Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions

by

Nuhash Kabir Neeha
21301025
Nuzhat Rahman
21301538
Imtela Islam
21341018

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
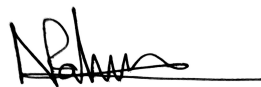Brac University
Month Year.

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

_____
Nuhash Kabir Neeha
21301538

_____
Nuzhat Rahman
21301538

_____
Imtela Islam
21301538

Approval

The thesis/project titled "Speech Emotion Recognition: Clustering and Deep Learning Approach to Detect Conflicting Emotions through Vocal Expressions" submitted by

1. Nuhash Kabir Neeha (21301025)

2. Nuzhat Rahman (21301538)

3. Imtela Islam (21341018)

of Summer 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in Fall 2024.

Examining Committee:

Supervisor:
(Member)

Dibyo Fabian Dofadar

Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Given the increasing dependency to automated systems, it is essential that machines understand human emotions more effectively in order to improve user experience as well as prevent extreme damages in multiple areas, starting from customer service and virtual assistants to emergency hotlines and criminal interrogation. As such, Speech Emotion Recognition (SER) is a crucial part of human-computer interaction(HCI) as it aims to improve a machine's ability to understand human emotions through their vocal expressions. However, despite advancements, current SER models face challenges to accurately recognize emotions due to one having multiple emotions(conflicting emotions) at a time. This study aims to cluster and then train deep learning models to accurately recognize such emotional nuances. In conclusion, we aim to make machines have more natural and effective interactions with humans by making technology more responsive to emotional cues.

Keywords: Sound Emotion Recognition, Human-Computer Interface, Conflicting Emotions, Deep Learning Models, Clustering, Emotional Cues, Natural Interactions, Machine Responsiveness, Vocal Expressions.

Table of Contents

List of Figures

Chapter 1

Introduction

In this era of technological advancements, human-computer interactions have become a part of daily life. Along with improving user experience and others, HCI also aids in creating a natural and efficient interface for human-computer communication . Additionally, play a role in criminal interrogations particularly for analyzing behaviours, detection processes, improving communications and so on. Emotions are a core part of human interaction. A critical aspect of these interactions depends on machines' ability to understand human emotions and respond to it in real time. Sound Emotion Recognition(SER) can be used to bridge the emotional gap between humans and machines. SER enables systems to perceive and interpret emotions based on vocal expressions. While the existing SER models have been progressing significantly over time; there are areas in this field that are yet to be explored. Many models may struggle with the dynamic nature of human emotions, as one may be experiencing multiple emotions at the same time (i.e. fear and anxiety, fear and excitement, deep sadness and anxiousness and so on.). These real-life overlapping emotions make it challenging for the existing SER models to recognize and accurately classify emotions, which in turn leads to reduced model performance. This study aims to face these challenges by employing advanced clustering and deep learning techniques and models to accurately catch the nuanced emotions one might be experiencing through speech. This study will contribute in creating a more natural and emotionally intelligent interaction between humans and automated systems, by enabling machines to interpret such intricate emotions. Consequently, making these systems more accurate, sensitive and precise.

1.1 Background

Emotions are complex psychological and physiological states triggered by different stimuli. This section is concerned with defining what emotions are, what models are used to understand emotions and human expression of emotions.

1.1.1 Emotions, Moods, Feelings

Speech is one of the primary ways of conveying information, while emotions are fundamental human expressions of different experiences. Although closely related, the concepts of emotion, feeling, mood are distinct.
According to a study[14] emotions are intense, short-lived reactions triggered by

internal or external stimuli(physiological and psychological). It does not occur consciously. In contrast, feelings are subjective experiences of an emotion. It is the interpretation of emotional responses. It occurs consciously. Lastly, moods are sustained emotional states with no specific triggers.

1.1.2 Emotional Models

Emotions can be categorised into 2 popular approaches. Firstly, the categorical (discrete) approach describes emotion as a discrete number of classes. One of the most popular theorists [2], listed 6 basic emotions: anger, happiness, surprise, sadness, fear and disgust. These are universal regardless of one's cultural and social background. Secondly, in Russel's 2D Model, the dimensional (continuous) approach describes emotions on a continuous scale of valence, which can be positive or negative, and arousal, which can be high or low. PAD (Russel's 3D Model, stands for Pleasure, Arousal, Dominance) is another widely used emotional state model, consisting of 3 dimensions: arousal, pleasure, dominance [1]. Discrete model distinctly categorises emotions giving them their own set of cognitive and psychological elements, whereas the dimensional model recognises the complexity of emotional experiences, influenced by many factors such as personal history, cultural background and so on. It uses 2D (arousal and valence) and 3D (arousal, valence and dominance) emotional space models.

