UNIVERSITY OF NOTTINGHAM MALAYSIA

PhD Thesis

Speech Emotion Recognition (SER) System for Late-Deafened Educators in Online Teaching

Student:

Aparna Vyakaranam
20351655

Supervisor:

Dr.Tomas Maul

Co-Supervisor:
Dr.Bavani Ramayah

School of Computer Science

Faculty of Science and Engineering

AUG 2025

Acknowledgements

I am deeply grateful to Dr Tomas for his support, guidance, and mentorship throughout my PhD journey. His expertise, structured approach, and insightful advice have been instrumental in shaping both my research and personal growth. I feel truly privileged to be his student, to work under his supervision, and to learn from such a remarkable scholar and guru.

I would also like to express my heartfelt gratitude to my co-supervisor, Dr Bavani, for her constructive feedback and insights, which have significantly contributed to my work. Her supportive and helpful nature has been invaluable throughout this journey. My sincere thanks go to Dr Zhiyuan Chen, my internal examiner, for her thorough evaluation and suggestions.

I want to express my deepest gratitude to those who have supported and encouraged me throughout this journey. To my husband for his unwavering support, advice, and belief in me, and to my parents, mother-in-law and daughters, thank you for your love, encouragement, and patience. A special note to my father, whose body of work has inspired me and whose support has been my greatest strength. Additionally, I would like to express my profound gratitude to two remarkable individuals who have inspired my journey: my friend Sanjukta and my mother, both of whom are late-deafened individuals with differing causes of hearing loss. Interacting with them and understanding their experiences have deeply motivated this research.

Sanjukta, a highly qualified academic affiliated with a prestigious university, is engaged in both research and teaching. She developed hearing difficulties later in life, which presented significant challenges in her professional journey. Despite these obstacles, she has shown remarkable resilience. Her struggles during online teaching—particularly due to limited or no access to students' facial expressions, lip movements, or body language—underscored the importance of innovative solutions for making education more inclusive, especially for individuals with disabilities. My conversations with her and the insights she shared have been instrumental in shaping my perspective and inspiring this work. I dedicate this work to her as a token of my gratitude for being a guiding light throughout my PhD journey.

My mother, who also has a hearing impairment, has been a constant source of strength and resilience. Her ability to adapt and thrive despite the challenges she faces has profoundly

influenced my outlook on life. Her determination and positivity have inspired me to pursue solutions that empower individuals with similar challenges.

Publications and Conferences

- A Review on Speech Emotion Recognition for Late Deafened Educators in Online Education
  Aparna Vyakaranam, Tomas Maul and Bavani Ramayah
  Published in: International Journal of Speech Technology (2024) 27:29–52
  https://doi.org/10.1007/s10772-023-10064-7
- Preliminary Study: Speech Emotion Recognition in Online Teaching From the Perspective of Educators Especially Late Deafened
  Aparna Vyakaranam, Bavani Ramayah and Tomas Maul
  Published in: 2024 2nd International Conference on Software Engineering and Information Technology (ICoSEIT)
  DOI:10.1109/ICoSEIT60086.2024.10497503
- Usability and User Experience of a Speech Emotion Recognition System with and without Hearing Impaired Educators
  Aparna Vyakaranam, Bavani Ramayah and Tomas Maul
  Submitted to: Educational Technology Research and Development
- Comparison of Three Hybrid Architectures Using 1D, 2D, and 3D CNNs for Speech Emotion Recognition
  Aparna Vyakaranam, Tomas Maul and Bavani Ramayah
  Published in: International Journal of Speech Technology. Year 2025.
  DOI: 10.1007/s10772-025-10204-1

Abstract

Speech emotion recognition (SER) involves predicting human emotions from speech signals, aiding in the understanding of human behaviour and offering opportunities in human-computer interaction (HCI). It is widely applicable across domains such as psychology, medicine, education, and entertainment. This research explores the development of an SER system to support late-deafened educators in online teaching environments.

A review of relevant literature highlighted the importance of emotional engagement, defined as students' emotional responses to academic content, which is essential for effective learning and often conveyed through vocal and behavioural cues. However, in online classes, such non-verbal cues are limited due to the lack of physical presence, resulting to what is referred to as emotional deficiency. This challenge is particularly significant for late-deafened educators, who may find it difficult to hear or interpret verbal feedback, making it harder to gauge student emotions and engagement. To address this, a real-world SER system was developed to detect and display student emotions from verbal feedback accurately and in real time. The aim was to help late-deafened educators better understand student engagement and adjust their teaching strategies accordingly during online classes.

A preliminary study indicated emotional deficiency in online classes and highlighted the value of integrating emotional feedback into online teaching environments. The proposed system extracted acoustic features such as Zero Crossing Rate (ZCR), Root Mean Square (RMS), Chroma-STFT, Mel Frequency Cepstral Coefficients (MFCCs), and Mel-spectrograms. Three hybrid CNN architectures combining 1D, 2D, and 3D layers were explored through a novel comparative analysis using fusion strategies: averaging, parallel merging, and sequential integration. These models were evaluated on five benchmark datasets—IEMOCAP, DEMoS, TESS, RAVDESS, and EMO-DB. The averaging fusion model consistently outperformed the others, achieving accuracies of 82% on IEMOCAP, 91% on DEMoS, EMO-DB, and RAVDESS, and 100% on TESS, and was therefore selected for implementation.

The final system featured a user-friendly graphical user interface (GUI) and was evaluated for usability and user experience through testing with educators both with and without hearing impairment. Quantitative results showed that 90% of users found the system intuitive and effective for real-time emotion detection, and 80% of late-deafened educators reported it accurately captured student emotions. Qualitative feedback further emphasized its value in helping educators tailor instruction based on emotional cues. This research demonstrates a very high practical value of integrating SER into online teaching to enhance late-deafened educators' awareness of student emotional engagement and to support more adaptive teaching strategies. However, the developed system relies solely only on five discrete, universally accepted human emotions, which are further classified into positive or negative emotions. It relies partly on acted speech datasets, which may not fully capture the subtle, diverse, and spontaneous expressions typical in real classrooms. The absence of multimodal cues, such as facial expressions or textual input, limits the system's ability to provide a holistic understanding of student emotions in real time. Future enhancements can include expanding emotion categories to cover education-specific states like confusion or boredom, and integrating multimodal cues (facial expressions, text and such) to improve real-time accuracy and contextual understanding across diverse learning environments.

Contents

List of Tables

List of Figures

Chapter 1

1. Introduction

1.1 Overview

Speech emotion recognition (SER) system for late-deafened educators in online teaching plays a crucial role in enhancing human-computer interaction. An emotional gap in online education impacts effective learning interactions [1]. Students' vocal responses in online classes convey emotions that reflect their level of understanding [2]. This insight can help educators adjust teaching strategies based on students' emotional and behavioural engagement. However, detecting student emotions during online classes is challenging [3], particularly for late-deafened educators with hearing limitations. Consequently, student feedback may go unheard and their emotions undetected. An effective SER system could bridge this gap by detecting emotions from students' verbal feedback and displaying them in an accessible format for the educator. Despite their potential, developing SER systems that perform reliably in real-world scenarios remains challenging [4]. This work presents a real-world SER system for late-deafened educators, aiming to bridge the emotional communication gap in online teaching. Our system leverages a deep learning-based approach using Convolutional Neural Networks (CNNs), which seeks to detect emotions accurately in real-time. This work compares a hybrid CNN architecture that integrates 1D, 2D, and 3D CNN layers using three fusion techniques: averaging, parallel merging, and sequential integration. The averaging fusion technique, which demonstrated the highest accuracy, was selected for its suitability in real-world applications. To enhance reliability, the model was trained with five datasets comprising both natural and acted emotional speech and employed a combination of temporal, spectral, and deep learning-based features to improve detection accuracy. The finalized real-time SER system, with an accessible Graphical User Interface (GUI), was evaluated by target respondents who were educators with and without hearing impairments. Results indicate that the SER system can significantly enhance teaching effectiveness and efficiency, making it a valuable tool for late-deafened educators in online settings.

The following sections provide background information on this research idea to establish a clear foundation before stating the research questions and objectives.

1.2 Online Teaching and Learning

1.2.1 Overview of Online Education, Teaching and Learning

“Online education is defined as education being delivered in an online environment through the use of the Internet for teaching and learning. This includes online learning on the part of the students that is not dependent on their physical or virtual co-location. The teaching content is delivered online, and the instructors develop teaching modules that enhance learning and interactivity in the synchronous or asynchronous environment.” [5]. This is a contrast with “Traditional classroom setting” or face-to-face (F2F) instruction, the main difference being the mode of interaction and physical presence. In a traditional classroom, instructors and students interact face-to-face in a shared physical space, allowing immediate feedback and non-verbal cues. On the other hand, online education takes place over the Internet, enabling participation without requiring physical presence, often relying on digital tools for interaction, which can be synchronous or asynchronous [5, 6].

Online learning, a subset of online education [5], has expanded significantly in recent years as an alternative or complement to traditional F2F education. A key factor in this growth is that students are active technology users who regularly engage with digital tools [7]. In fact, during the COVID-19 pandemic, there was a forced and largely unprepared move to fully online education for educators and students. High infection rates and social distancing measures forced the closure of educational institutions worldwide. This caused a forced transition to online classes. Educators had to quickly implement comprehensive online courses that mirrored traditional classes to minimize disruptions in the academic schedule (UNESCO IESALC, 2020) [8]. This presented challenges, as educators were required to design online classes resembling physical classrooms, which also demanded suitable devices and platforms for effective online teaching [9]. Before the pandemic, “Online Education” was primarily viewed as a supplement to traditional in-person classes. However, it became the primary mode of instruction for all educational institutions during the pandemic [10]. Numerous online teaching tools, including learning management systems (LMSs), have been developed over the past decades to support online education [7].

