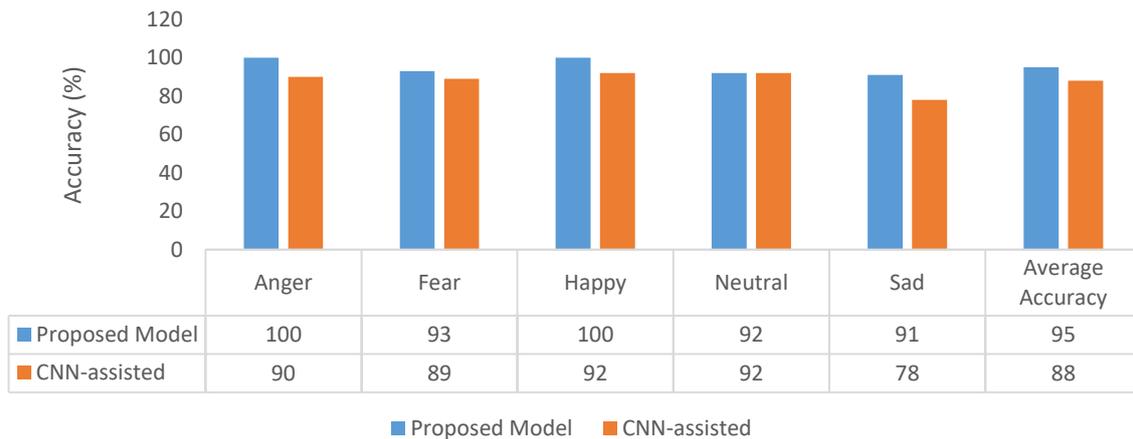


Figure 5.6: Performance comparison of the proposed model with CNN-Assisted on IESC database

Figure 5.6 makes it evident that the suggested model outperforms the CNN-Assisted SER model in terms of accuracy at the class level and by 7% on average. Figure 5.7 displays the proposed model's training and testing accuracy for the number of epochs for the EMOVO, EMOVB, and SAVEE corpora.

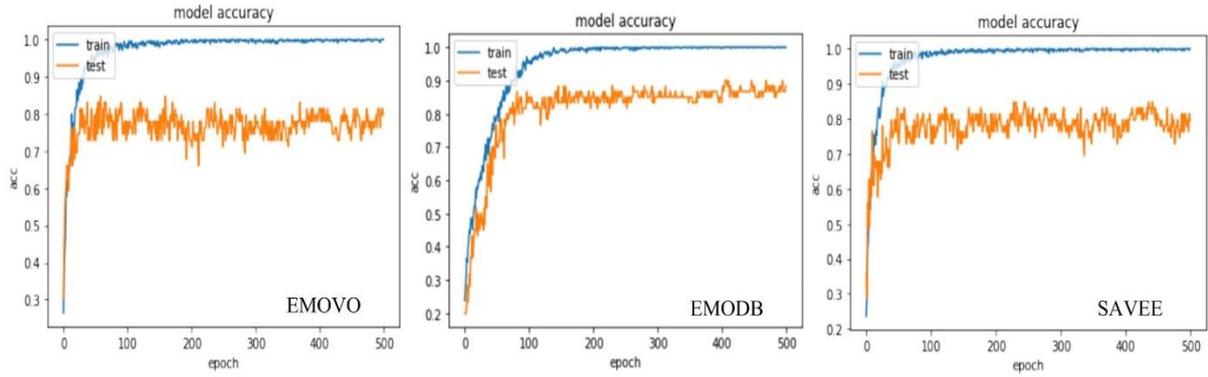


Figure 5.7: Proposed Model training and testing accuracy variation concerning the number of epochs on EMOVO, EMODB, and SAVEE databases

Table 5.4 presents the proposed model's training, testing, and validation average accuracy results on the EMODB, EMOVO, SAVEE, and IESC databases.

Table 5.4: The Experimental Results of the proposed Model

Speech Corpora	Average Accuracy (%)		
	Training	Validation	Testing
EMOVO	100	81.00	81.00
EMODB	99.58	87.04	87.04
SAVEE	100	80.00	80.00
IESC	99.81	95.00	95.00

Table 5.4 makes it evident that the suggested model testing accuracy for EMOVO, EMODB, SAVEE, and IESC database is 81 percent, 87.04 percent, 80 percent, and 95 percent, respectively. It is evident from Fig. 4 and Table 5 that the suggested model works better with the developed database IESC. For the EMODB, EMOVO, and SAVEE databases, Table 5.5 lists the class level precision, recall, and F1-score average accuracy of the proposed model.

Table 5.5: Performance of the proposed model on EMOVO, EMODB, and SAVEE databases

Emotion	EMOVO			EMODB			SAVEE		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Anger	0.88	0.88	0.88	1.00	0.94	0.97	0.75	1.00	0.86
Fear	0.60	1.00	0.75	1.00	1.00	1.00	0.88	0.70	0.78
Happy	0.91	1.00	0.95	1.00	0.83	0.91	1.00	0.90	0.95
Neutral	0.67	1.00	0.80	0.60	1.00	0.75	0.78	0.70	0.74
Sad	0.88	0.88	0.88	0.88	0.64	0.74	0.67	1.00	0.80
Boredom/	0.88	0.70	0.78	1.00	0.80	0.89	0.88	0.64	0.74

Surprise									
Disgust	0.88	0.88	0.88	0.75	1.00	0.86	0.88	0.88	0.88
Average			0.81			0.87			80

CNNs with ResNet-18 and ResNet-34 have 18 and 34 layers, respectively [148]. Figures 5.8a, 5.8b, 5.8c, and 5.8d, respectively, illustrate the confusion matrices of the ResNet-18 and ResNet-34 implementations with EMOVO, EMOVB, SAVEE, and IESC.