1.1.3 Expression of Emotions

Humans use different channels to express emotions. According to another study[6], facial expressions can communicate multiple emotions without the need for verbalizing. Some of the universal nonverbal emotions expressed through facial features are anger, surprise, fear etc. In addition to facial expressions, vocal cues—including pitch, amplitude, and frequency of sound waves—play a crucial role in identifying an individual's emotional state, forming the foundation of SER. SER uses speech to recognize emotions; however, body language, consisting of gestures, postures, eyemovent, hand placement or movement further enhances the communication of emotions. Likewise, physiological signals such as skin conductance, electrocardiogram(ECG), blood volume pulse (BVP), heart rate and so on, can be used as modes of expression of emotions for individuals who suffer from mental or physical illnesses; where they are unable to communicate efficiently through facial expression, vocal cues or body language[17]. Figure 1.1 shows a breakdown of this.
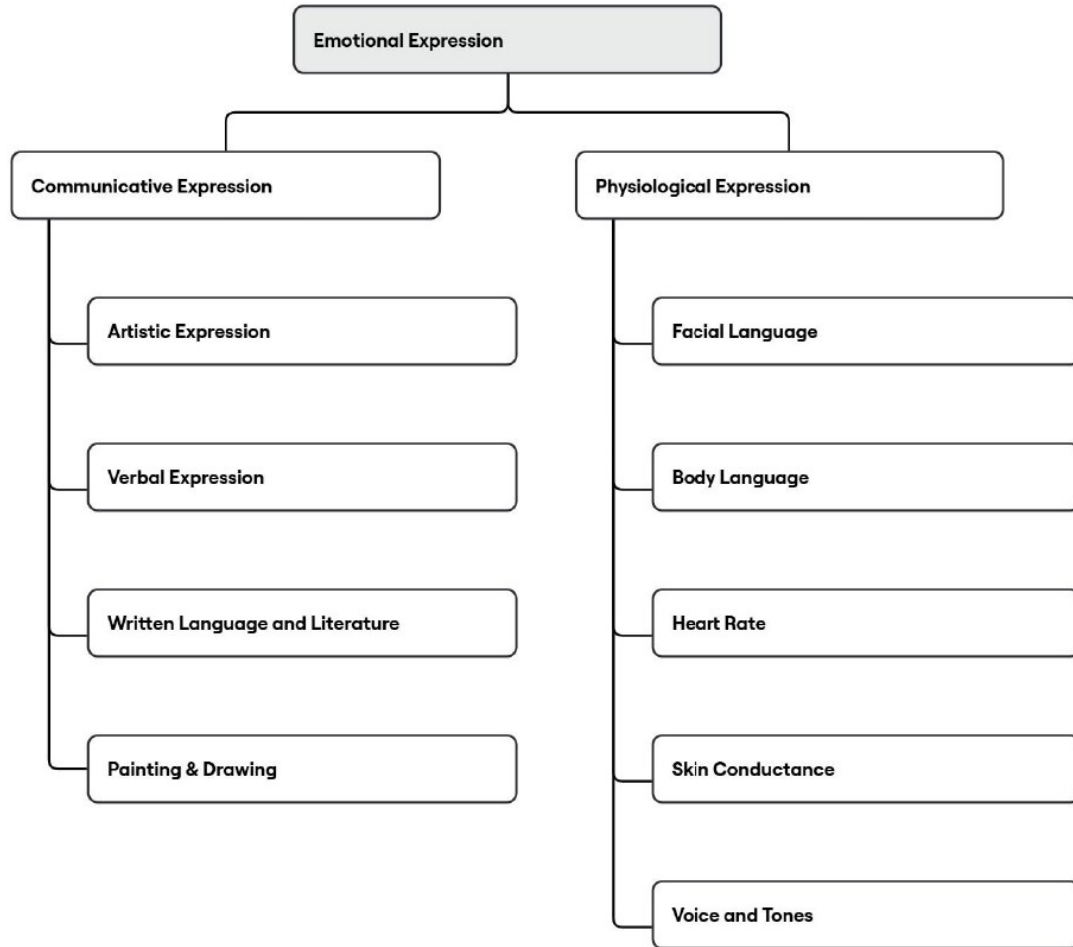
Figure 1.1: Human Emotional Expression