Online teaching, an essential element of online education, refers to instruction delivered through technology-based platforms like video conferencing (e.g., Zoom) [5, 6]. Online teaching tools offer features such as real-time video conferencing, the ability to upload class materials, conduct assessments, and facilitate live chat, among others. Students appreciate this format due to their comfort with technology and the flexibility to learn from anywhere, as was evident during the COVID-19 pandemic [9, 11]. Although both instructors and students have adapted to this mode and continue to utilize it, challenges remain. One such challenge is that educators find it difficult to gauge students' facial expressions and emotional states during online teaching sessions, making it harder to adapt their teaching style to enhance students' understanding [7, 9]. This limitation can also negatively affect educators' motivation.

1.2.2 Importance of Students' Emotions in Learning and Engagement

The American Psychological Association (APA) [12] defines emotion as " a complex reaction pattern, involving experiential, behavioural, and physiological elements, by which an individual attempts to deal with a personally significant matter or event". In the context of education, student emotions play a significant role in academic settings, impacting both academic success and personal growth. Students experience many emotions associated with understanding classroom instructions, engaging in learning, and achieving academic goals. Positive emotions can motivate students, enhancing their enjoyment of learning and pride in their accomplishments. Conversely, negative emotions may hinder performance and even increase the risk of student distress, like anxiety or disengagement [13, 14].

1.2.3 Student Emotions as Affective Feedback for Educators' Self-Regulation Practices in Online Teaching

Students' engagement during class interactions is closely linked to their conceptual understanding, particularly emotional engagement, which refers to students' emotional responses to academic subjects [2]. Given this, educators encourage interaction by seeking feedback or asking questions about the ongoing topic. Students' vocal responses and behaviour reflect their level of understanding. This feedback helps educators self-regulate and adjust their teaching strategies based on students' emotional and behavioural engagement. Feedback is crucial in teaching-learning [7] as it offers specific information about a learning task or process, helping bridge the gap between the desired and actual understanding of the content. In an online setting, it is even more critical due to the lack of physical presence and missing of many non-verbal cues.

1.2.4 Challenges Faced by Late-Deafened Educators in Interpreting Student's Emotions during Online Teaching

During online classes, students respond to educators' queries through voice, video, chat, or other options provided by the online platform. While educators receive these responses, they sometimes struggle to interpret the emotions behind them fully. Emotions are reactions influenced by mood, circumstances, and other factors expressed through body language, facial expressions, and voice [15]. These emotions are crucial in student learning and engagement [16]. However, online class often lacks the emotional depth of in-person interactions, leading to what researchers describe as an "emotional deficiency" [3]. This deficiency makes it challenging for educators to gauge how students feel about the class and to determine whether a student is on track or requires additional support [17]. For educators who are late-deafened, the difficulty is further compounded by the inability to access auditory cues, which limits their ability to interpret emotions from students' feedback. According to the National Deaf Centre, USA, "late-deafened" individuals are those who have acquired oral communication skills before losing their hearing [18]. While their verbal skills remain intact, hearing loss becomes a significant obstacle. Late-deafened educators fall into this category. Developing solutions for student emotion detection during online classes, for use in online teaching, is essential, as they would inherently be inclusive and particularly beneficial for late-deafened educators.

1.3 Problem Statement

As pointed out in Section 1.2.3, emotional engagement, reflected in students' vocal and behavioural feedback, is crucial for guiding teaching—especially in online settings where non-verbal cues are limited. However, as discussed in Section 1.2.4, there is an "emotional gap" in online classes, which is especially challenging for late-deafened educators. Due to their hearing loss, they often struggle to gauge student emotions through verbal feedback during online classes. These challenges highlight a need for solutions specifically supporting them, ultimately benefiting all educators. While an SER system holds promise in addressing this issue, its real-world application in supporting late-deafened educators remains largely unexplored. Building an effective SER system that can accurately detect student emotions in real time is still a significant challenge. This study aims to address this gap by developing a real-world SER system tailored to help late-deafened educators identify student emotions more accurately during online teaching. This would enhance the teaching and learning experience, achieving engagement levels comparable to physical classrooms.

1.4 Key Research Idea

1.4.1 Overview

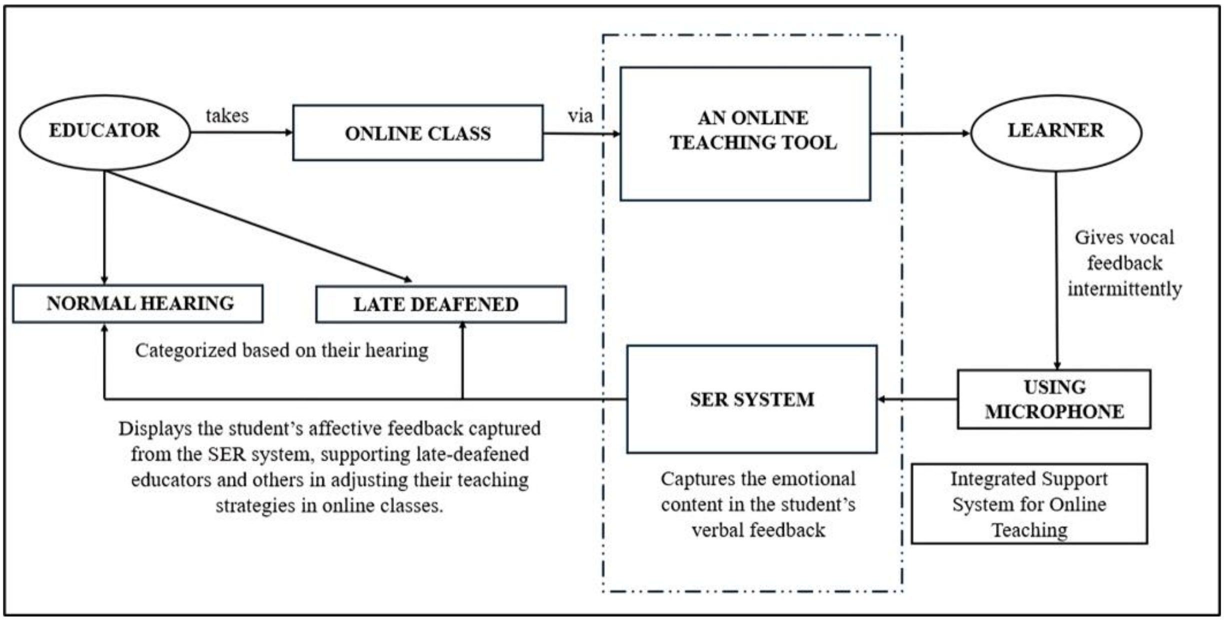

EDUCATOR
takes
ONLINE CLASS
via
AN ONLINE TEACHING TOOL
LEARNER
Gives vocal feedback intermittently
NORMAL HEARING
LATE DEAFENED
Categorized based on their hearing
SER SYSTEM
USING MICROPHONE
Displays the student's affective feedback captured from the SER system, supporting late-deafened educators and others in adjusting their teaching strategies in online classes.
Captures the emotional content in the student's verbal feedback
Integrated Support System for Online Teaching

Figure 1.1: Research idea overview

As shown in Figure 1.1, educators, including those with normal hearing and those who are late-deafened, conduct online classes using suitable teaching tools. Learners participate in this remote learning through the same platform. Educators perform various tasks through these tools, such as delivering lessons, uploading materials, providing feedback, and conducting assessments. During online classes, educators periodically gather verbal feedback from students to assess comprehension and determine whether further clarification or examples are needed. The emotions expressed in student feedback are crucial for understanding engagement and learning outcomes, as highlighted in Section 1.2.2. An SER system can automatically classify the emotions conveyed in student verbal feedback and display this information to educators. This would help educators adjust their teaching strategy when needed. The detected emotion would appear on the educator's screen as a visual representation with an appropriate message. This affective feedback is particularly beneficial for late-deafened educators, who may face challenges hearing verbal feedback and interpreting emotional cues. By responding to the detected emotions, these educators can adapt their teaching approaches as needed, thereby enhancing student engagement. The following section provides a brief overview of SER systems.

1.4.2 Speech Emotion Recognition

SER is a key challenge in Human-Computer Interaction (HCI). Its successful implementation as a real-world application can help address challenges faced by late-deafened educators, supporting them in their teaching and training activities, as outlined earlier. SER system is a set of methodologies that process and classify speech signals to identify underlying emotions [19]. These systems can help detect discrete emotions such as fear, anger, neutrality, sadness, happiness, and surprise.

Deep Learning, a branch of machine learning, is popularly used by researchers to train SER models, giving high accuracy [20]. Hence, deep learning has been used to develop the SER system for this research. Figure 1.2 illustrates a typical flow diagram for an SER system [21]. The process begins with acquiring datasets containing audio recordings of speech samples labelled with corresponding emotions. The next stage, feature selection, involves pre-processing the audio (e.g., noise removal) and extracting relevant features. Finally, using machine learning or deep learning, emotions are classified as anger, happiness, sadness, and such.