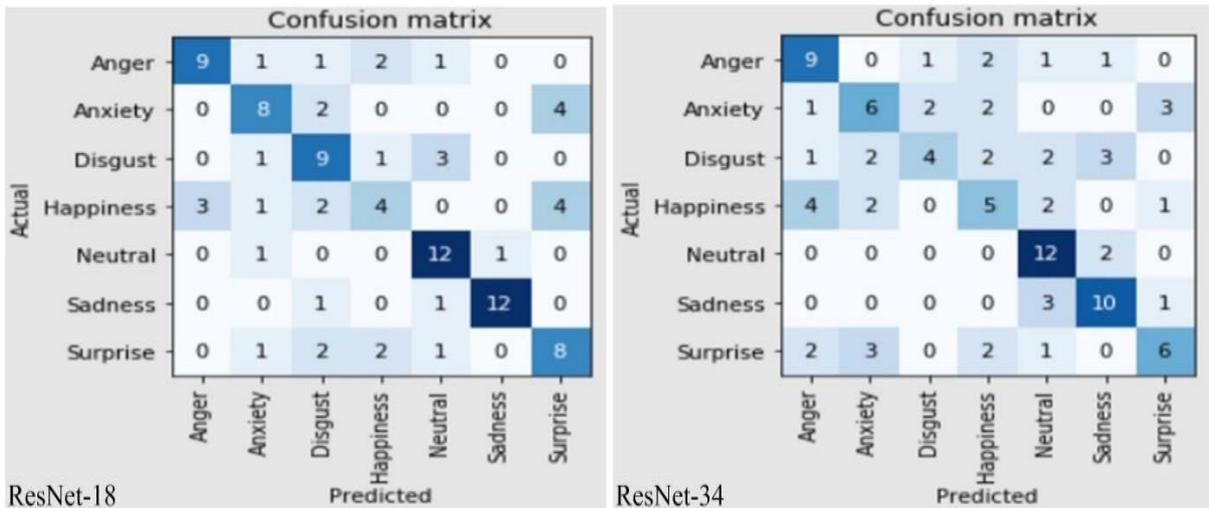


Figure 5.8a Confusion matrix of ResNet-18 and ResNet-34 on database EMOVO

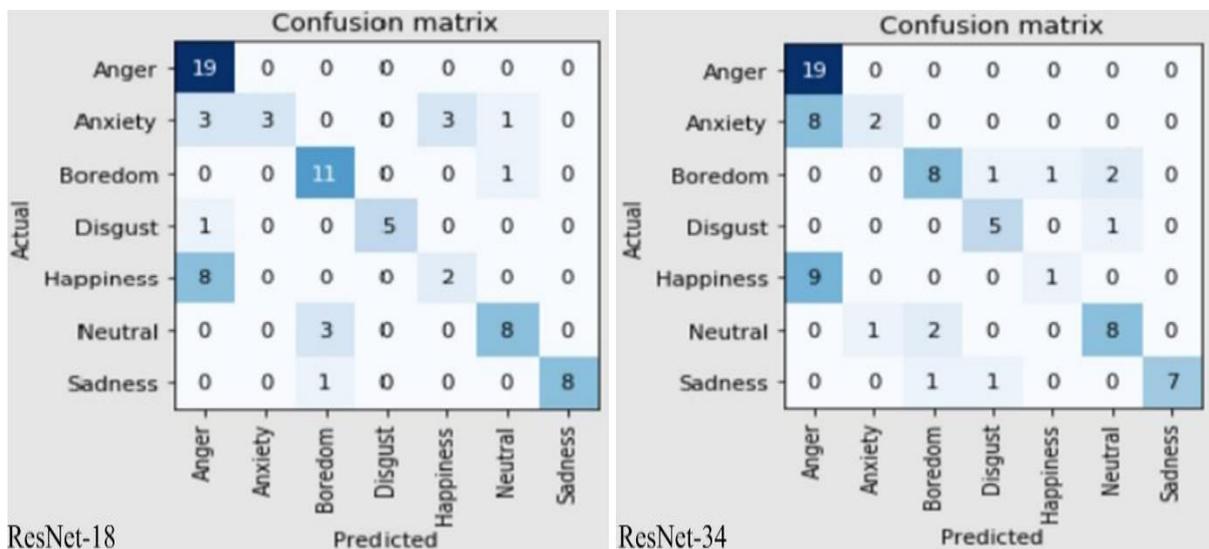


Figure 5.8b Confusion matrix of ResNet-18 and ResNet-34 on database EMOVB

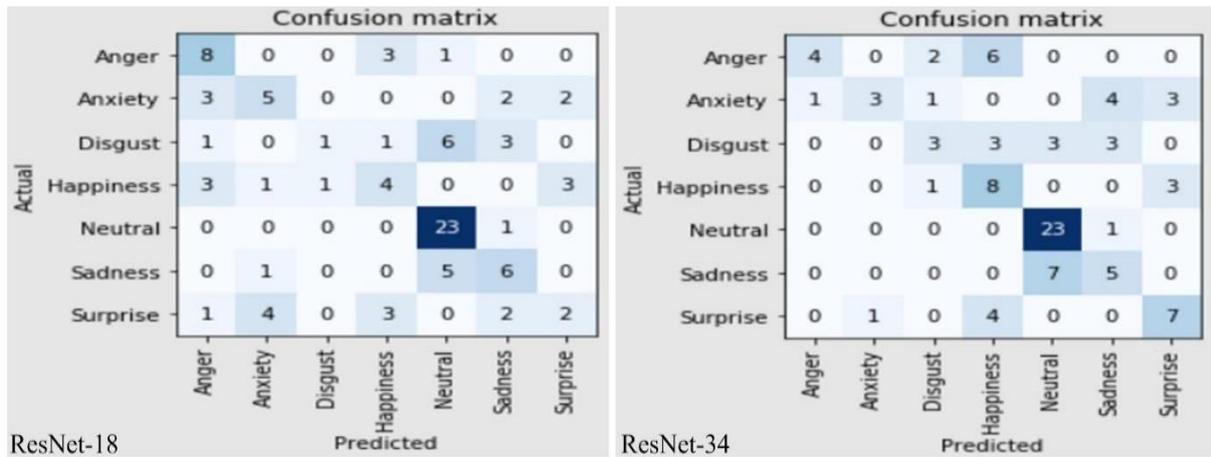


Figure 5.8c: Confusion matrix of ResNet-18 and ResNet-34 on database SAVEE

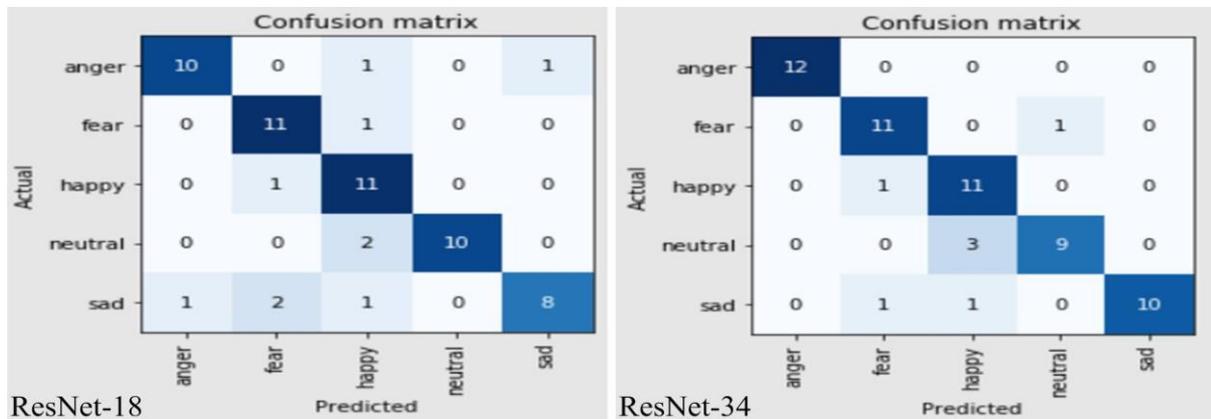


Figure 5.8d: Confusion matrix of ResNet-18 and ResNet-34 on database IESC

The accuracy of the suggested model is higher than that of the benchmark ResNet-18 and ResNet-34 deep learning models. The suggested model is also more efficient than ResNet-18 and ResNet-34 in terms of training time and model size. The suggested model's performance against other models is shown in Table 5.6.

5.3.4. Performance comparison of the proposed algorithm

According to Table 5.6 performance comparison with other state-of-the-art techniques, the recommended model outperforms the prior finding on the EMOVB, SAVEE, and EMOVO datasets using spectrograms as input.

Table 5.6 Comparison of the proposed model with the state-of-the-art SER approaches

	Method	Input	Average Accuracy (%)			
			EMOVO	EMOVB	SAVEE	IESC
Meng H. et al. [149]	RNN	MFCC	-	82.62%	-	-

Liu Zayene et al. [150]	CNN + LSTM	Spectrograms	65.80	84.22	73.83	-
Li S. et al. [68]	CNN	spectrograms	-	83.30	56.50	-
Ancilin et al. [53]	Linear Kernel based SVM	MFMC	73.30	81.50	75.63	-
ResNet-18[148]	CNN	Spectrogram	63.26	72.72	51.04	83.33
ResNet-34[148]	CNN	Spectrogram	53.06	64.93	55.20	88.33
Proposed Model	CNN	spectrogram	81	87	80	95

From Table 5.6, for EMOVO database the proposed model gives better accuracy (81%) as compared to CNN with LSTM (65.80%) [150], linear kernel based SVM (73.30 %) [151], ResNet-18 (63.27%) and ResNet-34 (53.06%). For EMODB database proposed model gives accuracy (87%) which is better than RNN (82.62%) [149], CNN with LSTM (84.22 %) [150], CNN (83.30%) [68], linear kernel SVM (81.50%) [153], ResNet-18 (72.72%) and ResNet-34 (64.93%). For SAVEE database the proposed model's average accuracy (80%) is better as compared to CNN with LSTM (73.83%) [150], CNN (56.50 %) [68], linear kernel based SVM (75.63%) [151], ResNet-18 (51.04%) and ResNet-34 (55.20%). For IESC database proposed model give 95% accuracy which is better than the ResNet-18 (83.33%) and ResNet-34 (88.33%).

5.4. Conclusion

In this chapter, an English emotional speech database (IESC) made up of 8 speakers from north India is established. More than 20 persons manually validate the produced database, and the claimed average accuracy of the validation is 89.96 percent. Convolutional neural network (CNN)-based deep learning model is suggested. The suggested model outperforms the SER CNN-Assisted model by 7% and achieves an average accuracy of 95% on the IESC database (average accuracy- 88 percent). The suggested model outperforms state-of-the-art SER techniques with an average accuracy of 81 percent, 87 percent, and 80 percent on the EMOVO, EMODB, and SAVEE datasets, respectively. When compared to the other basic deep learning models ResNet-18 and ResNet-34 for prediction, the suggested model performs better.

CHAPTER 6

1D CNN BASED ALGORITHM USING MFCC FEATURES

6.1. Overview

The study's contribution is that it proposes a 1D CNN-based technique in which features from audio files are extracted using the state-of-the-art audio features extraction method MFCC, and then the retrieved features are used as input for a 1D CNN model with the fewest possible layers. Because of the use of minimal convolutional layers and input size, the suggested technique boosts average accuracy while reducing computing cost, as demonstrated by the experiment.

Because of its widespread use, researchers are striving to develop an effective SER that can detect a speaker's emotions. Extraction of low-level and high-level important characteristics in SER is a difficult process. [152]. In this field, researchers have utilised a number of classification methods. Many researchers have recently applied deep learning ideas to increase SER identification accuracy. DBN is proposed, and it is discovered that it is more accurate than the baseline [153], [154]. Han et al. [130] presented the DNN extreme learning machine for SER, whereas Zeng et al. [128] employed the spectrogram with deep CNN in order to improve precision.

Many challenges arise in SER when utilising deep learning algorithms, such as determining which characteristics are more beneficial for recognising emotional state and increasing the number of layers to extract a huge number of features, which increases the computational cost. And there's no way of knowing which deep learning algorithms will result in a higher recognition rate. To overcome the constraints of SER, we proposed a 1D CNN-based technique. RAVDESS Database was used to test the suggested model.

6.2. Proposed Algorithm

The planned SER strategy is explained in this section. Using the state-of-the-art audio features extraction technique, MFCC features are extracted from audio clips. The retrieved features are sent into the 1D CNN network as input. Table 1 lists the steps for implementing the suggested technique, and Figure 1 depicts the complete architecture.

Table 6.1: Steps to implement the proposed model

Input: Audio files (RAVDESS speech corpora)
Output: Recognize the emotion as output (like calm, neutral, happy, angry, sad, fearful, surprised, disgust)

Step 1: Read .wav files from the RAVDESS speech corpora.