1.1.4 Emotion Recognition

There are two types of emotion recognition: non-automated recognition and automated recognition. Non-automated recognition refers to the natural human ability to perceive and interpret emotions. Researchers have developed structured techniques to assess emotions such as FACS (Facial Action Coding System), Geneva Emotion Wheel, SAM (Self-Assessment Manikin) along with psychological assessment questionnaires like PANAS (Positive and Negative Affect Schedule), BDI (Beck Depression Inventory), STAI (State-Trait Anxiety Inventory) and POMS (Profile of Mood States). In contrast, Automated recognition recognises emotions and interprets them with the help of machines. These systems can further be divided into 3 modes: unimodal, bimodal and multimodal. Unimodal takes into account 1 mode of expression, whereas Bimodal uses 2 modes of expression and Multimodal uses multiple modes of expression.[17]

1.2   Rational of the Study or Motivation

As technology advances and gets integrated into daily life, it becomes crucial for machines and humans to have natural and smooth communication, to improve user satisfaction, and ensure more intuitive and interactive computer interactions. Even though SER systems are high in demand and are useful in many aspects they still lack the ability to differentiate between the conflicting or overlapping emotions. Existing SER systems may face challenges classifying between frustration and determination or anxiety and excitement. A more adaptive and nuanced recognition technique is required to face said challenges.

1.3   Problem Statement

SER models can be indispensable in human-computer interaction, enabling human-machine communication to be emotional. However, traditional SER models struggle to identify conflicting and overlapping emotions, as most of the systems classify emotions into discrete categories such as happy, sad, angry. The models assume emotions to be static and independent. Although, in reality emotions are interdependent, dynamic, complex and multiple emotions can often be expressed simultaneously. For example, individuals may experience feeling both anxious and excited at the same time. Existing SER models, especially the models dependent on manually extracted features such as MFCCs, spectral features and so on fail to capture such emotional nuances, leading to misclassification and reduced performance. This research aims to use advanced clustering techniques in addition to deep learning models to identify said emotional nuances and create a model for better emotion recognition of conflicting emotions.

1.4   Objective

This study aims to solve the challenges of overlapping emotions faced by SER systems and optimizing the accuracy of said systems. The goal is to create a more robust and effective emotion recognition technique, in order to classify multiple emotions being expressed simultaneously. Key objective is analyzing limitations faced by SER models, developing multi-emotion classification approach, and evaluating its' effectiveness. Ultimately, this study aims to be able to engineer a more emotionally intelligent system based on SER for enhanced human-computer interaction.

1.5   Methodology in Brief

This research paper was based on reading many technical research articles and journals related to sound/speech emotion recognition. Firstly, the research focused on understanding the complexity of emotions. Also, how emotional nuances can be expressed using different channels. Moreover, different modes of emotional recognition. Secondly, review papers were analysed to look for research gaps and key challenges

faced by currently existing SER systems. Lastly, different models were compared and contrasted to see which models performed better and what were their consequent drawbacks. The papers, articles and so on that were read for this study were chosen based on relevance, large number of citations and more recent publication. These strategically selected papers will help understand the recent developments and applications of SER. This will lay the foundation for future research and model development.

1.6 Work Plan

The work plan for this thesis paper has been divided into 3 distinct parts: pre-thesis 1, pre-thesis 2 and defense. In pre-thesis 1, a research gap is discovered. The problem is further researched and defined undertaking a comprehensive study of the problem. This paper comprises literature review which helps build a solid theoretical foundation in order to understand the problem and research objectives. Furthermore, pre-thesis 2 encompasses data collection, feature extraction and development of basic SER model in order to solve the challenge of detecting conflicting or overlapping emotions. Lastly, defense will involve the advancement and further development of the hybridized model made using clustering and deep learning models. The model will be evaluated depending on the suitable evaluation metric and compared with state of the art models.
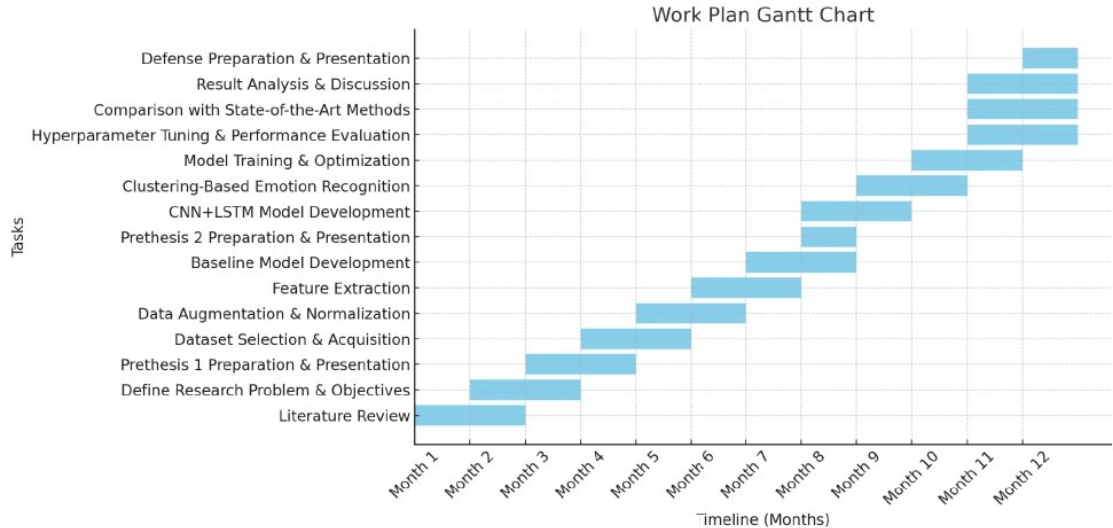