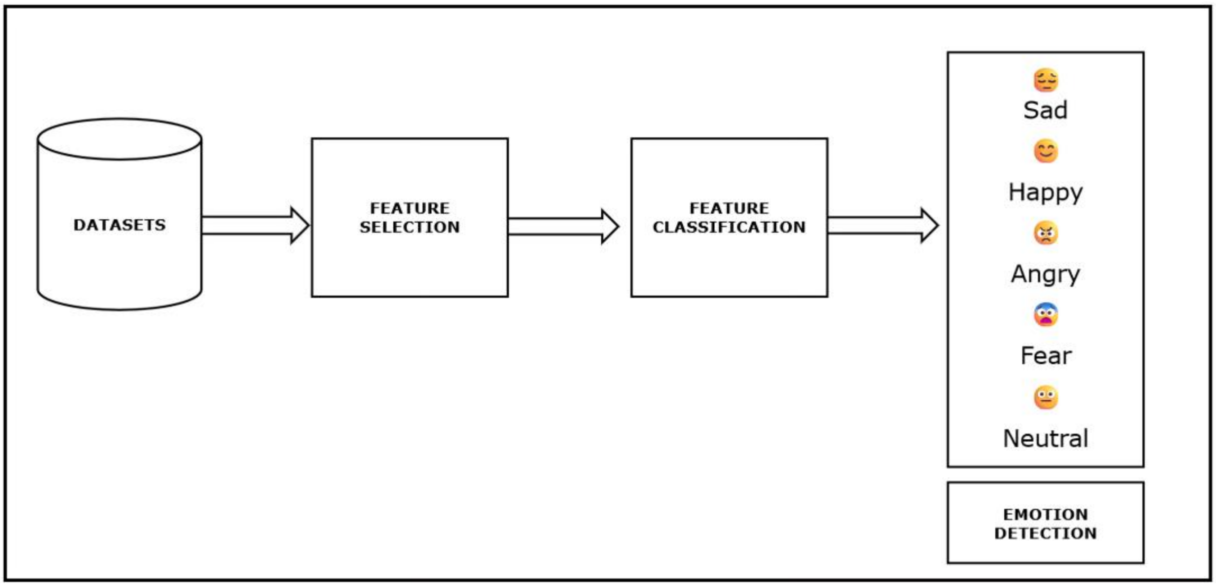


Figure 1.2. General flow diagram of an SER system

While SER systems offer powerful tools for emotion detection, their effective design and integration into real-world environments require alignment with HCI principles. HCI serves as the framework for ensuring these systems are user-friendly, efficient, effective, and capable of addressing the needs of late-deafened educators. The next section delves into the HCI framework and its role in this research.

1.4.3 Human-Computer Interaction (HCI): Designing for Late-Deafened Educators

HCI, a subfield of computer and information sciences, has been a significant area of study since the early 1960s [22]. HCI design aims to create a seamless alignment between the user, the machine, and the required services to achieve both the high-quality and optimal performance of those services [23]. As per P.Zhang and D.Galletta in their book *Human-computer interaction and management information systems: Foundations,* "Human-computer interaction is a discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them". HCI focuses on building a strong connection between humans and technology. It plays a crucial role in system design, where the design quality is measured in terms of effectiveness, efficiency, and satisfaction by the humans interacting with the technology [22, 23].

Audio-based HCI is a fascinating field of HCI that gathers information via various audio signals. An SER system is one such application of HCI design [23]. SER systems specifically address the interaction between humans and computers through speech to recognize and interpret the emotional content of spoken language. SER systems represent a significant area of HCI research that explores the intersection of human emotion, communication, and technology. SER is categorized as an intelligent HCI design as it integrates Artificial Intelligence techniques. The application of methods emerging from the Artificial Intelligence area allows the design of intelligent interfaces that try to bridge the gap between the user and the system [24]. By developing and evaluating these intelligent systems, HCI researchers aim to create more empathetic and effective human-computer interactions that enrich the user experience and address several real-world needs and challenges. The aim is to align the system's technological capabilities with user requirements. The general process in HCI, which can vary, includes (1) designing the research, (2) conducting data collection, and (3) reporting the findings [25]. For this research, a hybrid software development life cycle (SDLC) model has been employed, combining principles from both the Incremental and Iterative approaches for HCI development process [26].

Late-deafened educators who teach online are already adept at using online tools such as Zoom and MS Teams and such. The purpose of the SER system is to enhance their existing HCI experience and provide additional support. SER, which can be seamlessly integrated with any online teaching platform, can be adopted by all educators with minimal complexity. The system has been designed to accommodate user diversity.

1.5 Research Questions and Approach

Discussions in Section 1.3 highlight the need to address the lack of emotional depth in online class. Addressing this gap could enhance the teaching and learning experience, achieving engagement levels comparable to physical classrooms. An efficient SER system that accurately detects real-time emotions could bridge this gap. A real-world application of SER giving accurate results is still a big challenge [20, 27, 28]. This study focuses on developing an SER system tailored for late-deafened educators, enabling them to identify student emotions accurately during online education. Currently, there is a lack of computational solutions to support late-deafened educators in this context [29, 30]. This study aims to develop a real-world SER system for late-deafened educators to detect student emotions accurately during online classes, for use in online teaching. This study is directed to answer the following research questions:

- RQ1 - What are efficient and effective pre-processing and feature engineering approaches for a deep learning-based real-world speech emotion recognition (SER) system?
- RQ2 - In the context of deep learning, specifically convolutional neural networks (CNNs), what is an effective hybridization approach for combining 1D, 2D, and 3D convolutional layers to improve accuracy whilst maintaining efficiency?
- RQ3 - How can emotional insights from an SER system be shown in a way that educators, especially late deafened, find it easy to understand and use during online classes?
- RQ4 - What is the perceived impact of an SER system built upon RQ1-RQ3 on educators with hearing impairment and educators without hearing impairment regarding outcome and experience?

1.6 Research Objectives

The Objectives are:

- RO1 - To identify the key speech features and feature engineering approaches that enhance the efficiency and accuracy of emotion recognition in a deep learning-based real-world SER system.
- RO2 - To compare and assess the effectiveness of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers through three selected fusion techniques, identifying the approach with the highest accuracy and efficiency and adapting it to build a real-world SER application.
- RO3 - To design and develop a graphical user interface (GUI) that integrates SER-derived emotional feedback and displays emotions to late-deafened educators accurately and in real-time.
- RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency and perceived impact on late deafened educators for an effective engagement in online teaching.

1.7 Significance and Scope of the Study

1.7.1 Research Significance

This research is significant as it enables real-time detection of student emotions from speech via a SER system during an online class. This system helps late-deafened educators overcome the communication challenges posed by hearing loss, enhancing their ability to interpret student engagement and adjust teaching strategies accordingly. Built on a hybrid CNN architecture and validated across multiple datasets, the system not only demonstrates high accuracy but also delivers practical value through a user-friendly interface. Its positive reception among educators both with and without hearing impairment highlights its potential as an inclusive tool that enhances teaching effectiveness by helping educators better understand and respond to students' emotional behaviours, thereby supporting accessibility in digital classrooms.

1.7.2 Research Scope

This research focuses on a unimodal speech-based emotion recognition system that specifically supports the teaching needs of late-deafened educators in online educational settings. The system explores three distinct hybrid CNN architectures that integrate 1D, 2D, and 3D CNN layers to effectively capture and process various acoustic features of speech. It is validated across five datasets: two semi-natural and three acted. The emotion recognition component is built around five basic discrete emotions—happy, sad, angry, fear, and neutral. For clarity and efficiency in feedback delivery, the system further categorizes these emotions into positive and negative classes, enhancing its usability for real-time teaching support.

1.8 Novel Contributions

The contributions of this study are:

1. This study assessed the necessity and potential impact of an SER system in online teaching. It uniquely addresses the perspectives of late-deafened educators, offering insights into an area that has not been extensively explored.
2. This study contributes to the development of an SER system through a unique comparison that integrates 1D, 2D, and 3D CNNs. It uses three different fusion techniques—averaging, parallel merging, and sequential integration—to identify the most suitable model for achieving high accuracy. The selected model was further developed into a real-world application for live testing.
3. A user-friendly interface was designed and integrated to display SER-derived emotional information accurately. This interface bridges the emotional gap experienced during online teaching, encouraging affective engagement between late-deafened educators and students—an area that has not been sufficiently addressed.
4. The system's usability and user experience were evaluated with educators both with and without hearing impairments, showing its potential to improve teaching effectiveness in online teaching.

1.9 Structure of this Thesis

This thesis is structured as follows:

In Chapter 2, a literature review is conducted to comprehensively understand SER systems and identify gaps in their implementation for real-time emotion detection in online education for late-deafened educators. The first section of the review introduces the fundamentals of SER, followed by an analysis of existing research and identifying potential pathways to develop real-world SER applications capable of accurately detecting emotions. The second section focuses on HCI approaches, addressing late-deafened educators' unique challenges and requirements.

Chapter 3 outlines the methodology related to the proposed SER system in response to the limitations of existing SER systems identified in Chapter 2. This section is divided into four parts: 1). methodology for a preliminary study that assesses the potential benefits of using an

SER system for late-deafened educators in online classes; 2). the methodology for developing the SER system, where three hybrid CNN (deep learning) architectures are proposed and compared, with the model that achieves the highest accuracy in emotion detection being selected as the final model; 3). the methodology for transforming the selected SER system into a functional, real-time application capable of detecting and displaying emotions; 4). and lastly, the methodology for evaluating the usability and user experience of the developed SER system in supporting late-deafened educators during online teaching.