Step 2: Extract features from each .wav file using state-of-the-art audio feature extraction method (MFCC).

Step 3: Divide the data sets into

- Train=80% (of all data sets)
- Test =10% (of all data sets)
- Validation=10% (of all data sets)

Step 4: Apply the 1D CNN model to recognise the emotions.

Step 5: The proposed model Train and Validate

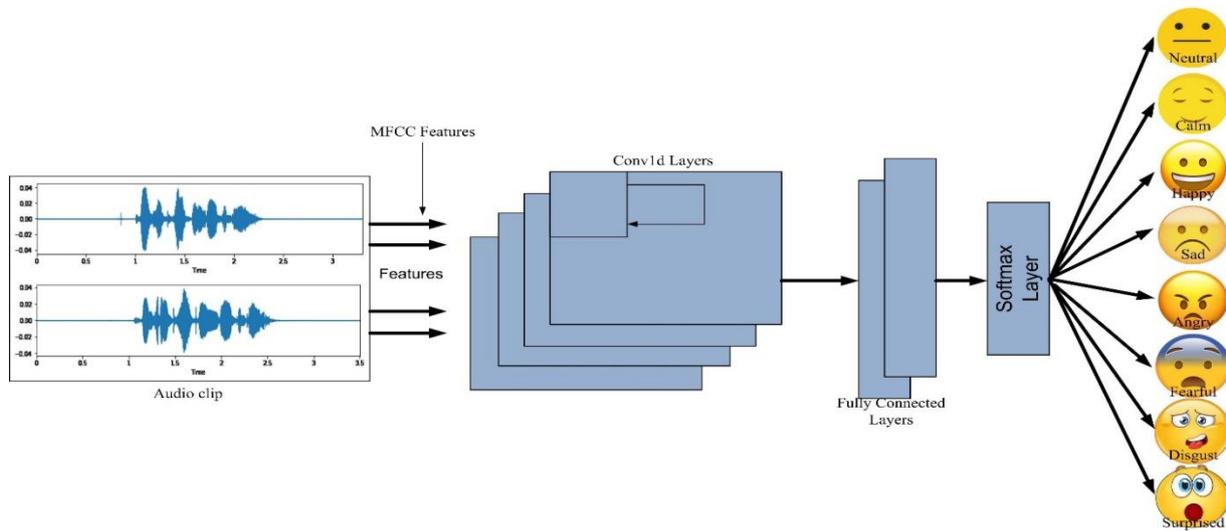


Figure 6.1: 1D CNN Based Algorithm for SER

The initial convolutional layers Conv1d-1 in our proposed model comprise 64 filters of size (5) that are applied to the collected MFCC features (40,1). Stride (1) is applied to the output feature map produced by Conv1d-1. Similarly, the second layer Conv1d-2 employs 128 size (3) filters, with the output passing via stride (2). Conv1d-3 employs 256 filters of size (3), with the output passing via stride (2). And the output of conv1d-4 travels via stride (2), which contains 512 filters of size (3). ReLU activation function is employed after each convolutional layer in this suggested model activation function to correct the output features map. The corrected output layers are followed by batch normalisation with momentum 0.9 to regularise the proposed model. The Conv1D-4 layer is then supplied to a fully - connected layer. After the

fully - connected layer, a 50% dropout rate is employed to minimise validation loss. Finally, after the fully connected layer, the soft-max function is utilised to compute the likelihood of each emotion.

6.3. Experimental Details

On RAVDESS datasets, we examine our proposed algorithm for emotion identification from speech in this chapter. Experiments are carried out using a Windows 10 laptop with a 2.70 GHz Intel(R) CoreTM i5-7200U processor and 8 GB RAM, which is an x64-based processor. The implementation specifics will be covered in the following sections.

6.3.1. Data Set

The RAVDESS Livingstone and Russo (2018) database is an actor-based speech corpus that is commonly used for emotive speech and song detection in English. For the database recording, 24 professional actors (12 men and 12 females) have participated in eight emotions. Table (2) of the audio files of speech and song statistics from the RAVDESS database.

Table 6.2: Number of samples in each emotion

RAVDESS Database	Neutral	Happy	Calm	Sad	Angry	Fearful	Surprised	Disgust	Total
Speech samples	96	192	192	192	192	192	192	192	1440
Song samples	92	184	184	184	184	184	-	-	1012

6.3.2. Experimental Results

On the benchmark database RAVDESS dataset, we evaluated our proposed algorithm. The speech samples are divided into three groups: training, validation, and testing, in the ratio of 80:10:10. For 500 epochs, we train the proposed algorithm with a learning rate of 10-4 and a batch size of 16. Figure 6.2 depicts the proposed model's validation and training loss variation.

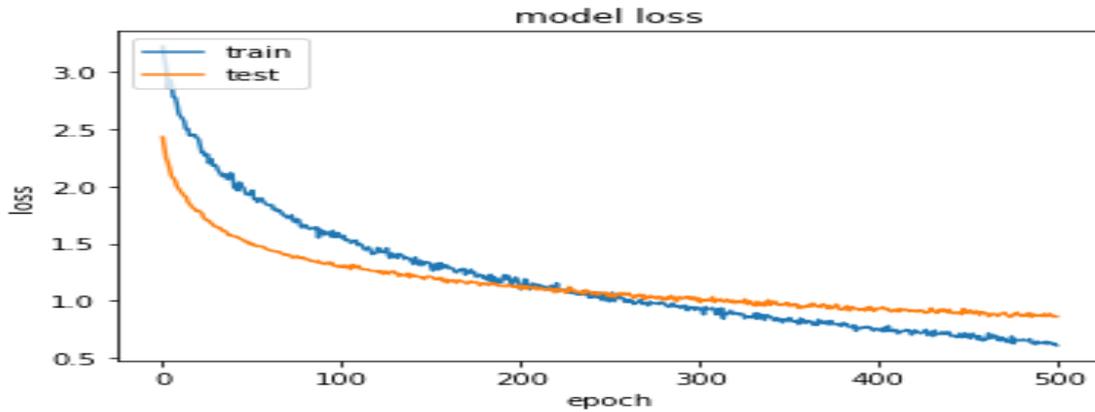


Figure 6.2: Training loss and validation loss variation of the model

Table 6.3 shows the proposed model's training, validation, and testing accuracy. The experimental outcomes of the suggested model in terms of accuracy, recall, f1-score, and support are shown in Table 6.4. Table 6.5 shows the confusion matrix.

Table 6.3: Experimental Results

Dataset	Training Accuracy	Validation Accuracy	Testing Accuracy
RAVDESS	97.01%	82.93%	82.93%

Table 6.4: Experimental Result of the proposed model on RAVDESS data sets

Emotion	Precision	Recall	F1-score	Support
Neutral	0.80	0.83	0.81	29
Calm	0.94	0.84	0.89	38
Happy	0.76	0.90	0.82	39
Sad	0.77	0.79	0.78	29
Angry	0.90	1.00	0.95	36
Fearful	0.89	0.71	0.79	48
Disgust	0.73	0.79	0.76	14
Surprised	0.69	0.69	0.69	13
Average/Total			0.83	246

Table 6.5: Confusion matrix for emotion recognition on RAVDESS

Emotional Class	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
Neutral	0.84	0.03	0.07	0.03	0.03	0.00	0.00	0.00
Calm	0.05	0.84	0.05	0.00	0.00	0.00	0.05	0.00
Happy	0.02	0.00	0.91	0.00	0.00	0.05	0.00	0.02
Sad	0.03	0.00	0.07	0.80	0.00	0.03	0.00	0.07
Angry	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00

Fearful	0.02	0.02	0.02	0.12	0.04	0.74	0.04	0.00
Disgust	0.00	0.00	0.07	0.00	0.07	0.00	0.79	0.07
Surprised	0.07	0.00	0.15	0.00	0.00	0.07	0.00	0.71
Overall Accuracy								82.93%

6.3.3. Performance Comparison of the Proposed Algorithm

The proposed algorithm's performance is compared to that of state-of-the-art techniques, and it is discovered that the proposed model provides superior accuracy (82.93 percent). The suggested model is compared to the state-of-the-art model in Table 6.6. Table 6.6 shows that the proposed model outperforms DSCNN (79.50 percent) proposed by Kwon (2020) [78], Deep BiLSTM (82.01 percent) proposed by Sajjad and Kwon (2020) [79], and 1D CNN (76.66 percent) proposed by Yulan et al. (2019) [155] in terms of accuracy (82.93 percent).

Table 6.6: Comparison of the proposed algorithm with state-of-the-art methods in term of average Accuracy

	Methods	Input	Average Accuracy (%) (RAVDESS Database)
Kwon (2020) [78]	Deep Stride CNN (DSCNN)	Clean_spectrograms	79.50%
Sajjad and Kwon (2020) [79]	Deep BiLSTM	Spectrogram	82.01%
Yulan et al. (2019) [155]	1D CNN	MelSpectrogram and MFCC	76.66%
Proposed Algorithm	1D CNN	MFCC	82.93%

6.4. Conclusion

This chapter introduces a 1D CNN-based method for voice emotion identification. In this work, the MFCC approach is employed to extract features from an audio file, and the collected features that are fed into a 1D CNN model as input. The suggested model is tested using RAVDESS dataset and is found to be more accurate (82.93 percent) than existing methods. The experimental results reveal that, when compared to other approaches for emotion identification from speech, the 1D CNN layers capture more efficient emotional information and have a lower computational cost than the 2D and 3D CNN.