Figure 1.2: GANTT chart of work plan

Chapter 2

Literature Review

2.1 Sound/ Speech Emotion Recognition

A review on SER[15] states that it is a field of study that deduces human emotions from verbal expression. The system focuses on identifying and classifying voice input to various defined categories of emotion. It identifies the acoustic and linguistic features of a speech signal to capture these emotional states as such plays a crucial role in human-computer interaction(HCI) as well as behavioral analysis.

Definition of SER
It[16] is stated that speech emotion recognition is a part of speech processing which analyses speech signals and recognises patterns of speech such as prosody, pitch, frequency, rhythm in order to determine the emotional state of the speaker. Classifying emotions using sound/speech of an individual has various applications, majorly in human-computer interaction.

Relevance of Using SER
In today's day and age, increased dependency on technology gives rise to the need for better technical systems. People in recent years have become more reliant on technology, mostly human-computer interaction (HCI), artificial intelligence (AI) and so on, necessitating more emotionally-aware computer systems in order to enhance user satisfaction. SER could be applied in multiple fields, namely, emergency hotline, virtual assistants, criminal investigation etc..

2.2 Evolution of SER

2.2.1 Early Research and Traditional Methods

The research on sound emotion recognition had begun by using acoustic features and statistical classification models. Initially, models like support vector machine (SVM), hidden markov model (HMM) were used along with prosodic and spectral features (pitch, energy, mel-frequency cepstral coefficients). Although these methods did provide insights into emotion classification, they struggled to accurately identify the emotions in the presence of noise, speakers of different languages as well as in the cases of conflicting emotions.

2.2.2 Transition to Machine Learning

Due to the struggles faced by the traditional methods, Machine Learning(ML) approaches were introduced to SER. ML improved SER by allowing automated feature selection and classification. Models such as Random Forests and Gaussian Mixture Models (GMMs) demonstrated higher adaptability and accuracy. Even after these improvements, automatic feature selection is not completely reliable as it is highly dependent on feature engineering quality as well as dataset variability. As such, manually extracted features remain as a limitation.

2.2.3 Adoption of Deep Learning

The adoption of deep learning (DL) improved on the limitations previously faced by the earlier methods by eliminating the need for handcrafted features. Models such as Convolutional Neural Networks (CNNs) effectively extracted spatial patterns from spectrograms, while Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, modeled temporal dependencies in speech. Hybrid CNN-LSTM architectures improved the recognition accuracy as such improving generalization to real-world emotional speech. Compared to ML-based methods, DL models demonstrated a higher level of robustness and scalability, even though they required extensive data and computational resources.

2.2.4 Role of Benchmark Datasets

The availability of standardized datasets has been crucial in advancing SER. Corpora such as RAVDESS, IEMOCAP, Emo-DB, CREMA-D, SAVEE and TESS provide diverse linguistic and emotional variations, facilitating the development of more adaptable models. These datasets have enabled direct comparisons between ML and DL approaches, demonstrating the effectiveness of deep learning in handling complex emotional variations. SER has evolved from statistical methods to deep learning-driven approaches, significantly improving emotion recognition accuracy. However, challenges related to cross-corpus generalization, speaker variability, and computational efficiency remain key areas for further research.