In Chapter 4, results are presented, offering insights into each phase of the research. The results are divided into four sections in the following sequence: 1) The first section discusses the findings from the preliminary study, focusing on the needs and perspectives of late-deafened educators. 2) The following section presents the results from experiments conducted using three hybrid models, comparing their performance and identifying the model with the highest speech emotion recognition accuracy. 3) The next section covers the results of integrating the selected SER model with the graphical user interface, emphasizing its functionality as a real-time emotion detection application. 4) and the final section discusses the outcomes of evaluating the developed SER system for its effectiveness, efficiency, and perceived impact on late-deafened educators' engagement in online education

In Chapter 5, conclusions are drawn regarding the impact and the potential future research opportunities made possible by the contributions of this work.

Chapter 2

2. Review of Literature

2.1 Introduction

In Chapter 1, the background of this research is outlined, highlighting the challenges faced by late-deafened educators in understanding student emotions during online classes. The proposed solution is a real-world SER system based on a deep learning approach capable of detecting student emotions in real-time. A literature review was conducted to thoroughly explore SER systems as potential support mechanisms for late-deafened educators in online education. This review is systematically divided into two sections. The first section introduces the fundamentals of SER, detailing key components such as feature extraction and classification techniques and explaining its underlying principles. This foundational overview provides the context for analysing existing research and identifying pathways to developing SER applications. A key objective is exploring deep learning approaches, especially using hybrid CNN architecture to build an efficient SER system for real-time emotion detection. The second section of this review delves into HCI approaches, particularly emphasising the unique challenges and needs of late-deafened educators. It explores the gaps between technology and accessibility for late-deafened educators. It highlights the requirements of this group, provides an overview of hearing impairments and their impact, and examines the specific difficulties faced by individuals with hearing loss in workplace environments. It also reviews existing HCI solutions designed for late-deafened individuals, emphasizing the need for more inclusive and effective technologies. This discussion sets the stage for identifying areas where advancements in HCI can bridge the gap between accessibility and technological support.

2.2 Speech Emotion Recognition System

Among humans, interactions happen through multiple channels for information exchange, such as facial expressions, body language or physical gestures, verbal communication, and such. Of these sources, speech is likely the most readily accessible [27]. Speech or verbal communication is the most natural form of communication. It is very significant as it holds a lot of paralinguistic and linguistic information. It conveys explicit ideas and contains implicit cues that carry significant meaning. It is a complex signal that contains information about the message, speaker, language, and emotion. This information can be found in the tone, volume, and pace of speech and the emotional demeanour of the individuals involved. Emotion is a feedback process, with some stimuli acting as key triggers that set off the emotional response. Emotion makes the speech expressive and, hence, effective [14]. Since emotions facilitate a deeper understanding between humans, extending this comprehension to computers would be a natural consequence [31]. Basic emotions—happiness, anger, fear, boredom, sadness, disgust, and neutrality—are foundational in evolutionary theory. Emotion recognition, a growing multidisciplinary field, enhances human-machine interaction by enabling computers to interpret and respond to emotions [32, 33]. As an important part of HCI, SER helps to recognise emotions from speech signals. It extracts emotions from a speakers' speech signal by pre-processing the raw signal, classifying key features, and displaying the processed results [34, 35]. As illustrated in figure 2.1, the SER process can be divided into many distinct components [34, 36]. In the figure 2.1, emotions embedded in the datasets on the left are extracted by a classifier positioned on the far right.

Overview of SER system

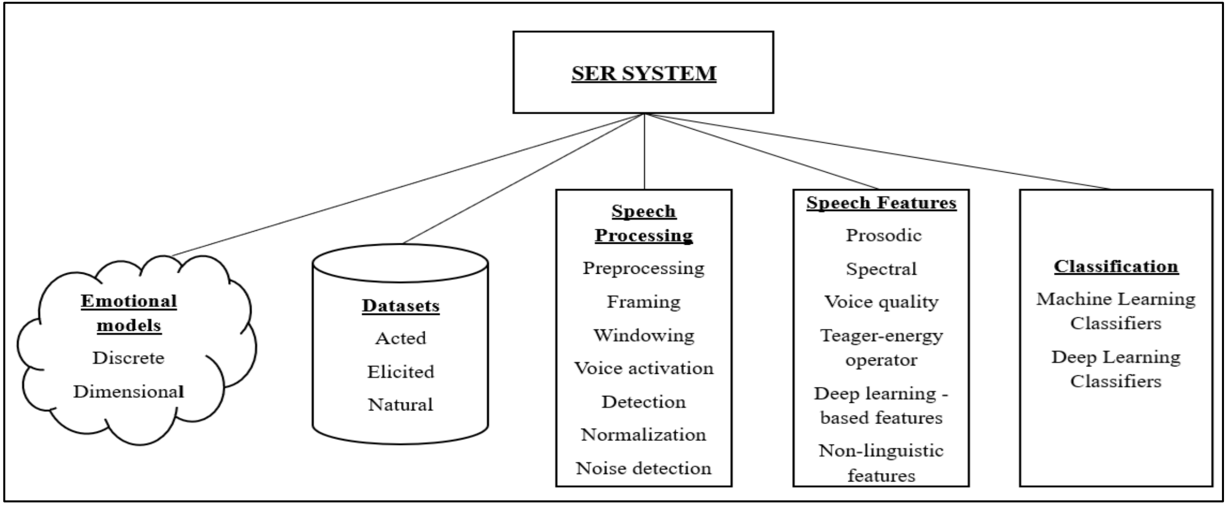


Figure 2.1: Overview of key SER components [34, 36]

Each component is essential to the emotion recognition process; hence, each component is discussed in the next section. This understanding is crucial to gaining a complete perspective of the process and its key elements, enabling a systematic review of the SER literature and identifying existing research gaps. This study further explores traditional SER techniques and more recent, widely adopted deep learning approaches, focusing on the techniques used to extract emotions from speech signals. The following section covers all areas of SER, followed by a review.

2.3 SER Components

2.3.1 Emotional Models

Emotions are complicated as they depend on personal experience, physiology, behaviour and communication responses [14, 15, 33]. As given in [33], many definitions are proposed for emotions. Analysing emotions that can be used for any application needs to be defined and modelled properly. The most popular models that can be used in SER consist of the discrete and dimensional emotional models. The discrete emotion model, as described by Ekman et al. in their book "Emotion in the human face: Guidelines for research and an integration of findings", consists of six categories of basic emotions, which are "sadness, happiness, fear, anger, disgust, and surprise". These universal and culturally independent emotions, which are experienced briefly, were first described by Ekman and further elaborated by Ekman and Oster and Ekman et al. [19, 37, 38]. Since they are widespread in everyday life, they are easy to model and label in SER systems. The acoustic features of these speech signals, used for emotion recognition, exhibit similarities across different languages, allowing for the utilization of a common classification model [27].

On the other hand, the dimensional emotional model has several disadvantages for representation as it focuses on a few latent dimensions like valence, arousal, control, power, and such. Characterization of emotions with this model is very challenging as the emotions in this case are not intuitive. Labelling these emotions requires special training [19, 39, 40].

2.3.2 Datasets

In the SER system, the effectiveness of the classification process relies heavily on the volume and quality of the data [41]. Noisy, inaccurate, or incomplete data can significantly reduce SER accuracy. Therefore, data for training and testing must be meticulously gathered, and annotated. Based on the emotion-labelled speech data, the datasets are classified as acted, elicited or natural. The three types of databases or datasets that can be used for SER [20, 21, 27, 42, 34] are discussed below:

2.3.2.1 Acted (Simulated) Datasets

These datasets are generated by trained performers who read the exact text with varying emotions in soundproof studios. While they are the simplest type to produce, they often fail to capture genuine real-life emotions, as performances can sometimes be exaggerated, leading to reduced recognition accuracy for natural emotions. However, they are highly valuable in theoretical research. Such datasets include CASIA, EMOVO, SAVEE, TESS, and CREMA-D [32, 43, 44].

2.3.2.2 Elicited or Semi-Natural (Induced) Datasets

Although simulated, these datasets are more naturalistic, creating an artificial emotional scenario in which the speaker simulates various emotions. These emotions closely resemble real ones, making this dataset highly suitable for real-time applications. Examples include eNTERFACE05 and IEMOCAP [43, 45, 46].

2.3.2.3 Natural Datasets

These datasets consist of speech captured from live recordings, such as TV shows, YouTube videos, call centre interactions and radio broadcasts, and are often referred to as spontaneous speech. Collecting this type of data can be challenging due to potential legal and ethical issues in processing or distributing it. Examples include BAUM-1, call centre recordings, and CHEAVD [21, 27, 47].

Acted datasets provide clear and exaggerated emotions that are easy to label and widely available, making them useful for training and evaluating SER models. Semi-natural datasets, captured in controlled but less scripted settings, offer more authentic expressions while retaining some experimental control. They are valuable for real-world applications as they strike a balance between emotional authenticity and controlled conditions, enabling more realistic yet manageable model training. Both acted and semi-natural data are easier to collect and annotate compared to natural datasets. Although natural datasets offer the highest realism for building models that generalise well to real-life scenarios, they are limited in availability, pose potential copyright and privacy concerns, and are often noisy or ambiguously labelled [20, 34, 43, 44]. Thus, selecting a dataset type involves a trade-off between realism that is needed for practical applicability and control, which ensures consistency and quality in data collection and annotation.