CHAPTER-7

CONCLUSION AND FUTURE SCOPE

7.1. Conclusion

This chapter wraps up the study's contribution, which centred on investigating the building of an emotional speech database in Indian accents and the detection of emotions from speech. In this study, we examined and analysed the prior research in SER for the years 2000–2021 in terms of the emotional speech database, speech features, standard ML techniques, and DL approaches. Additionally, we developed the IESC emotional speech database and explored new SER techniques. To examine the unique research challenges and the most current studies, we have included a detailed review of the SER techniques in this study. It is discovered that several difficulties, such as databases, features, and classifiers, are the focus of existing research. Four publicly accessible, labelled speech corpora in distinct languages—EMODB, EMOVO, SAVEE, and RAVDESS—have been used to assess the proposed SER model. Because the suggested technique employs an ideal learning rate rather than a random learning rate, it produced superior outcomes with any type of language and actor. The study's findings are summarised in the following paragraphs:

- A study of the literature on SER's limits, speech features, and other types of emotional speech databases. We've selected some of the speech datasets that are most often utilised. analyses and reviews the many forms of emotional speech databases, including actor-based, induced/semi-natural, and natural emotional databases. A quick assessment of current speech characteristics, including local and global aspects, is provided. The Reviews' findings are then discussed.
- Analysed and evaluated the material that is currently available in terms of the traditional ML and DL methodologies used for SER from 2000 to 2021. We have provided a thorough analysis of SER techniques to examine certain research challenges and current findings. Additionally, they summarised the results of literature reviews that addressed specific research concerns and current investigations in terms of their declared study subjects and the accuracy attained by various deep learning techniques of the widely used databases. Finally, assessments of the deep learning approaches are provided regarding SER's justification and limits.

- Suggest a DCNN-based algorithm that, independent of the language or the gender of the speakers in voice corpora, can more accurately extract high-level attributes from raw data. The DCNN model is trained twice. The ideal learning rate is established in stage 1 using the Learning Rate(LR) range test, and in stage 2 the model has been retrained using the ideal learning rate. Three significant public speech corpora in different languages, EMODB (German), EMOVO (Italian), and SAVEE (British English), are used to test the proposed model. The claimed accuracy is better than prior studies for speakers of all languages. We may infer that the model would be quite good at recognising emotions.
- Created an emotional speech database (IESC) with eight speakers from north India. The generated database is manually verified by more than 20 individuals, with a stated 89.96% accuracy rate. A deep learning model based on convolutional neural networks (CNN) is suggested. On the IESC database, the recommended model outperforms the SER CNN-Assisted model by 7% and achieves an average accuracy of 95%. (average accuracy- 88 percent). On the EMOVO, EMODB, and SAVEE datasets, respectively, the recommended model beats cutting-edge SER approaches with an average accuracy of 81, 87, and 80 percent. The recommended model outperforms the other fundamental deep learning models ResNet-18 and ResNet-34 for prediction.
- A 1D CNN-based technique for identifying speech emotions has been suggested. The MFCC technique is used in this study to extract features from an audio file, and the features that are obtained are then supplied as input into a 1D CNN model. Using RAVDESS data sets, the proposed model was evaluated and shown to be more accurate (82.93%) than the standard techniques. The experimental results show that the 1D CNN layers capture emotional information more effectively and at a lower computational cost than the 2D and 3D CNN when compared to other techniques for emotion recognition from speech.

7.2. Future Scope

The goal of this research is to examine how to identify emotions using speech signals. In this study, we reviewed, analysed, traditional ML and DL approaches and proposed an efficient algorithm for SER as well as created the IESC emotional speech database. The existing approaches of SER are analysed in terms of features of speech, emotional speech database, traditional ML approaches, and DL approaches. The study of emotion identification from

speech might go in several different paths in the future. This section interprets the review's conclusions as well as the SER's existing limitations in the following ways:

- **Deciding emotion labels and filter quality speech from given speech utterances:** Emotions are regularly mingled in everyday conversation, making it difficult to determine emotion labels from spoken utterances. One research [156] looked at the effects of progressively adding additional judges while making the emotion annotation. The majority vote approach was employed in most research to determine the emotion annotation. This work did not include the annotation and identification of mixed emotions in genuine human-to-human dialogue or many other real interactions, but it may be a future avenue for SER.
- **Label subjectivity:** Emotion perception is highly individualised and dependent on factors such as listener's gender, age, and culture. Some researchers have taken into account all possible replies, but the majority have utilised strict labelling. To accommodate for label ambiguity and annotator idiosyncrasy, a joint learning technique has incorporated both hard and soft emotion label annotation, as well as individual and crowd annotator modelling [157].
- **Domain mismatch/adaptation between speech styles (acted, induced, and natural) and languages/cultures:** When we apply a model developed on data from one domain to data from another domain, performance degrades. This makes SER less helpful in practical applications. Consideration should be given to transfer learning as a potential remedy. To enhance SER performance, sparse auto-encoder-based feature transfer learning has been proposed using tiny data [158].
- **Using multiple corpora:** Since SER databases are frequently tiny, it is important to employ numerous databases. However, it is not simple since the emotion classes vary between corpora even if their domain may be the same. Multitask learning has been employed in several research to get around this problem. To increase identification rates and make the most of both labelled and unlabelled data, semi-supervised auto-encoding has been applied [159].
- The dropout and multi-task learning strategies have been found to be successful for multilingual speech emotion recognition, and common normalisation across two languages has led to further improvement under all circumstances, suggesting that even

when two highly heterogeneous languages are combined, better generalisation is still possible [160].

- **Using un-labelled speech utterances:** When the training and testing domains are not compatible, SER systems often suffer degradation. The collection of a sizable database for SER is not simple. Then, it's crucial to successfully use unlabelled utterances. This has been accomplished via semi-supervised learning [159], [161]. A significant quantity of unlabeled data has been used to successfully investigate self-supervised learning from speech [162].

LIST OF PUBLICATIONS

(A) List of International Journals

(i) Published

- Y.B. Singh, S. Goel. An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning. *Multimedia Tools and Applications* **80**, 14001–14018 (2021). <https://doi.org/10.1007/s11042-020-10399-2>. (SCIE, IF= 2.75)
- Y.B. Singh, S. Goel. (2022). A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492 (2022), 245-263. <https://doi.org/10.1016/j.neucom.2022.04.028>. (SCIE, IF=5.719)

(ii) Communicated (Major Revision Submitted)

- Y.B. Singh, S. Goel. A lightweight 2D CNN based approach for speaker-independent emotion recognition from speech with new Indian Emotional Speech Corpora. *Multimedia Tools And Applications* (SCIE, IF= 2.75)

(B) List of International Conferences

- Y.B. Singh, S. Goel. (2021). 1D CNN based approach for speech emotion recognition using MFCC features. In *Artificial Intelligence and Speech Technology* (pp. 347-354). CRC Press.
- Y.B. Singh, S. Goel. (2018, October). Survey on Human Emotion Recognition: Speech Database, Features and Classification. In *2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)* (pp. 298-301). IEEE.

REFERENCES

- [1] P. Gangamohan, S.R. Kadiri, B. Yegnanarayana, Analysis of emotional speech-a review, *Toward Robotic Socially Believable Behaving Systems-Volume I*, (2016) 205-238.
- [2] H. Altun, G. Polat, Boosting selection of speech related features to improve performance of multi-class SVMs in emotion detection, *Expert Systems with Applications*, 36(4) (2009) 8197–8203. <https://doi.org/10.1016/j.eswa.2008.10.005>.
- [3] A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu, N. Amir, The automatic recognition of emotions in speech, In *Emotion-oriented systems*, Berlin Heidelberg: Springer, (2011) 71–99. https://doi.org/10.1007/978-3-642-15184-2_6
- [4] C.N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial Intelligence Review*, 43(2) (2012) 155–177. <https://doi.org/10.1007/s10462-012-9368-5>.
- [5] M. El Ayadi, M.S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, 44(3) (2011) 572–587. <https://doi.org/10.1016/j.patcog.2010.09.020>. https://doi.org/10.1007/978-3-319-31056-5_11.w.
- [6] J.T. Senders, M. M. Zaki, A.V. Karhade, B. Chang, W.B. Gormley, M.L. Broekman, T.R. Smith, O. Arnaout, An introduction and overview of machine learning in neurosurgical care, *Acta Neurochirurgica*, 160(1) (2018) 29-38. <https://doi.org/10.1007/s00701-017-3385-8>.
- [7] D. Ververidis, C. Kotropoulos, Emotional speech recognition: resource, feature, and method, *SPC*, 48 (2006) 1162-1181. <https://doi.org/10.1016/j.specom.2006.04.003>.
- [8] C. Williams, K. Stevens, Emotional and speech: some acoustical correlates, *Journal of Acoustic Society of America*, 52(2) (1972) 1238-1250. <https://doi.org/10.1121/1.1913238>.
- [9] A. Batliner, J. Bucknow, H. Nieman, E. Noth, Volker Warnke. *Vermobile: Foundations of speech to speech translation*, ISBN 3540677836, 9783540677833: springer, 2000
- [10] S. Bansal, A. Dev, Emotional Hindi speech database, In 2013 International Conference Oriental COCOSDA held jointly with the 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE) (2013) 1-4. IEEE. <https://doi.org/10.1109/ICSDA.2013.6709867>.