2.3 Taxonomy of Methods in SER
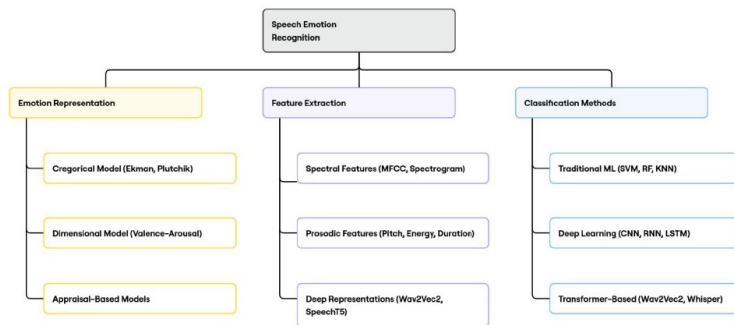
The taxonomy in SER can be shown as follows:

Figure 2.1: Flow chart for SER taxonomy

2.4   Key Challenges in SER

The key challenges faced by SER systems, according to some studies [3]–[7], [13], were as follows:

1. **Dataset Limitations:** almost all the articles that were analysed in order to write this paper used the same datasets available. This can be addressed by data augmentation. A small dataset can lead to overfitting and limited generalizability.

2. **Noise and Distortions in Speech Data:** background noise and distortions reduce the quality of the data and introduce added complexities.

3. **Audio Feature Extractions:** extracting relevant audio features is crucial for effective SER

4. **Class Similarity and Misclassification:** overlapping emotions can cause 2 different emotional vocal cues to sound similar leading to an increased risk of misclassification.

5. **Scalability and Real-time Processing:** implementing operational SER systems that can process audio information remains highly challenging.

6. **Linguistic Variations:** Impact of cultural, personal, and contextual factors affect the differences between emotional tone and expression.

7. **Generalization:** Models performing poorly on unseen datasets or real-world scenarios

8. **Model Selection and performance:** different ML, DL models have different strengths and weaknesses. Model interpretability can also be a limiting factor.

2.5   Application of SER

According to multiple studies[4], [6], there are multiple applications for SER.
Practical applications:

1. **HCI:** SER enhances human computer interaction by allowing machines to identify and interpret human emotions for more natural communication.

2. **Healthcare:** SER can help with mental health assessments by using it during therapy session to monitor their emotional response

3. **Entertainment:** if SER is used in gaming or interactive media, understanding player emotions can lead to a higher user satisfaction.

Academic applications:

1. **Research & Development:** studying and using SER contributes in the field of interactive computing and emotional AI. Consequently, advancing machine learning, signal processing and cognitive psychology.

2. **Emotion Theory:** exploration of SER helps improve the existing theory on emotions, which in turn advances studies on psychological and linguistic studies depending on emotional states through vocal expression.

Societal impact:

1. **Communication improvement:** SER can improve social interactions by recognizing emotions.

2. **Educational enhancement:** SER can further help teachers understand students' emotional state and create a work plan that will best enhance their performance.

3. **Inclusion and accessibility:** SER can be used to aid individuals with disabilities by providing them with a smoother and better communication through emotionally aware systems.

SER has become very popular due to its multiple applications. However, many SER systems face challenges to accurately identify and interpret emotions due to conflicting and overlapping emotions at a time, limiting their effectiveness. This complexity makes it difficult for traditional categorical emotion classification making it insufficient and thus requiring more adaptive and nuanced recognition techniques.

2.6   Existing Model Architecture

One of the studies[13] uses a one-dimensional convolution network (Conv1D). It is a deep learning technique that processes audio features to learn the complex patterns using convolution layers. It is a good model for capturing local, temporal dependencies in audio signals. Additionally, they also use random forest (RF). It is a tree-based machine learning algorithm that works by constructing multiple decision trees. Using it alongside feature pruning increases performance, provides dimensionality reduction, and reduces computational cost. The comparative analysis of Conv1D and RF provides insight on the quality of performance of the two models. Consequently, becoming the highlight of the study.

In comparison, another study[5] uses an ASR model at its core for audio-to-text mappings, utilizing dilated convolution and gated convolutional units. Also, in this study the authors use the ASR model as a feature extractor for emotion recognition tasks. This is done by computing mean activations for different layers. Furthermore, a linear regression model is used to predict the arousal and valence values from the extracted features. This approach made their study novel by utilizing ASR as feature extractor, exploration of layer contributions, and end-to-end training considerations.

Conversely, another[6] uses deep learning models like RNN (Recurrent Neural Network) and LSTM (Long Short-Time Memory). RNN works by keeping a record of old information which makes it suitable for capturing temporal information or dependencies of sequential data such as speech signals. LSTMs are extended versions of RNN. LSTM allows for long-term dependency learning. In this study, the authors also used SVM (Support Vector Machine) for classification. Even thougH, SVM works with highly dimensional data it is a more traditional model used for SER.