2.3.3 Speech Processing

Speech processing involves manoeuvring the speech signals to extract essential characteristics from them, which is achieved by the following steps -

2.3.3.1 Pre-processing

The first step after selecting the relevant datasets is to preprocess the samples available in the datasets. Pre-processing involves unwanted noise removal, silence removal, normalization [27, 48], and other procedures to be performed on the speech samples. This would help get higher-quality speech sequences that can be used for feature extraction, which further gives way to emotion recognition.

2.3.3.2 Framing

Framing is a pre-processing technique that divides the speech signal into small, fixed-length segments known as frames [44]. This segmentation is essential because the speech signal is continuous, and emotions vary. By breaking the signal into frames, features can be extracted from each segment, facilitating accurate analysis [36, 49]. The frame size significantly influences emotion recognition performance and is dataset-dependent, as SER datasets contain audio samples of varying lengths. Segmenting the speech signals with a consistent frame size is crucial to achieve a high recognition rate. Even a 1-millisecond difference in frame size can impact recognition success.

2.3.3.3 Windowing

As described in the framing process, the speech signal should not be processed as a whole but should be divided into regular interval sections, known as frames [44]. However, framing can lead to information loss at the transitions between frames. An overlap rate is applied to mitigate this, called windowing [49]. Each signal is initially framed and then windowed to smooth out any discontinuities at the beginning and end of each frame. Applying this window function helps to minimize information loss during framing [19].

2.3.3.4 Voice Activation Detection

Speech signals are classified into voiced, unvoiced, and silence segments, a process that is essential for speech analysis [48]. Voiced speech occurs when the vocal cords vibrate to produce sounds, typically vowels. Unvoiced speech is produced when only air is expelled from the lungs, creating turbulence without vocal cord vibration. Silence is defined by the absence of any vocal cord activity. Detecting these segments, known as endpoint or voice activity detection, is critical for system accuracy. Common techniques for voice activity detection include the auto-correlation method, zero crossing rate, and short-time energy analysis [43].

2.3.3.5 Normalization

Normalization typically involves transforming the speech signal to a normal volume [48]. First, the maximum value of the signal is calculated, followed by dividing the full sequence of the signal by the computed maximum value. This ensures that each sentence has the same volume. Normalization does not compromise on the speech features' discriminative strength but decreases the speaker and recording inconsistencies. Several normalization methods have been proposed, as highlighted in [50, 51]. Z-normalization is one of the most commonly used techniques [19, 36].

2.3.3.6 Noise Reduction

Background noise from the environment is often captured alongside the speech signal, especially in natural, real-time speech data recorded across diverse settings. This noise can negatively impact recognition accuracy, resulting in lower performance. Therefore, applying noise reduction techniques is essential [52]. Commonly used methods for reducing background noise include the Log-spectral amplitude MMSE (LogMMSE) and Minimum Mean Square Error (MMSE) estimators.

2.3.4 Speech Features

Following speech signal processing, the subsequent crucial step is feature extraction, which significantly impacts recognition performance [53]. Speech feature extraction aims to identify features that effectively highlight and distinguish different emotions [19]. The extracted features from the audio speech signal are called acoustic features. Emotion is contained in a speech signal which is varying in nature. The features selected and how they characterize each emotion enhance the emotion recognition rate. There is no guarantee which extracted feature is suitable for which classifier. The extracted features typically fall into one of these categories: frequency-domain features, time-domain features, statistical features, hybrid features, and deep features [34, 43].

Traditional hand-crafted features rely on acoustic characteristics extracted from each frame of a speech signal. These features can be derived from local and global characteristics [45]. Local features, also known as Low-Level Descriptors (LLDs), are extracted from individual frames or quasi-stationary segments. They capture essential information from specific parts of the signal, as emotional features are not uniformly distributed throughout the speech. Thus, these features are often called segmental or short-term features, encapsulating temporal dynamics. Global or long-term features, on the other hand, are derived from local features and are represented by statistical measures such as mean, standard deviation, maximum, and minimum values. These are also known as high-level features or high-level statistical functions (HSFs).

A separate category comprises deep features obtained by feeding raw signals or spectra into a deep neural network. Finally, a hybrid approach can combine all types of features to create a comprehensive feature set [43]. A detailed description of these speech features is provided below.

2.3.4.1 Prosodic Features

Prosodic or paralinguistic features are described as features that people can understand, such as rhythm and intonation. They are categorized as long-term features as their frame duration for speech analysis is large [45]. Prosodic features are primarily based on frequency, energy, and duration. They can help differentiate between high and low arousal (like happy and sad) but not between emotions of the same arousal (like happy and angry) [27].

2.3.4.2 Spectral Features

Spectral features are often referred to as vocal tract features. A sound made by a person is filtered by the vocal tract [36]. The shape of the vocal tract determines the resulting sound, and accurately simulating this shape leads to precise representations of both the vocal tract and the generated sound. These features are displayed in the frequency domain, with the Fourier transform enabling their extraction by converting the time-domain signal into frequency components [43]. The most commonly used spectral features for emotion detection, as described in [27], are outlined below.

1. Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are among the most commonly used spectral representations of speech in automatic speech recognition. They are known for their effectiveness in identifying the emotional state in spoken utterances [54]. MFCCs represent the short-time power spectrum and capture the shape of the vocal tract. To calculate MFCCs, the speech signal is divided into segments, each transformed into the frequency domain using a short-term discrete Fourier transform [45]. A Mel-frequency filter bank is then applied to map the linear frequency scale to the Mel scale, which is designed to reflect the way the human ear perceives sound frequencies. This Mel scale is logarithmic, making it more sensitive to lower frequencies than higher ones. In the cepstral analysis stage, the Mel spectrum is transformed back into the time domain using the Discrete Cosine Transform (DCT), resulting in the Mel-Frequency Cepstral Coefficients (MFCCs) [55].

2. Linear Prediction Cepstral Coefficients

Linear prediction cepstral coefficients (LPCCs) with vocal tract characteristics gather emotion-specific information [36]. LPCCs with a recursive method can be attained from the Linear Prediction Coefficient (LPC) [19, 56]. LPC models the vocal tract and analyses the spectral properties of speech. LPC calculates coefficients by predicting each speech signal sample as a linear combination of past samples. LPCCs are then obtained by transforming these LPC coefficients to the cepstral domain, which helps to capture the speech signal's important spectral features more effectively.

3. Gammatone Frequency Cepstral Coefficient

Gammatone Frequency Cepstral Coefficients (GFCCs) are auditory features derived from a bank of Gammatone filters, which are designed to closely model the frequency response of the human cochlea. Similar to the Mel-Frequency Cepstral Coefficients (MFCCs), GFCCs are computed by passing the speech signal through a series of filter banks—however, in this case, Gammatone filters are used instead of Mel filters to capture sub-band energies. These energies are then logarithmically scaled and processed using the Discrete Cosine Transform (DCT) to obtain the final cepstral coefficients. This approach preserves important perceptual cues and provides improved robustness to background noise, making GFCCs particularly useful in tasks like speech emotion recognition and speaker verification [19, 57].

2.3.4.3 Voice Quality Features

Voice quality features are called sub-segmental level features because the speech analysis segment duration is typically shorter than 10 ms [27]. Due to the compression in the vocal cords, when air is expelled through the glottis, an emotion is generated. Varying glottal aperture can generate different tones. For example, a harsh voice is linked with anger. These prosodic features are not primary but secondary to SER systems. The voice quality influences emotional content in speech. Changes in voice quality, such as harmonics-to-noise ratio, jitter, and shimmer, may cause some differentiating emotions in a small variation [58].

2.3.4.4 Teager Energy Operator Based Features (TEO)

Stress in speech can impact speech recognition accuracy, which, in turn, affects emotion recognition. Techniques are employed to reduce or offset stress in speech to enhance robustness

[59]. Teager and Teager [19] introduced a technique for detecting stress within speech using the Teager Energy Operator (TEO), a nonlinear process for measuring energy in speech signals. TEO-Auto-Env, introduced by Zhou et al. [59], is highly effective for stress classification and assessment.

2.3.4.5 Deep Learning-Based Features

In recent years, deep learning algorithms have surged in popularity due to their ability to automatically learn intricate patterns and representations from vast datasets. They extract acoustic features from raw speech in an end-to-end fashion. These algorithms can discern subtle variations and nuances in audio signals, producing more precise and robust emotion recognition models. The LLD features can be directly fed into deep learning models, enabling these models to learn more intricate representations based on these features [20]. They learn hierarchically, progressively extracting higher-level abstractions from lower-level features. Automatic feature extraction from raw signals using deep learning models has generally outperformed traditional handcrafted features in SER [27]. However, deep networks can also benefit from incorporating handcrafted features [60].

2.3.4.6 Non-linguistic Utterances

These non-verbal behaviours, such as laughter, breathing patterns, irregular sounds, crying, and pauses, often accompany verbal utterances and are valuable for emotion recognition. Known as speech disfluencies, these behaviours can help identify various emotions, including happiness and sadness. They can be detected using automated speech recognition technology [27].