- [11] B. Rambabu, K.K. Botsa, K.K. G. Paidi, S.V. Gangashetty, IIIT-H TEMD Semi-Natural Emotional Speech Database from Professional Actors and Non-Actors. In Proceedings of 12th Language Resources and Evaluation Conference (2020) 1538-1545. Marseille, France. European Language Resources Association. <https://aclanthology.org/2020.lrec-1.192>
- [12] S. Koolagudi, S. Maity, V. Kumar, S. Chakrabarti, K. Rao, IITKGP-SESC: speech database for emotion analysis, In international Conference On Contemporary Computing (2009) 485-492. Doi:10.1007/978-3-642-03547-0_46.
- [13] S. Livingstone, F. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English, PLOS ONE (2018) 13 e0196391. <https://doi.org/10.1371/journal.pone.0196391>.
- [14] S. Haq, P.J.Jackson, Multimodal emotion recognition. In Machine audition: principles, algorithms, and systems (2011) 398-423. IGI Global. <https://doi.org/10.4018/978-1-61520-919-4.ch017>
- [15] O. Martin, I. Kotsia, B. Macq, I. Pitas, The eNTERFACE' 05 Audio-Visual Emotion Database. 22nd International Conference on Data Engineering Workshops (ICDEW'06), Atlanta, GA, USA (2006) 8. <https://doi.org/10.1109/ICDEW.2006.145>.
- [16] Pichora-Fuller, M. Kathleen, Dupuis, Kate, Toronto emotional speech set (TESS), Scholars Portal Dataverse, V1 (2020). <https://doi.org/10.5683/SP2/E8H2MF>.
- [17] S. Poria, D.Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, Meld: A multimodal multi-party dataset for emotion recognition in conversations (2018). arXiv preprint arXiv:1810.02508.
- [18] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, S. Narayanan, IEMOCAP: Interactive emotional dyadic motion capture database, Journal of Language Resources and Evaluation, 42(4) (2008) 335-359. <https://doi.org/10.1007/s10579-008-9076-6>.
- [19] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset, in IEEE Transactions on Affective Computing, 1 Oct.-Dec. 2014, vol. 5, no. 4, pp. 377-390. <https://doi.org/10.1109/TAFFC.2014.2336244>.
- [20] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, T. Yang, HEU Emotion: a large-scale database for multimodal emotion recognition in the

- wild, *Neural Computing and Applications* (2021) 1-17. <https://doi.org/10.1007/s00521-020-05616-w>.
- [21] Y. Li, J. Tao, L. Chao, W. Bao, Y. Liu, CHEAVD: a Chinese natural emotional audio–visual database, *Journal of Ambient Intelligence and Humanized Computing*, 8(6) (2017) 913-924. <https://doi.org/10.1007/s12652-016-0406-z>.
- [22] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, J. Tao, Building a Chinese natural emotional audio-visual database, In 2014 12th International Conference on Signal Processing (ICSP) IEEE (2014) 583-587. <https://doi.org/10.1109/ICOSP.2014.7015071>.
- [23] M. Grimm, K. Kroschel, S. Narayanan, The Vera am Mittag German audio-visual emotional speech database. In 2008 IEEE international conference on multimedia and expo (pp. 865-868). <https://doi.org/10.1109/ICME.2008.4607572>.
- [24] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, B.W. Schuller, DEMoS: An Italian emotional speech corpus, *Language Resources and Evaluation*, 54(2) (2020) 341-383. <https://doi.org/10.1007/s10579-019-09450-y>.
- [25] G. Costantini, I. Iaderola, A. Paoloni, M. Todisco, EMOVO corpus: an Italian emotional speech database. In *International Conference on Language Resources and Evaluation, European Language Resources Association (ELRA)*, (2014) 3501-3504. http://www.lrec-conf.org/proceedings/lrec2014/pdf/591_Paper.pdf
- [26] N. Lubis, R. Gomez, S. Sakti, K. Nakamura, K. Yoshino, S. Nakamura, K. Nakadai, Construction of Japanese audio-visual emotion database and its application in emotion recognition. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'2016)* 2180-2184. Portoro, Slovenia. <https://aclanthology.org/L16-1346>
- [27] T.L.B. Khanh, S.H. Kim, G. Lee, H.J. Yang, E.T. Baek, Korean video dataset for emotion recognition in the wild, *Multimedia Tools and Applications*, 80(6) (2021) 9479-9492. <https://doi.org/10.1007/s11042-020-10106-1>.
- [28] F. Burkhardt, A. Paeschke, M. Rolfes, W.F. Sendlmeier, B. Weiss, A database of German emotional speech, In *Ninth European Conference on Speech Communication and Technology (INTERSPEECH 2005)*.
- [29] A.H. Meftah, M.A. Qamhan, Y. Seddiq, Y.A. Alotaibi, S.A. Selouani, King Saud University Emotions Corpus: Construction, Analysis, Evaluation, and Comparison, *IEEE Access*, 9 (2021) 54201-54219. <https://doi.org/10.1109/ACCESS.2021.3070751>.

- [30] A. Adigwe, N. Tits, K.E. Haddad, S. Ostadabbas, T. Dutoit, the emotional voices database: Towards controlling the emotional dimension in voice generation systems (2018). arXiv preprint arXiv:1806.09514.
- [31] E. Parada-Cabaleiro, G. Costantini, A. Batliner, A. Baird, B. Schuller, Categorical vs Dimensional Perception of Italian Emotional Speech, in Proc. of Interspeech, Hyderabad, India, (2018) pp. 3638-3642. <https://doi.org/10.5281/zenodo.1326428>.
- [32] S. Guo, L. Feng, Z.B. Feng, Y.H. Li, Y. Wang, S. L. Liu, H. Qiao, Multi-view laplacian least squares for human emotion recognition. *Neurocomputing*, 370 (2019) 78-87. <https://doi.org/10.1016/j.neucom.2019.07.049>.
- [33] S. Mo, J. Niu, Y. Su, S.K. Das, A novel feature set for video emotion recognition, *Neurocomputing*, 291 (2018) 11-20. <https://doi.org/10.1016/j.neucom.2018.02.052>.
- [34] Liu, Zhen-Tao, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, Guan-Zheng Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, *Neurocomputing* 273 (2018): 271-280. <https://doi.org/10.1016/j.neucom.2017.07.050>.
- [35] T.M. Wani, T.S. Gunawan, S.A.A. Qadri, M. Kartiwi, E. Ambikairajah, A Comprehensive Review of Speech Emotion Recognition Systems, *IEEE Access*, 9 (2021) 47795-47814. <https://doi.org/10.1109/ACCESS.2021.3068045>.
- [36] M. Fleischer, S. Pinkert, W. Mattheus, A. Mainka, D. Mürbe, Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall, *Biomech. Model. Mechanobiol*, 14(4) (2015) 719-733. <https://doi.org/10.1007/s10237-014-0632-2>.
- [37] D. Gupta, P. Bansal, K. Choudhary, The state of the art of feature extraction techniques in speech recognition, in *Speech and Language Processing for Human-Machine Communications*. Singapore: Springer, (2018). https://doi.org/10.1007/978-981-10-6626-9_22.
- [38] K. Gupta, D. Gupta, An analysis on LPC, RASTA and MFCC techniques in automatic speech recognition system, in Proc. 6th Int. Conf.- Cloud Syst. Big Data Eng. (Con_uence), Jan. (2016) 493_497. <https://doi.org/10.1109/CONFLUENCE.2016.7508170>.
- [39] M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, A. M. Kulkarni, J. A. Christensen, Voice emotion recognition by cochlear-implanted children

- and their normally hearing peers, *Hearing Res.*, 322(2015) 151-162. <https://doi.org/10.1016/j.heares.2014.10.003>.
- [40] T.L. Pao, Y.T. Chen, J.H. Yeh, W.Y. Liao, Combining Acoustic Features for Improved Emotion Recognition in Mandarin Speech. In international conference on Affective Computing and Intelligent Interaction. ACII 2005. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 3784 (2005) 279-285. https://doi.org/10.1007/11573548_36.
- [41] T.L. Pao, Y.T. Chen, J.H. Yeh, Y.M. Cheng, C.S. Chien, Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech, In International conference on Affective Computing and Intelligent Interaction. ACII 2007. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 4738 (2007) 741-742. https://doi.org/10.1007/978-3-540-74889-2_77.
- [42] N. Kamaruddin, A. Wahab, Features extraction for speech emotion. *Journal of Computational Methods in Sciences and Engineering*, 9(s1) (2009) S1-S12. <https://doi.org/10.3233/JCM-2009-0231>.
- [43] Chul Min Lee and S. S. Narayanan, Toward detecting emotions in spoken dialogs, in *IEEE Transactions on Speech and Audio Processing*, 13(2) (2005) 293-303. <https://doi.org/10.1109/TSA.2004.838534>.
- [44] M. Schroder, R. Cowie, Issues in emotion-oriented computing-towards a shared understanding. In *Workshop on Emotion and computing* (2006).
- [45] T.L. Nwe, S.W. Foo, L.C. De Silva, Speech emotion recognition using hidden Markov models, *Speech communication*, 41(4) (2003) 603-623. [https://doi.org/10.1016/S0167-6393\(03\)00099-2](https://doi.org/10.1016/S0167-6393(03)00099-2).
- [46] M. Schröder, Emotional speech synthesis: A review. In *7th European Conference on Speech Communication and Technology*, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7 (2001). <https://dblp.uni-trier.de/db/conf/interspeech/interspeech2001.html#Schroder01>
- [47] S.G. Koolagudi, K.S. Rao, Emotion recognition from speech A review. *International Journal of Speech Technology*, 15(2) (2012) 99–117. <https://doi.org/10.1007/s10772-011-9125-1>.
- [48] W. Chung-Hsien, W.B. Liang, Emotion recognition of affective speech based on multiple classifiers using acoustic prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1) (2011) 10–21. <https://doi.org/10.1109/T-AFFC.2010.16>.