In addition, CNN (Convolution Neural Network) can also be a good model to use when using automated feature extraction, i.e. spectrograms or mel-spectrograms. Spectrograms are computed using STFT(Short-Time Fourier Transform) and mel-spectrograms are computed by applying mel filter banks on spectrograms. Lastly, it was discovered that using a hybrid model enhances model performance. This hybrid approach introduces novelty in their study. Hybris models can be created by integrating attention mechanisms in RNNs and LSTMs, consequently, aiding the model focus on emotionally rich parts of speech signal. It was also discovered that multimodal models classified emotions more accurately than unimodal or bimodal models as they considered multiple modes of expression

Similarly, a different source[7] also reviews various deep learning techniques, including Deep Belief Networks (DBNs), and Deep Boltzmann Machines (DBMs) along with more commonly used Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks. CNNs extracts spatial features directly from input data through convolutional layers. These deep learning techniques leverage automated feature extraction, which in turn helps RNNs and LSTMs identify temporal dependencies in speech signals. Consequently, enhances long-term dependency recognition. They introduce novelty through hybridization of multiple models. The implementation of their model was executed using Python, leveraging libraries such as TensorFlow, Keras, and Librosa, among others.

In contrast, another research[3] used Deep Neural Network(DNN) and Stochastic Gradient Descent (SGD). DNN uses a combination of convolution, pooling and fully connected layers; whereas, SGD is an optimization approach. This leads to a more efficient training and accurate classification, updating the model parameter in order to reduce the loss function. The use of automated feature extraction is the highlight of this study. To accomplish this they enhanced temporal resolution by using segmented 20 ms frames and implemented the model using python leveraging libraries such as Keras and Theano.

Subsequently, another study[4] also used a DNN. In contrast, the author hybridizes DNN (Deep Neural Networks) with KNN (K-Nearest Neighbor) rather than SGD (Stochastic Gradient Descent). As mentioned before, DNN uses convolution and multiple fully connected layers in order to solve complex problems. Whereas, KNN is a machine learning algorithm that classifies data depending on the distance between the data points of the number of neighbours chosen. KNN was used to improve accuracy, also, as a secondary validation method. Hybridization of ML and DL, where incorporating statistical features enhances the performance of the model, and makes it stand out amongst all the others.

On the other hand, a different study[8] used CNN and LSTM. They used 3 convolution layers for the CNN model for spectrograms and mel-spectrograms, with different feature maps and dropout layers each and softmax function was used for output. Furthermore, they used a variant of LSTM, Bi-LSTM, where the layers are bidirectional and fully connected, including the softmax layers. This BLSTM model uses feature vectors for input and trains the models using sequence-to-one

modelling. They have also used an ensemble of the two models where they used CNN followed by BLSTM, in order to automate feature extraction using CNN and feeding it directly to BLSTM. This is done to reduce the inefficiency of modelling the complexities of emotional nuances in a single utterance. Here, the CNN is used to extract high-level features, whereas the BLSTM is used to model the utterance-level long term dependencies subsequently. Finally, to compare and classify the models, a cross-entropy loss function was used along with the Adam optimiser to ensure the model is not too computationally expensive.

In a subsequent paper[10], supervised models - Regression and Artificial Neural Networks (ANN) - were used on the IADS dataset. The similarity of features in this dataset are compared using Pearson's r correlation coefficient, which were then used as inputs in the regression and ANN models. The emotional dimension of valence and arousal were predicted using the output from the following models. For regression analysis, RMSE was used, where the lower the RMSE value, the better the performance of the model. To ensure unbiasedness, principal component analysis (PCA) was used on data with and without dimension reduction in order to use five and ten fold cross validation. In contrast, they also used a two-layer feed forward ANN , consisting of one hidden layer. Additionally, the arousal and valence for this model were determined using two separate ANNs. The training-validation-testing ratio for the data set used was 70-15-15.

Furthermore in a differnt study[9], the authors study different models of SER including traditional as well as the ones created in recent times. Different features extractors, for example, openSmile, Voice Quality Features, MFCCs, Pulse Code Modulation (PCM), Prosodic Features and more traditional sStatistical Methods were also explored. And it was stated that the type of extractor used depends on the specific context it is to be used in. The feature extracted using these methods are used as inputs into models such as Hidden Markov Models (HMM) - which is a more traditional model , SVMs, ANNs, CNNs, RNNs - especially LSTM networks, and Generative Adversarial Networks (GANs) and Variational Autoencoders (VAE) - which are more advanced architectures used in the more recent years . HMM is a statistical model which consists of a finite set of hidden states, transitions (probabilities of moving from one state to another), emissions (probabilities of an observable output), and initial probabilities (the starting probabilities in each of the hidden states). On the contrary, GANs and VAEs are both generative deep learning models. GAN works by training two neural networks (generator and discriminator) at the same time in a competitive manner, and VAEs consist of an encoder and a decoder -which are trained to minimise the difference between the input and the decoded output data.