Feature extraction can generate a wide range of possible speech features, but there is no universally ideal set for modelling emotions. To ensure optimal performance, it is important to identify and focus on the most relevant features. Without this step, the classifier may be overwhelmed by an excessive number of features, leading to the curse of dimensionality, which can result in longer training times and a higher risk of overfitting—both of which can negatively impact emotion recognition accuracy in speech [19].

The description of speech features was conducted to comprehensively understand their relevance and potential in improving the performance of a real-world SER system. As per the discussion, each feature type offers unique insights into speech characteristics. In conjunction with these, pre-processing techniques were also described to understand how signal quality can be enhanced, ensuring that the extracted features are clean and reliable for emotion detection. These features and feature engineering techniques help decide which ones would effectively enhance the efficiency and accuracy of the proposed SER system, aligning with the research objectives. This would further assist in developing a real-world SER system that can detect emotions accurately in practical scenarios. In the next section, the last component of SER, as illustrated in Figure 2.1, is discussed, which are the classifiers.

2.3.5 Classifiers

The next component in SER is the classification of emotions in speech. This section begins by discussing traditional machine learning classifiers, followed by deep learning classifiers. While deep learning techniques have shown exceptional performance across various applications [21, 28, 34, 41, 61], it is important to examine both traditional machine learning and deep learning approaches to gain a comprehensive understanding of their strengths and limitations for SER. Given the complexity of the task, there is no universally agreed-upon algorithm, making it crucial to explore both techniques to identify the most suitable solution.

Current research in SER relies heavily on empirical studies to determine effective approaches. SER systems use classification algorithms involving an input *X*, an output *Y*, and a function that maps them as $f(X) = Y$. This mapping function predicts the class of new inputs. A learning algorithm uses labelled data to categorize samples into classes [36]. Typically, data is divided into training, validation, and test sets: the training set is used to optimize the model's central parameters (connection weights, node biases, and such), the validation set to select the best model, and the test set to evaluate performance [62]. Common traditional machine learning algorithms for SER include Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Support Vector Machines (SVM), Decision Trees (DT), k-Nearest Neighbors (k-NN), k-means, and Naive Bayes Classifiers. Some of the traditional classifiers are presented first, as below.

2.3.5.1 Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a probabilistic approach that models the probability distribution of a set of data points. The data points are generated from a mixture of finite Gaussian densities [19]. While GMMs are applied in unsupervised and supervised learning, they are less commonly used as classifiers. However, GMMs are well-suited for frame-level analysis. According to Wan et al. [63], GMM variants (AIC-GMM, BIC-GMM, and VBGM) demonstrate improved performance compared to the original GMM classifier. AIC-GMM (Akaike Information Criterion - Gaussian Mixture Model) and BIC-GMM (Bayesian Information Criterion - Gaussian Mixture Model) are model selection criteria that help balance model fit and complexity, with BIC applying a stronger penalty for model complexity. VBGM (Variational Bayesian Gaussian Mixture Model) is a GMM variant that uses variational inference to estimate parameters efficiently, especially for large or sparse datasets. SC-GMM (Structured Covariance Gaussian Mixture Model) assumes a specific structure for the covariance matrices, reducing complexity and preventing overfitting. Results show that SC-GMM significantly outperforms the original GMM classifier and is comparable in accuracy to AIC-GMM, BIC-GMM, and VBGM but more efficient than AIC-GMM and BIC-GMM. Moreover, SC-GMM achieves competitive performance compared to KNN, SVM, decision tree, and naive Bayes classifiers [63].

2.3.5.2 Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) have been widely applied in automatic speech recognition (ASR) due to their ability to capture the time-dependent characteristics of speech. This makes them useful for identifying emotional content in speech [64]. HMMs were among the earliest methods used for SER, initially focusing on temporal features alone. However, classification results can be improved when both temporal and spectral features are included. In an SER system, HMMs are trained for each emotion label after feature maps are obtained. The system's effectiveness is evaluated by calculating the log-likelihood of emotions in test data using the trained models. The model's performance is fine-tuned by adjusting the Markov model's number of iterations and states [64]. Additionally, HMMs effectively utilize contextual information and require minimal data for training, making them well-suited for SER with natural datasets [27].

2.3.5.3 Support Vector Machine (SVM)

SVMs are based on the statistical learning theory of structural risk minimization, aiming to reduce the empirical risk on training data [65]. A key factor in SVM performance is the choice of kernel function, which maps the original feature space to a higher-dimensional space for effective classification. Their popularity stems from strong generalization capabilities, effective handling of outliers, and adaptability to high-dimensional features, which have been successfully applied in various real-world scenarios [65]. Additionally, SVMs are well-suited for small datasets [27].

2.3.5.4 K-Nearest Neighbors (KNN)

The KNN algorithm is a straightforward supervised learning technique for regression and classification tasks. It operates based on the distance or similarity between data points, often using the Euclidean distance as a metric. In classification, KNN assigns a new data point the label of the most common class among its k-nearest neighbors, with k representing the number of neighbors considered. If k equals 1, the label is based on the nearest neighbor's class alone. Research in [36] demonstrates that KNN, when combined with other classifiers, can produce strong SER results for multilingual databases, as it inherently captures the nonlinear relationships between emotional features [27].

2.3.5.5 Decision Trees

Decision trees are a supervised machine learning approach for regression and classification tasks. A decision tree is a nonlinear classifier where leaves represent classification outcomes, and branches denote the combinations of features that lead to these outcomes [32]. A decision tree progressively splits the input data into smaller subsets as it traverses the tree toward the leaves. This process continues until each subset reaches a desired level of homogeneity, at which point the algorithm stops [36].

2.3.5.6 Naive Bayes Classifier

The Naive Bayes Classifier is a supervised probabilistic algorithm based on Bayes' theorem, a core principle of probability theory that assumes conditional independence between features given a class. It encompasses a family of algorithms relying on this conditional independence assumption [66]. According to [67], Naive Bayes classifiers, when trained on certain datasets such as EmoDB, often yield superior results compared to many existing systems for SER.

2.3.5.7 Ensemble of Classifiers

An effectively constructed ensemble or combination of multiple classifiers generally outperforms a single classifier [68, 69, 70, 71, 72]. Combining classifiers enhances predictive performance, as each model captures different facets of emotional data. In ensemble learning, multiple machine learning algorithms are combined to enhance predictive performance. Each algorithm in the ensemble contributes to the final result, typically through a voting mechanism. The performance of ensemble models is often superior to that of individual classifiers. There are various ensemble classifier architectures. One approach involves feeding the same data to each classifier and comparing the results to make a final decision. Another approach uses a hierarchical classifier, where input data is first processed by one algorithm, and its output is then passed to another classifier sequentially and hierarchically to make the final decision [19]. Ensembles have been shown to work effectively for SER [69, 73].

In the next section, we describe deep learning techniques. We begin by defining deep learning and the activation functions related to deep learning techniques, followed by describing some commonly used deep learning techniques.

Deep Learning

Deep learning is a subset of machine learning that relies on deep neural networks (DNNs), consisting of multiple layers of neurons designed to automatically learn complex patterns from large amounts of data. Deep learning techniques offer several advantages over traditional machine learning methods. These techniques are especially effective for speech-related applications, as they can detect intricate structures and features without requiring manual feature extraction or adjustment. Deep learning models can extract low-level features from data and even work with unlabelled datasets. Deep learning solutions are essentially artificial neural networks with multiple representational layers (hence ‘deep’) and have proven effective across various applications, including SER. The deeper the architecture, the more complex and abstract the representations it can learn. Due to the vast number of parameters in DNNs, they often require substantial amounts of data [19, 27, 41, 74]. Figure 2.2 illustrates a simple block diagram representing an example of DNN architecture for an SER system [75]. As shown in Figure 2.2, the architecture of the DNN for SER begins with an input layer that allows for pre-processed speech features extracted from raw audio signals. Following the input layer, several hidden layers are stacked, where each layer contains a number of neurons consisting of different types of vector operations and nonlinear activation functions. These hidden layers allow the network to learn hierarchical representations of the input features, enabling it to capture intricate patterns related to emotional content in the speech. The architecture culminates in an output layer, often utilizing a SoftMax activation function to produce a probability distribution across multiple emotion classes. This allows the model to predict the likelihood of each emotion, facilitating effective classification based on the learned features. DNNs in speech emotion recognition leverage their depth and non-linear transformations to improve recognition performance.

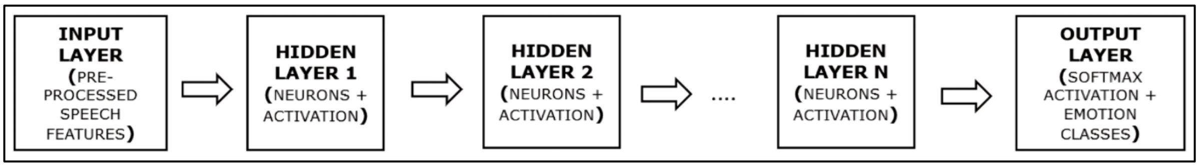

INPUT LAYER (PRE-PROCESSED SPEECH FEATURES)
HIDDEN LAYER 1 (NEURONS + ACTIVATION)
HIDDEN LAYER 2 (NEURONS + ACTIVATION)
....
HIDDEN LAYER N (NEURONS + ACTIVATION)
OUTPUT LAYER (SOFTMAX ACTIVATION + EMOTION CLASSES)

Figure 2.2. Block diagram of a DNN architecture for an SER system

Some activation functions commonly used in deep learning architectures are described in the next section, as they are essential for understanding how neural networks learn and process information [75]. These descriptions will aid in understanding deep learning models and their applications in the later sections of the study.