- [49] K.S. Rao, R. Reddy, S. Maity, S.G. Koolagudi, Characterization of emotions using the dynamics of prosodic. In Proceedings of speech prosody, Chicago, 4 (2010). <http://www.isle.illinois.edu/speechprosody2010/program.php#100941>
- [50] M. Swain, A. Routray, P. Kabisatpathy, Databases, features and classifiers for speech emotion recognition: A review, *Int. J. Speech Technol.*, 21(1) (2018) 93-120. <https://doi.org/10.1007/s10772-018-9491-z>.
- [51] A. Guidi, C. Gentili, E. P. Scilingo, N. Vanello, Analysis of speech features and personality traits, *Biomed. Signal Process. Control*, 51(2019) 1-7. <https://doi.org/10.1016/j.bspc.2019.01.027>.
- [52] O.W. Kwon, K. Chan, J. Hao, T.W. Lee, Emotion recognition by speech signals. In 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland, September 1-4, (2003).
- [53] N. Sato, Y. Obuchi, Emotion Recognition using Mel-Frequency Cepstral Coefficients, *Journal of Natural Language Processing*. 14 (2007) 83-96. https://doi.org/10.5715/jnlp.14.4_83.
- [54] S.A. Firoz, S.A. Raji, A.P. Babu, Automatic Emotion Recognition from Speech Using Artificial Neural Networks with Gender-Dependent Databases. International Conference on Advances in Computing, Control, and Telecommunication Technologies, Trivandrum, Kerala (2009) 162-164. <https://doi.org/10.1109/ACT.2009.49>.
- [55] K.B. Khanchandani, M.A. Hussain, Emotion recognition using multilayer perceptron and generalized feed forward neural network, *CSIR* 68(2009) 367-371. <http://hdl.handle.net/123456789/3787>.
- [56] P. Shen, Z. Changjun, X. Chen X, Automatic Speech Emotion Recognition using Support Vector Machine. Proceedings of International Conference on Electronic & Mechanical Engineering and Information Technology, Harbin (2011) 621-625. <https://doi.org/10.1109/EMEIT.2011.6023178>.
- [57] P. Henríquez, J.B. Alonso, M.A. Ferrer, C.M. Travieso, J.R. Orozco-Aroyave, Nonlinear dynamics characterization of emotional speech. *Neurocomputing*, 132 (2014) 126-135. <https://doi.org/10.1016/j.neucom.2012.05.037>.
- [58] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audio visual emotion recognition. IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, (2013) 3687-3691. <https://doi.org/10.1109/ICASSP.2013.6638346>.

- [59] W. Zheng, M. Xin, X. Wang, B. Wang, A Novel Speech Emotion Recognition Method via Incomplete Sparse Least Square Regression. *IEEE Signal Processing Letters*, 21(2014) 569-572. <https://doi.org/10.1109/lsp.2014.2308954>.
- [60] K. Wang, N. An, B.N. Li, Y. Zhang, L. Li, Speech Emotion Recognition Using Fourier Parameters. In *IEEE Transactions on Affective Computing*, 6 (2015) 69-75. <https://doi.org/10.1109/TAFFC.2015.2392101>.
- [61] S. Prasomphan, Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. *International Conference on Systems, Signals and Image Processing (IWSSIP)*, London, (2015) 73-76. <https://doi.org/10.1109/IWSSIP.2015.7314180>.
- [62] S. Motamed, S. Setayeshi, A. Rabiee, Speech emotion recognition based on a modified brain emotional learning model. *Biologically Inspired Cognitive Architectures*. 19 (2017) 32-38. <https://doi.org/10.1016/j.bica.2016.12.002>.
- [63] Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* 521 (2015) 336-444. doi:10.1038/nature14539
- [64] Williams, J. Ronald, Hinton, E. Geoffrey, D.E. Rumelhart, Learning representations by back-propagating errors". *Nature*. 323 (1986) 533–536. <https://doi.org/10.1038/323533a0>
- [65] X. Yu, X. Wu, C. Luo, P. Ren, Deep learning in remote sensing scene classification: a data augmentation enhanced convolutional neural network framework. *GIScience & Remote Sens.*, 54(5) (2017) 741-758. <https://doi.org/10.1080/15481603.2017.1323377>.
- [66] D. Lopez, E. Rivas, O. Gualdrón, Primary user characterization for cognitive radio wireless networks using a neural system based on deep learning. *Artif Intell Rev*, 52 (2019) 169–195. <https://doi.org/10.1007/s10462-017-9600-4>.
- [67] M.Y. Azar, L. Hamey, Text summarization using unsupervised deep learning. *Expert Syst Appl.*, 68 (2017) 93–105. <https://doi.org/10.1016/j.eswa.2016.10.017>.
- [68] S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, X. Xu, Spatiotemporal and frequential cascaded attention networks for speech emotion recognition, *Neurocomputing*, 448 (2021) 238-248. <https://doi.org/10.1016/j.neucom.2021.02.094>.
- [69] Z. Lian, B. Liu, J. Tao, DECN: Dialogical emotion correction network for conversational emotion recognition. *Neurocomputing*, 454 (2021) 483-495. <https://doi.org/10.1016/j.neucom.2021.05.017>.

- [70] K.A. Araño, P. Gloor, C. Orsenigo, C. Vercellis, When Old Meets New: Emotion Recognition from Speech Signals. *Cognitive Computation*, 13(3) (2021) 771-783. <https://doi.org/10.1007/s12559-021-09865-2>.
- [71] S. Kwon, MLT-DNet: Speech emotion recognition using 1D dilated CNN based on a multi-learning trick approach. *Expert Systems with Applications*, 167 (2021) 114177. <https://doi.org/10.1016/j.eswa.2020.114177>
- [72] M. Chourasia, S. Haral, S. Bhatkar, S. Kulkarni, Emotion recognition from speech signal using deep learning. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI*, Singapore Springer, (2021) 471-481.
- [73] S. Kwon, Att-Net: Enhanced emotion recognition system using the lightweight self-attention module. *Applied Soft Computing*, 102 (2021) 107101. <https://doi.org/10.1016/j.asoc.2021.107101>
- [74] Y.B. Singh, S. Goel, An efficient algorithm for recognition of emotions from speaker and language independent speech using deep learning. *Multimedia Tools Appl.*, 80 (2021) 14001–14018. <https://doi.org/10.1007/s11042-020-10399-2>.
- [75] L. Pepino, P. Riera, L. Ferrer, Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings. *arXiv e-prints*. 2021. Apr:arXiv-2104.
- [76] P. Singh, R. Srivastava, K.P.S. Rana, V. Kumar, A multimodal hierarchical approach to speech emotion recognition from audio and text. *Knowledge-Based Systems*, 229 (2021) 107316. <https://doi.org/10.1016/j.knosys.2021.107316>.
- [77] D. Issa, M.F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59 (2020) 101894. <https://doi.org/10.1016/j.bspc.2020.101894>.
- [78] Mustaqeem, S Kwon, A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition, *Sensors*, 20(1) (2020) 183. <https://doi.org/10.3390/s20010183>
- [79] Mustaqeem, M. Sajjad, S. Kwon, Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM, in *IEEE Access* 8 (2020) 79861-79875. <https://doi.org/10.1109/ACCESS.2020.2990405>.
- [80] E. Guizzo, T. Weyde, J.B. Leveson, Multi-time-scale convolution for emotion recognition from speech audio signals. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020) 6489-6493. IEEE. [10.1109/ICASSP40776.2020.9053727](https://doi.org/10.1109/ICASSP40776.2020.9053727).