2.7 Model Comparison

One of the studies[6] employed 2 models RNN and SVM and compared it against MLR. The features extracted were MFCC and MS (Modulation Spectral). Feature selection (FS) was used to find the most relevant feature subset. It was observed that the RNN model when run on the Spanish database obtained an accuracy of 94%, performing better than SVM OR MLR. It was also observed that employing speaker

normalisation improved model performance when the model was run on the Berlin database but had insignificant effect when run on the Spanish database. Moreover, feature selection enhanced performance by decreasing dimensionality. On the other hand, another[5] explored models like ASR model with transfer learning and linear Regression for emotion estimation. The ASR system was used as a features extractor which in turn feeds the features into the linear regression model that estimates the valence-arousal values. Since, ASR- feature extractor was used instead of using manually extracted features which aided the model to outperform previously engineered models. This study was conducted on an IEMOCAP dataset, and the study also found that certain parts of the ASR system correspond differently to certain emotional expressions. From that fact, it can be inferred that an ASR based model is an unsupervised learner, where it naturally picks up on emotional trends in speech. In addition, a different paper[13] explored ML and DL models like Conv1D and RF. the extracted features used were MFCCs, chromograms, mel-spectrogram, tonnetz, ZCR and more. After feature selection was used RF achieved better accuracy than Conv1D (69%). Moreover, it had a precision of 72% for fear and recall of 84% for calm. RF performed well whereas Conv1D misclassified emotions like anger, disgust and fear, with happy, neutral, sad respectively due to overlapping emotional features.

A different research[3] observed an accuracy of 96.97% after employing convolutional and fully connected layered DNN. Even though the model performed well it still faced challenges. The lack of feature selection and the model needing substantial hyperparameter tuning makes it dependent on larger datasets. Similarly, another study[4], also used a DNN model, although it was employed along with KNN as a hybrid model, in order to detect distress signals which consequently improves classification accuracy. Nevertheless, the hybridization introduced complexity which in turn raised computational cost. Alternatively, a different research[7], used deep learning techniques to automate feature extraction. In conclusion, it was observed that there was always a trade off between accuracy, precision computational cost and modular complexity. This highlights the challenge of balancing these factors in order to engineer an efficient system for optimal SER.

In a paper from a different set of studies, authors[8] found that a hybridized model of CNN and BLSTM outperformed the models on only CNN or BLSTM. An accuracy of 82.35% was obtained when MFCC features were applied on EmoDB dataset; on the other hand, 50.05% accuracy was obtained using mel-spectrogram for IEMO-CAP dataset. For the former dataset, the confusion matrix showed that happy speech was not detected as accurately, which could be either because there were not enough instances of happy speech in the dataset or due to happy and angry being part of the same category (arousal), the subtle differences were not captured accurately. In contrast, the confusion matrix for the latter dataset showed that the model successfully recognised all the four basic emotions, despite the more natural instances in this dataset. They speculated that the staggering difference (82% vs 50%) in the success rates between the two datasets can be due to the naturalness of the emotions elicited by IEMOCAP.

A different research[10] showed that the regression model had unexpectedly poor outcomes in Audio Emotion Recognition, in contrast to the previous research findings

which suggested regression models work well for MER. Their ANN model showed variance for 64.4% and 65.4% for arousal and valence prediction respectively.

Finally, an alternate paper[9] concludes with the finding that CNNs are the better at emotion recognition due to its higher low-level and short term discriminative capabilities. This, combined with the addition of LSTM networks not only allows the model to identify long-term paralinguistic patterns, it also has shown higher capabilities of speaker-independent emotion recognition.

2.8   Recent Trends

The analysis of review papers suggests that SER has made significant advancements in recent years. According to authors[12], traditional SER models use manually extracted features such as MFCCs, pitch, spectral features whereas modern approaches depend on deep representation learning to learn features directly from raw audio using neural networks reducing the need for feature extraction; consequently, reducing computational cost of a model. On the other hand, employment of self-supervised learning (SSL) for feature extraction have also been observed. Additionally, models with SSL (unsupervised learning) can easily work on unseen dataset without needing additional fine-tuning. VAEs (Variational Autoencoders) and GANs (Generative Adversarial Networks) are used to modify the dataset for better emotion recognition (i.e. data augmentation and emotion transformation). To introduce further generalization in recent models, transfer learning and domain adaptation have been used. A futer study found[11] how incorporating attention mechanisms with SER models (i.e. RNN, LSTM, Transformer models) improves efficiency as it focuses more meaningful parts of speech

2.9   Gaps in Research Field

While there has been progress, there are still some unresolved issues. Present models do not consistently recognize tonal variations across languages and accents. Present models also become less effective in noisy environments, limiting areas of real-world application. Deep learning-based SER systems also lack interpretability, leaving users with no explanations for model decisions and, as a result, making it difficult to earn users' trust[18]. Systems also fail to consistently detect conflicting emotions, and mixtures of emotional expressions have resulted in misclassification and inaccuracy. For SER to be applied practically, these limitations must be overcome.