Activation Functions

The input layer of a deep learning model accepts data for training, while hidden layers process this data by detecting patterns and consolidating similar features. The output layer generates predictions or classifications, typically using a SoftMax function to provide probabilities. The placement of activation functions within the network depends on their role. Activation functions placed after the hidden layers transform linear mappings into non-linear ones, while the ones placed in the output layer handle predictions. Activation functions in neural networks are typically applied after computing the weighted sum of the inputs and biases, determining whether and how a neuron should be activated. The network typically learns by tuning weights and biases via a gradient descent procedure that shapes the overall function computed by the network according to the data. Activation functions are typically non-linear, and their purpose is to control the outputs of neural networks, improving performance across domains like object recognition and speech recognition [74]. Choosing the right activation function is critical for enhancing the accuracy of a neural network. In linear models, the input is mapped to the output through an affine transformation, expressed as:

$$f(x) = w^T x + b \tag{2.1}$$

x represents input, w is the weights, and b is the bias.

Neural networks produce linear results by default, and activation functions are introduced to transform these linear outputs into non-linear ones, which helps the network learn complex patterns in data. The formula for the output is:

$$Y = (w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b) \tag{2.2}$$

In multilayer networks like deep neural networks (DNNs), the outputs from each layer are passed to the next, and activation functions are used to introduce non-linearity, which is essential for learning higher-order relationships. Non-linear activation functions are applied to the outputs of linear models, as shown in the formula:

$$Y = \alpha\,(w_1 x_1 + w_2 x_2 + \ldots + w_n x_n + b) \tag{2.3}$$

where α represents the activation function. These functions convert linear inputs into non-linear outputs, enabling the network to learn complex, higher-degree relationships in the data. DNNs consist of multiple hidden layers, strategically placing activation functions between them to ensure proper learning and prediction. Some of the commonly used activation functions are described below [74]:

a) ReLU (Rectified Linear Unit)

This is a widely used activation function in deep learning due to its simplicity and effectiveness and is defined as:

$$f(x) = \max(0, x) \tag{2.4}$$

The function outputs the input directly if it's positive. Otherwise, it outputs zero, introducing non-linearity into the network. This allows ReLU to help neural networks learn complex patterns while maintaining computational efficiency, as it avoids the need for complex operations like exponentiation and addresses the vanishing gradient problem. Additionally, ReLU creates sparsity by setting negative inputs to zero, which can reduce computational load and improve representational quality. However, it faces the "dying ReLU" problem, where neurons can stop learning if they output zero consistently. Despite this, ReLU remains popular for its balance of simplicity and performance, and its variants, like Leaky ReLU, are used to address potential issues.

b) Sigmoid

The sigmoid is a commonly used activation function in neural networks, especially for binary classification tasks. It has an S-shaped curve and is defined as:

$$f(x) = \frac{1}{1+ e^{-x}} \tag{2.5}$$

The sigmoid function maps any real-valued input to a value between 0 and 1, making it useful for representing probabilities. When the input is large and positive, the output approaches 1, and when the input is large and negative, the output approaches 0. This non-linearity helps the neural network capture complex relationships in the data. However, the sigmoid function can suffer from issues like vanishing gradients, where the gradient becomes very small for inputs with relatively large magnitudes, slowing down the learning process in deep networks. Despite these drawbacks, the sigmoid function is still used, particularly in the output layer of binary classification models, to produce probabilities.

c) SoftMax

This is a commonly used function for the final layer in neural networks for multiclass classification tasks. It converts the raw output of a model into a probability distribution, where the values represent the likelihood of each class. The SoftMax function takes a vector of raw scores (logits) from the preceding layer and normalizes them to sum to 1. This makes the outputs interpretable as probabilities. Each element in the output vector corresponds to a specific class, and the class with the highest probability is typically selected as the predicted class. The SoftMax function is defined as:

$$\text{SoftMax}(z_i) = \frac{e^{zi}}{\sum_j e^{zj}} \tag{2.6}$$

where z_i is the input score for class i, and the denominator sums the exponential of all class scores z_j, ensuring that the output is a valid probability distribution. This makes the SoftMax layer particularly suitable for classification problems with multiple classes.

In the next section, we describe some of the DNNs commonly used for SER, which include Multilayer perceptron's**,** Recurrent Neural Networks, Deep Belief Networks, Deep Boltzmann Machines, Long-Short Term Memory, and such.

2.3.5.8 Multilayer Perceptrons (MLPs)

MLPs are a feedforward artificial neural network inspired by the structure of the brain, commonly used for classification and regression tasks. MLPs generally consist of input, hidden, and output layers, each containing multiple nodes. The nodes in the input and output layers vary based on the nature of the input data and labelled classes. In contrast, the number of hidden layers, nodes within them, and activation functions are typically tailored to the specific problem. MLPs have demonstrated strong performance in SER when combined with other classifiers [19]. Their fast prediction time suits them well-suited for real-world applications [27].

2.3.5.9 Deep Boltzmann Machine (DBM)

DBMs are named after the Boltzmann distribution attributed to Ludwig Boltzmann. It is an unsupervised probabilistic model with visible units and multiple hidden layers. The visible units connect to the data, while the hidden units assist in modeling the data distribution [76]. Through layer-by-layer pre-training, DBMs achieve rapid learning and efficient data representation, often yielding strong results in SER.

2.3.5.10 Restricted Boltzmann Machine (RBM)

The computational complexity of the Boltzmann Machine led to the development of the RBM, a streamlined variant where connections within the same layer are restricted. This means that neurons within the input layer or hidden layer cannot connect to each other, while connections are allowed between the hidden and visible layers [76]. As two-layer neural networks, RBMs can automatically detect underlying patterns in data by reconstructing the input [36].

2.3.5.11 Deep Belief Networks (DBN)

DBNs are highly beneficial for speech emotion recognition (SER), as they consist of stacked Restricted Boltzmann Machines (RBMs) capable of learning complex, high-level representations. The process begins by training data on the first RBM, with its output then serving as input to the next RBM, and so on. This sequential approach creates a deep hierarchical model that learns progressively from low-level to high-level features. The features extracted by DBNs can then be used as inputs to supervised algorithms for further processing [77].

2.3.5.12 Recurrent Neural Networks (RNN)

RNNs are robust neural networks renowned for their effectiveness, especially due to their internal memory, which allows them to retain input information and predict future sequences. This capability makes RNNs particularly useful for sequential data like speech [19]. However, one major limitation of RNNs is their struggle with the vanishing gradient problem and difficulty in representing relationships between signal elements that are far apart in the sequence. Long Short-Term Memory (LSTM) networks address these issues by incorporating a specialized memory cell within the recurrent connections, enhancing their performance in handling long-term dependencies [20].

2.3.5.13 Long Short-Term Memory (LSTM)

LSTM is an advanced type of RNN designed for sequential data, excelling at retaining and managing information over extended periods. Its popularity is growing across various applications, particularly in time series tasks such as speech emotion recognition and speech processing. LSTMs can identify long-term paralinguistic features and have proven effective in speaker-independent emotion recognition [43, 78]. LSTMs feature a memory unit that functions as an input mechanism, allowing data to be retained for a certain duration. What sets LSTMs apart is their ability to recall the most recently computed value from this memory unit, enabling them to address the short-term memory limitations commonly encountered in traditional RNNs.

2.3.5.14 Convolutional Neural Networks (CNN)

CNNs are deep learning models that utilize a feed-forward architecture, and are primarily employed for classification, although they can equally be applied to regression problems [79]. They are widely used in various computer vision applications, including image classification, face recognition, object detection, and more [20]. CNNs are structured as hierarchical neural networks with multiple sequential layers, typically featuring several convolutional layers. These layers extract local features by convolving the input with different localized filters learned during training. Following the convolutional layers are pooling layers, which typically aggregate the maximum activation features from the convolutional feature maps, effectively reducing their spatial resolution. CNNs may also include fully connected layers, where each neuron in the input layer connects to every neuron in the subsequent layer. This combination of convolutional, pooling, and fully connected layers forms a pipeline for feature extraction that effectively models the input data. Finally, the SoftMax layer performs the classification task [80].

Exploring traditional and deep learning classifiers was conducted to assess their effectiveness in speech-emotion recognition (SER). The next section presents a literature review on research in the SER field. The application of the speech features and classifiers discussed earlier will be examined across various studies. The goal is to identify the most suitable approach for a real-world SER system that ensures accurate and efficient emotion detection. This analysis will help determine which classifier best integrates with the selected features and pre-processing techniques, ensuring reliable and accurate emotion recognition in practical scenarios.

2.4 Related Work in SER

SER has been an active research field for the last three decades, with the motivation of finding innovative ways to extract emotional context from speech appropriately. Feature engineering steps involving feature selection, extraction, and classification have been described in Section 2.2. These steps are intended to contribute to the development of accurate and effective models. Although deep learning models automate certain steps in feature engineering, the importance of feature selection, and feature engineering remains valuable in traditional machine learning and deep learning methods.