- [81] S.A. Kwon, CNN-assisted enhanced audio signal processing for speech emotion recognition. *Sensors*, 20(1) (2020) 183. <https://doi.org/10.3390/s20010183>.
- [82] M. Hao, C. Wei-Hua, L. Zhen-Tao, M. Wu, X. Peng, Visual-audio emotion recognition based on multi-task and ensemble learning with multiple features, *Neurocomputing* 391 (2020) 42-51. <https://doi.org/10.1016/j.neucom.2020.01.048>.
- [83] Z. Yao, Z. Wang, W. Liu, Y. Liu, J. Pan, Speech emotion recognition using a fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN, and LLD-RNN. *Speech Communication*, 120 (2020) 11-19. <https://doi.org/10.1016/j.specom.2020.03.005>.
- [84] J. Liu, Z. Liu, L. Wang, L. Guo, J. Dang, Speech emotion recognition with local-global aware deep representation learning. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020) 7174-7178. <https://doi.org/10.1109/ICASSP40776.2020.9053192>.
- [85] T. Anvarjon, S. Kwon, Deep-net: A lightweight CNN-based speech emotion recognition system using deep frequency features. *Sensors*, 20(18) (2020) 5212. <https://doi.org/10.3390/s20185212>.
- [86] J. Wang, M. Xue, R. Culhane, E., Diao, J. Ding, V. Tarokh, 2020, Speech emotion recognition with dual-sequence LSTM architecture. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2020) 6474-6478. <https://doi.org/10.1109/ICASSP40776.2020.9054629>.
- [87] M.C. Lee, S.Y. Chiang, S.C. Yeh, T.F. Wen, Study on emotion recognition and companion Chatbot using deep neural network. *Multimedia Tools and Applications*, 79 (2020) 19629–19657. <https://doi.org/10.1007/s11042-020-08841-6>.
- [88] J. Parry D. Palaz, G. Clarke, P. Lecomte, R. Mead, M. Berger, G. Hofer, Analysis of Deep Learning Architectures for Cross-corpus Speech Emotion Recognition. *Proc. Interspeech*, (2019) 1656-1660. <https://doi.org/10.21437/Interspeech.2019-2753>.
- [89] M. Ren, W. Nie, A. Liu, Y. Su, Multi-modal Correlated Network for emotion recognition in speech. *Visual Informatics*, 3(3) (2019) 150-155. <https://doi.org/10.1016/j.visinf.2019.10.003>.
- [90] J. Sebastian, P. Pierucci, September. Fusion Techniques for Utterance-Level Emotion Recognition Combining Speech and Transcripts. In *Interspeech* (2019) 51-55. <http://dx.doi.org/10.21437/Interspeech.2019-3201>.

- [91] J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control*, 47(2019) 312-323. <https://doi.org/10.1016/j.bspc.2018.08.035>.
- [92] L. Sun, B. Zou, S. Fu, J. Chen, F. Wang, Speech emotion recognition based on DNN-decision tree SVM model. *Speech Communication*, 115 (2019) 29-37. <https://doi.org/10.1016/j.specom.2019.10.004>
- [93] M.A. Jalal, E. Loweimi, R.K. Moore, T. Hain, Learning temporal clusters using capsule routing for speech emotion recognition. In *ISCA Proceedings of Interspeech (2019)* 1701-1705. <https://doi.org/10.21437/interspeech.2019-3068>.
- [94] N. Hajarolasvadi, D. Demirel, 3D CNN-based speech emotion recognition using k-means clustering and spectrograms. *Entropy*, 21(5) (2019) 479. <https://doi.org/10.3390/e21050479>.
- [95] L. Kerkeni, Y. Serrestou, K. Raouf, M. Mbarki, M.A. Mahjoub, C. Cleder, Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO. *Speech Communication*, 114 (2019) 22-35. <https://doi.org/10.1016/j.specom.2019.09.002>.
- [96] I. Shahin, A.B. Nassif, S. Hamsa, Emotion recognition using hybrid Gaussian mixture model and deep neural network. *IEEE Access*, 7 (2019) 26777-26787. DOI: 10.1109/ACCESS.2019.2901352.
- [97] B.T. Atmaja, M. Akagi, Speech emotion recognition based on speech segment using LSTM with attention model. In *2019 IEEE International Conference on Signals and Systems (ICSigSys)* (2019) 40-44. <https://doi.org/10.1109/ICSIGSYS.2019.8811080>.
- [98] M. Chen, X. He, J. Yang H. Zhang, 3-D convolutional recurrent neural networks with attention model for speech emotion recognition. *IEEE Signal Processing Letters*, 25(9) (2018) 1440-1444. <https://doi.org/10.1109/LSP.2018.2860246>.
- [99] J. Zhao, X. Mao, L. Chen, Learning deep features to recognize speech emotion using merged deep CNN. *IET Signal Processing*, 12(6) (2018) 713-721. <https://doi.org/10.1049/iet-spr.2017.0320>.
- [100] S. E. Eskimez, Z. Duan, W. Heinzelman, Unsupervised learning approach to feature analysis for automatic speech emotion recognition, in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. (2018) 5099_5103. <https://doi.org/10.1109/ICASSP.2018.8462685>.

- [101] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma, N. Dehak, Emotion identification from raw speech signals using DNNs, in Proc. Interspeech, (2018) 3097-3101. <http://eprints.whiterose.ac.uk/167268>.
- [102] J. Lee, S. Kim, S. Kim, K. Sohn, Audio-Visual Attention Networks for Emotion Recognition. In Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia (2018) 27-32. <https://doi.org/10.1145/3264869.3264873>
- [103] S. Latif, R. Rana, S. Younis, J. Qadir, J. Epps, Transfer learning for improving speech emotion classification accuracy, (2018) arXiv:1801.06353. <https://arxiv.org/abs/1801.06353>.
- [104] S. Sahu, R. Gupta, G. Sivaraman, W. AbdAlmageed, C. Espy-Wilson, Adversarial auto-encoders for speech based emotion recognition, (2018) arXiv:1806.02146. <https://arxiv.org/abs/1806.02146>
- [105] C. W. Lee, K. Y. Song, J. Jeong, W. Y. Choi, Convolutional attention networks for multimodal emotion recognition from speech and text data, (2018) arXiv:1805.06606. <https://arxiv.org/abs/1805.06606>.
- [106] S. Tripathi, H. Beigi, Multi-modal emotion recognition on IEMOCAP dataset using deep learning, (2018) arXiv:1804.05788. <https://arxiv.org/abs/1804.05788>.
- [107] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, J. Vepa, Speech emotion recognition using spectrogram & phoneme embedding, in Proc. Interspeech, (2018) 3688-3692. DOI: 10.21437/Interspeech.2018-1811
- [108] E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, S. Wermter, On the robustness of speech emotion recognition for human-robot interaction with deep neural networks, in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), Oct. (2018) 854-860.
- [109] H.M. Fayek, M. Lech, L. Cavedon, Evaluating deep learning architectures for Speech Emotion Recognition. Neural Networks, 92 (2017) 60-68. <https://doi.org/10.1016/j.neunet.2017.02.013>.
- [110] H. Tang, W. Liu, W.L. Zheng, B.L. Lu, Multimodal emotion recognition using deep neural networks. In International Conference on Neural Information Processing, (2017) 811-819. Springer, Cham. DOI: 10.1007/978-3-319-70093-9_86.
- [111] Q. Zhang, X. Chen, Q. Zhan, T. Yang, S. Xia, Respiration-based emotion recognition with deep learning. Computers in Industry, 92 (2017) 84-90. <https://doi.org/10.1016/j.compind.2017.04.005>.

- [112] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, S. Yang, Deep learning and SVM-based emotion recognition from Chinese speech for smart effective services, *Softw., Pract. Exper.*, 47(8) (2017) 1127-1138. <https://doi.org/10.1002/spe.2487>.
- [113] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. Schuller, S. Zafeiriou, End-to-end multimodal emotion recognition using deep neural networks, *IEEE J. Sel. Topics Signal Process.*, 11(8) (2017) 1301-1309. <https://doi.org/10.1109/JSTSP.2017.2764438>.
- [114] Q. Mao, G. Xu, W. Xue, J. Gou, Y. Zhan, Learning emotion-discriminative and domain-invariant features for domain adaptation in speech emotion recognition. *Speech Communication*, 93 (2017) 1–10. <https://doi.org/10.1016/j.specom.2017.06.006>.
- [115] S. Zhang, S. Zhang, T. Huang, W. Gao, Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching, in *IEEE Transactions on Multimedia*, 20(6) (2018) 1576-1590. <https://doi.org/10.1109/TMM.2017.2766843>.
- [116] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Universum autoencoder-based domain adaptation for speech emotion recognition, *IEEE Signal Process. Lett.*, vol. 24(4) (2017) 500-504. <https://doi.org/10.1109/LSP.2017.2672753>.
- [117] A. M. Badshah, J. Ahmad, N. Rahim, S. W. Baik, Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network, 2017 International Conference on Platform Technology and Service (PlatCon), (2017) 1-5. <https://doi.org/10.1109/PlatCon.2017.7883728>.
- [118] J. Han, Z. Zhang, F. Ringeval, B. Schuller, Prediction-based learning for continuous emotion recognition in speech, 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (2017) 5005-5009. <https://doi.org/10.1109/ICASSP.2017.7953109>.
- [119] Z. Wang, I. Tashev, Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks, 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), (2017) 5150-5154. <https://doi.org/10.1109/ICASSP.2017.7953138>.
- [120] S.E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, R.C. Ferrari, EmoNets: Multimodal deep learning approaches for emotion recognition in video. *J Multimodal User Interfaces* 10 (2016) 99–111. <https://doi.org/10.1007/s12193-015-0195-2>.