Chapter 3

Conclusion

This study gives us an overview of a research that is recent in Speech Emotion Recognition (SER), which displays its progress from acoustic-based models to machine learning and in a more recent form, deep learning approaches. Techniques involving feature extraction have also evolved in a similar way, which transitioned from manual extraction to the integration of deep neural networks beside the traditional methods, which enhanced model robustness. Besides these advancements, SER systems continue to face problems, especially in differentiating overlapping or conflicting emotions, which leads to misclassification and therefore reduces the effectiveness of the models. The mission of this research is to make a model that facilitates more natural and emotional communication between humans and computers, this will improve human-computer interaction and advance affective computing.

Bibliography

[1] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, 1980.

[2] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3-4, pp. 169–200, 1992. DOI: 10.1080/02699939208411068. [Online]. Available: https://doi.org/10.1080/02699939208411068.

[3] P. Harar, R. Burget, and M. K. Dutta, "Speech emotion recognition with deep learning," in *2017 IEEE International Conference on Signal Processing and Integrated Networks (SPIN)*, 2017, pp. 137–140. DOI: 10.1109/SPIN.2017.8049931.

[4] K. Tarunika, R. Pradeeba, and A. P, "Applying machine learning techniques for speech emotion recognition," Jul. 2018, pp. 1–5. DOI: 10.1109/ICCCNT.2018.8494104.

[5] N. Tits, K. E. Haddad, and T. Dutoit, "Asr-based features for emotion recognition: A transfer learning approach," in *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 48–52. DOI: 10.18653/v1/W18-3307. [Online]. Available: https://aclanthology.org/W18-3307/.

[6] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, and C. Cléder, "Automatic speech emotion recognition using machine learning," in *Social Media and Machine Learning [Working Title]*, IntechOpen, 2019. DOI: 10.5772/intechopen.84856. [Online]. Available: https://hal.science/hal-02432557.

[7] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117 327–117 345, 2019. DOI: 10.1109/ACCESS.2019.2936124.

[8] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, 2019, pp. 1–6. DOI: 10.1109/RADIOELEK.2019.8733432.

[9] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041249. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1249.

[10] S. Cunningham, H. Ridley, J. Weinel, and R. Picking, "Supervised machine learning for audio emotion recognition," *Personal and Ubiquitous Computing*, vol. 25, no. 4, pp. 637–650, 2021. DOI: 10.xxxx/xxxx.

[11] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021, ISSN: 2079-9292. DOI: 10.3390/electronics10101163. [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1163.

[12] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Qadir, and B. Schuller, "Survey of deep representation learning for speech emotion recognition," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1634–1654, 2023. DOI: 10.1109/TAFFC.2021.3114365.

[13] M. M. Rezapour Mashhadi and K. Osei-Bonsu, "Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest," *PLOS ONE*, vol. 18, no. 11, pp. 1–13, Nov. 2023. DOI: 10.1371/journal.pone.0291500. [Online]. Available: https://doi.org/10.1371/journal.pone.0291500.

[14] M. Vallejo, *Emotions vs. feelings vs. moods: Key differences*, Accessed: 2025-2-1, 2023. [Online]. Available: https://mentalhealthcenterkids.com/blogs/articles/emotions-vs-feelings-vs-moods.

[15] S. M. George and P. M. Ilyas, "A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise," *Neurocomputing*, vol. 568, 2024, ISSN: 0925-2312. DOI: 10.1016/j.neucom.2023.127015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231223011384.

[16] S. W. II, *What is speech emotion recognition?* Accessed: 2025-02-01, 2024. [Online]. Available: https://klu.ai/glossary/speech-emotion-recognition.

[17] S. Kalateh, L. A. Estrada-Jimenez, S. Nikghadam-Hojjati, and J. Barata, "A systematic review on multimodal emotion recognition: Building blocks, current state, applications, and challenges," *IEEE Access*, vol. 12, pp. 103 976–104 019, 2024. DOI: 10.1109/ACCESS.2024.3430850.

[18] M. Ramaswamy and S. Palaniswamy, "Multimodal emotion recognition: A comprehensive review, trends, and challenges," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, Oct. 2024. DOI: 10.1002/widm.1563.