- [121] M. Erdal, M. Kächele, F. Schwenker, Emotion Recognition in Speech with Deep Learning Architectures. In: Schwenker F., Abbas H., El Gayar N., Trentin E. (eds) Artificial Neural Networks in Pattern Recognition. ANNPR 2016. Lecture Notes in Computer Science, Springer, Cham., (2016) 298-311. https://doi.org/10.1007/978-3-319-46182-3_25.
- [122] X. Zhou, J. Guo, R. Bie, Deep Learning-Based Affective Model for Speech Emotion Recognition, 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCCom/IoP/SmartWorld), (2016) 841-846. <https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCCom-IoP-SmartWorld.2016.0133>.
- [123] Y. Huang, M. Hu, X. Yu, T. Wang, C. Yang, Transfer learning of deep neural network for speech emotion recognition. In Chinese Conference on Pattern Recognition, Springer, Singapore 663 (2016) 721-729. https://doi.org/10.1007/978-981-10-3005-5_59.
- [124] R. Xia, Y. Liu, DBN-vector Framework for Acoustic Emotion Recognition. In INTERSPEECH (2016) 480-484. DOI:10.21437/Interspeech.2016-488
- [125] Q. Mao, W. Xue, Q. Rao, F. Zhang, Y. Zhan, Domain adaptation for speech emotion recognition by sharing priors between related source and target classes, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), Mar. (2016) 2608-2612. <https://doi.org/10.1109/ICASSP.2016.7472149>.
- [126] S. Chen, Q. Jin, Multi-modal dimensional emotion recognition using recurrent neural networks. In Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge (2015) 49-56. <https://doi.org/10.1145/2808196.2811638>.
- [127] H. M. Fayek, M. Lech, L. Cavedon, Towards real-time speech emotion recognition using deep neural networks, in Proc. IEEE 9th Int. Conf. Signal Process. Commun. Syst. (ICSPCS), Dec. (2015) 1-5. <https://doi.org/10.1109/ICSPCS.2015.7391796>.
- [128] W. Q. Zheng, J. S. Yu, Y. X. Zou, An experimental study of speech emotion recognition based on deep convolutional neural networks, in Proc. Int. Conf. Affect. Comput. Intell. Interact. (ACII), Sep. (2015) 827-831. <https://doi.org/10.1109/ACII.2015.7344669>.
- [129] P. Barros, C. Weber, S. Wermter, Emotional expression recognition with a cross-channel convolutional neural network for human-robot interaction, 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), (2015) 582-587. <https://doi.org/10.1109/HUMANOIDS.2015.7363421>.

- [130] K. Han, D. Yu, and I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, in Proceedings of INTERSPEECH, ISCA, Singapore, pp. 223–227, 2014.
- [131] Z. Huang, M. Dong, Q. Mao, Y. Zhan, November. Speech emotion recognition using CNN. In Proceedings of the 22nd ACM international conference on Multimedia (2014) 801-804. <https://doi.org/10.1145/2647868.2654984>.
- [132] Q. Mao, M. Dong, Z. Huang, Y. Zhan, Learning salient features for speech emotion recognition using convolutional neural networks, *IEEE Trans. Multimedia*, 16(8) (2014) 2203-2213. <https://doi.org/10.1109/TMM.2014.2360798>.
- [133] J. Niu, Y. Qian, K. Yu, Acoustic emotion recognition using deep neural network, in Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Pro-cess. (ISCSLP), Sep. (2014) 128-132. <https://doi.org/10.1109/ISCSLP.2014.6936657>.
- [134] Y. Kim, H. Lee, E. M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, (2013) 3687-3691. <https://doi.org/10.1109/ICASSP.2013.6638346>.
- [135] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, G. Rigoll, LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2) (2013) 153-163. <https://doi.org/10.1016/j.imavis.2012.03.001>.
- [136] S. He, S. Wang, W. Lan, H. Fu, Q. Ji, Facial Expression Recognition Using Deep Boltzmann Machine from Thermal Infrared Images, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, (2013) 239-244. <https://doi.org/10.1109/ACII.2013.46>.
- [137] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, B. Schuller, Deep neural networks for acoustic emotion recognition: Raising the benchmarks, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), (2011) 5688-5691. <https://doi.org/10.1109/ICASSP.2011.5947651>.
- [138] Yu D, Deng D (2016) Automatic Speech Recognition. Springer London limited
- [139] Ververidis D, Kotropoulos C (2006) Emotional speech recognition: Resources, features, and methods. *Speech Communication*. 48:1162-1181. <https://doi.org/10.1016/j.specom.2006.04.003>.
- [140] Polzehl T, Sundaram S, Ketabdard H, Wagner M, Metze F (2009) Emotion classification in children's speech using fusion of acoustic and linguistic features. in: Intenth Annual Conference of The International Speech Communication Association.

- [141] Özseven T (2019) A novel feature selection method for speech emotion recognition, *Applied Acoustics*. 146:320-326. <https://doi.org/10.1016/j.apacoust.2018.11.028>.
- [142] Badshah A, Rahim N, Ullah N, Ahmad J, Muhammad K, Lee M et al.(2019) Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*. 78:5571-5589. <https://doi.org/10.1007/s11042-017-5292-7>.
- [143] Daneshfar, F., & Kabudian, S. J. (2020). Speech emotion recognition using discriminative dimension reduction by employing a modified quantum-behaved particle swarm optimization algorithm. *Multimedia Tools and Applications*, 79(1), 1261-1289.
- [144] Zvarevashe K, Olugbara O (2020) Ensemble learning of hybrid acoustic features for speech emotion recognition. *Algorithms* 3(3):70. <https://doi.org/10.3390/a13030070>
- [145] Nakatsu R, Nicholson J, Tosa N (2000) Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge-BASED Systems* 13:497-504. [https://doi.org/10.1016/s0950-7051\(00\)00070-8](https://doi.org/10.1016/s0950-7051(00)00070-8)
- [146] TENBOSCH L (2003) Emotions, speech and the ASR framework. *Speech Communication* 40:213-225. [https://doi.org/10.1016/s0167-6393\(02\)00083-3](https://doi.org/10.1016/s0167-6393(02)00083-3)
- [147] Polzin T, Waibel A (1998) Detecting emotions in speech, In *Proceedings of the CMC*.
- [148] He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. *IEEE conference on Computer Vision and Pattern Recognition* 770-778.
- [149] Meng H, Yan T, Wei H, Ji X (2021) Speech emotion recognition using wavelet packet reconstruction with attention-based deep recurrent neural networks, *Bull. Pol. Ac.: Tech* 69(1):136300. <https://doi.org/10.24425/bpasts.2020.136300>
- [150] Zayene B, Jlassi C, Arous N (2020) 3D Convolutional Recurrent Global Neural Network for Speech Emotion Recognition. *5th IEEE International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)* 1-5. <https://doi.org/10.1109/ATSIP49331.2020.9231597>.
- [151] Ancilin J, Milton A (2021) Improved speech emotion recognition with Mel frequency magnitude coefficient, *Applied Acoustics* 179:108046. <https://doi.org/10.1016/j.apacoust.2021.108046>
- [152] Wei, B., Hu, W., Yang, M., and Chou, C.T., 2019. From real to complex: Enhancing radio-based activity recognition using complex-valued CSI. *ACM Transactions on Sensor Networks (TOSN)*, 15(3), pp.1-32.

- [153] Kim, Y., Lee, H. and Provost, E.M., 2013, May. Deep learning for robust feature generation in audiovisual emotion recognition. In 2013 IEEE international conference on acoustics, speech and signal processing (pp. 3687-3691). IEEE.
- [154] Zheng, W.L., Zhu, J.Y., Peng, Y. and Lu, B.L., 2014, July. EEG-based emotion classification using deep belief networks. In 2014 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.
- [155] Li, Y., Baidoo, C., Cai, T. and Kusi, G.A., 2019. Speech Emotion Recognition Using 1D CNN with No Attention. In 2019 23rd International Computer Science and Engineering Conference (ICSEC) (pp. 351-356). IEEE.
- [156] A. Burmania, C. Busso, A Stepwise Analysis of Aggregated Crowdsourced Labels Describing Multimodal Emotional Behaviors. In INTERSPEECH (2017, August) 152-156. <https://doi.org/10.21437/Interspeech.2017-1278>.
- [157] H. Chou, C. Lee, Every Rating Matters: Joint Learning of Subjective Labels and Individual Annotators for Speech Emotion Classification, ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5886-5890. <https://doi.org/10.1109/ICASSP.2019.8682170>.
- [158] J. Deng, Z. Zhang, E. Marchi B. Schuller, Sparse Autoencoder-Based Feature Transfer Learning for Speech Emotion Recognition, 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 511-516. <https://doi.org/10.1109/ACII.2013.90>.
- [159] J. Deng, X. Xu, Z. Zhang, S. Frühholz, B. Schuller, Semisupervised Autoencoders for Speech Emotion Recognition, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, Jan. 2018, vol. 26, no. 1, pp. 31-43. <https://doi.org/10.1109/TASLP.2017.2759338>.
- [160] S. Lee, "The Generalization Effect for Multilingual Speech Emotion Recognition across Heterogeneous Languages," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 5881-5885. <https://doi.org/10.1109/ICASSP.2019.8683046>.
- [161] S. Parthasarathy, C. Busso, Semi-Supervised Speech Emotion Recognition With Ladder Networks, in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, vol. 28, pp. 2697-2709. <https://doi.org/10.1109/TASLP.2020.3023632>.
- [162] S. Evain, H. Nguyen, H. Le, M.Z. Boito, S. Mdhaffar, S. Alisamir, Z. Tong, N. Tomashenko, M. Dinarelli, T. Parcollet, A. Allauzen, Y. Estève, B. Lecouteux, F. Portet,

S. Rossato, F. Ringeval, D. Schwab, L. Besacier, LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. Proc. Interspeech 2021, 1439-1443. <https://doi.org/10.21437/Interspeech.2021-556>.