SER requires careful selection of the speech emotion corpora (datasets), identifying various features inherited in speech, and a flexible model for classifying those features. The goal is to enhance the performance and accuracy of emotion recognition systems. Many speech features have been proposed, and classification algorithms have been developed for effective emotion recognition, but still, there is fertile ground for future research opportunities. Studies have been done to explore various acoustic features, either independently or in combination, to discern emotions effectively. Machine learning (ML) studies rely on feature selection techniques, while deep learning (DL) investigations focus on end-to-end feature learning. There is no definitive set of features guaranteeing precise emotion classification. Studies mostly rely on the outcomes from experiments conducted. The classifiers must also be explored for their potential utilization to provide reasonable accuracy. While discussing the various approaches in the studies below, the emphasis is on the factors that are helpful for SER to be applied to a real-world application. Since SER has emerged as a pivotal element in HCI, its applications have been widespread in real-world scenarios.

The literature review in the next section aims to analyse SER systems with a focus on emotional datasets, the suitability of selected features, the deployment of appropriate classifiers leveraging ML or DL technologies, and their applicability to real-world scenarios. To trace advancements in SER over the years, a relevant set of papers was selected from different periods, encompassing a variety of features and classifiers to identify those most suitable for developing our real-world SER system. A search query combining relevant keywords ("speech emotion recognition," "real-time emotion detection," "hybrid CNN models," and "accessibility in SER systems") was used to explore key databases such as Google Scholar, IEEE Xplore, and Scopus. Key articles were shortlisted based on their use of ML and DL classifiers. Inclusion criteria were applied to focus on peer-reviewed articles, conference papers, and other credible sources published within the last 10 years for machine learning-based SER systems and the last 6 years for deep learning-based approaches. The review focused on the datasets utilized, speech features used for emotion recognition, classifiers employed, and the accuracy of the trained SER models in recognizing emotions. Collectively, these factors are evaluated for their potential in real-world applications. The review also identified and highlighted limitations in the selected papers to address the first research objective.

First, a review of selected studies applying machine learning techniques for SER is presented below and summarized in Table 2.1 [refer to Appendix 1]. The review focuses on the key components that influence model performance and their applicability in real-world settings.

2.4.1 Datasets Utilized

Most of the selected works reviewed under machine learning approaches utilize acted datasets such as the EMO-DB [81, 82, 83, 84, 86], RAVDESS [56, 84, 85], SAVEE [83, 85, 87], and EMOVO [85]. While these datasets provide well-labelled emotional categories, they consist of exaggerated expressions that lack the natural variability found in real-life interactions. Some work utilised self-created databases [88].Only a few studies consider multilingual datasets—for instance, [89] examined emotion recognition in native Odia languages, [85] used an Urdu database, and [83] employed the Persian Drama Radio Emotional Corpus (PDREC) alongside EMO-DB and SAVEE. Some studies used mixed corpora for cross-validation purposes like

CASIA and EMODB [82], IEMOCAP and EMODB [86], RAVDESS, EMO-DB and IEMOCAP [84], EMO-DB, SAVEE and PDREC [83] and some considered semi-natural datasets like IEMOCAP and e-NTERFACE [84, 85, 86].

However, a clear limitation in these studies is the reliance on acted datasets alone. Cross-dataset evaluation is rarely performed. To improve adaptability for real-world SER systems, focus should be on collecting and utilizing semi-natural and natural datasets, as well as implementing cross-corpus evaluation.

2.4.2 Speech Features used for Emotion Recognition

Most SER systems rely heavily on spectral and prosodic features. In this reviewed work of the selected studies under machine learning approaches, MFCCs appearing most frequently [81, 89, 88, 56, 84, 85]. In addition to MFCCs, features such as pitch, energy, Zero-Crossing Rate (ZCR), Delta MFCC, LPCC, RMS and wavelet-based features have been used [56, 81, 85, 88, 89]. Novel techniques include the Mel Frequency Magnitude Coefficient (MFMC), which replaces the energy spectrum with the magnitude spectrum and omits the discrete cosine transform to enhance spectral representation [85]. This feature is used along with three conventional spectral features—MFCC, Log Frequency Power Coefficient (LFPC), and Linear Prediction Cepstral Coefficient (LPCC). Another novel contribution was the use of adaptive time-frequency features derived from the Fractional Fourier Transform to capture more nuanced speech dynamics [83]. Some studies demonstrated that combining deep and acoustic features using pre-trained CNNs (e.g., ResNet) followed by feature selection using the Relief algorithm can significantly enhance emotion recognition [84]. In this work [84], Acoustic features like Root Mean Square (RMS) energy, Mel-Frequency Cepstral Coefficients (MFCC), and Zero-Crossing Rate are first extracted from voice recordings. Spectrogram images of the original sound signals are then input into pre-trained deep network architectures, including VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201, to obtain deep features. Feature selection using metaheuristic approaches, such as Cuckoo Search and a modified NSGA-II algorithm, was shown to reduce dimensionality while maintaining high accuracy [86]. These feature selection algorithms successfully eliminated unnecessary features and compacted the input to a smaller, more effective subset of features.

Despite these advances, many studies still rely on handcrafted features or static feature combinations, missing opportunities to fuse complementary feature types such as traditional prosodic descriptors and deep embeddings, which can enhance the emotion recognition of the SER systems in real-world application. Feature fusion remains largely underutilized.

2.4.3 Classifier(s) Employed: Performance and Limitations

The reviewed work of the selected studies uses a range of machine learning classifiers for SER, including Gaussian Mixture Models (GMM), k-Nearest Neighbors (KNN), Hidden Markov Models (HMM), Support Vector Machines (SVM), Decision Trees, Random Forests, and Naïve Bayes classifiers [56, 81, 82, 83, 84, 85, 86, 87, 88, 89]. Traditional classifiers such as GMM and KNN were applied independently in [81] to identify six emotions from the EMO-DB dataset. In terms of precision and F-score, GMM outperformed KNN. GMM and KNN algorithms were applied independently on the EMO-DB dataset without exploring a fusion of these classifiers, which could potentially improve recognition rates. HMM and SVM were compared for emotion recognition in native Odia languages in [89], where SVM achieved 82.41% accuracy and HMM 78.81% for speaker-independent conditions. The SVM classifier with MFCC also outperformed HMM in the speaker-dependent system. Although the study shows overall good performance, its computational complexity remains a concern, and it lacked detection of a broader range of emotions. In [88], a Naïve Bayes classifier trained on MFCCs, pitch, and energy features yielded accuracies of 81% (angry), 78% (happy), 76% (sad), and 77% (neutral) on a 2000-utterance custom dataset. The classifier could perform recognition with minimal datasets, however, features like voice quality and prosody were not considered, which could have potentially enhanced the classifier's performance. A novel Random Forest technique combined with Decision Trees was used in [87], achieving an overall recognition rate of 78% on the SAVEE dataset, which includes six emotions. While this approach demonstrates good accuracy, it is time-consuming, highlighting a trade-off between accuracy and computational efficiency. Decision Tree, SVM, and LDA were compared in [56], using RAVDESS, with Decision Tree achieving the highest accuracy at 85%, followed by SVM (70%) and LDA (65%). During feature selection, a global feature algorithm was employed to eliminate redundant information, and machine learning classifiers are used to identify emotions from the selected features. The proposed feature extraction algorithms enhanced the speech emotion recognition rate. Comparing these results with existing work indicates that the proposed system offers greater accuracy, suggesting it effectively extracts the necessary signal characteristics to recognize emotions more accurately and efficiently. However, the model was not tested with other datasets, and an ensemble of classifiers was not considered. In [82], a fuzzy C-means clustering and multiple random forests were adopted to subclass the high-dimensional emotional features. Integrating these methods, known as the two-layer fuzzy multiple random forest (TLFMRF), improves emotion identification. The experiments utilize the CASIA corpus and Berlin EmoDB datasets. TLFMRF demonstrated a 1.39%–7.64% and 4.06%–4.30% higher recognition rate than backpropagation neural networks and random forests. Though the experiments were conducted on a mobile robot, it took into consideration only six basic emotions. SVM was the primary classifier in [83], achieving high accuracies of 97.57% (EMO-DB), 80% (SAVEE), and 91.46% (PDREC) using adaptive time–frequency features. Cepstral features have been combined with adaptive Time-Frequency features to extract novel features. Potential integration of deep features with the extracted time features and adopting deep learning classifiers for constructing a speaker-independent SER system would have been ideal. Similarly, SVM was used in [84] to classify acoustic and deep features. Six pre-trained popular CNNs - VGG16, ResNet18, ResNet50, ResNet101, SqueezeNet, and DenseNet201 were used to extract deep features from spectrogram images, reaching 90.21% (EMO-DB), 85.37% (IEMOCAP), and 79.41% (RAVDESS). For the final classification only machine learning classifier was considered. In [86], SVM achieved 87.66% (EMO-DB) and 69.30% (IEMOCAP) when paired with modified feature selection techniques. The feature selection algorithms used have successfully eliminated unnecessary features and compacted the input to a smaller, more effective subset of features. However, better feature selection algorithms must be considered with applicability to more datasets. Lastly, multiclass SVM was used in [85] to classify data from six different datasets. A modified spectral feature, the Mel Frequency Magnitude Coefficient (MFMC), was proposed by replacing the energy spectrum with the magnitude spectrum and removing the DCT step from MFCC extraction. Evaluation with MFMC alone achieved accuracies of 81.50% (Berlin), 64.31% (RAVDESS), 75.63% (SAVEE), 73.30% (EMOVO), 56.41% (eNTERFACE), and 95.25% (Urdu), outperforming traditional spectral features across most datasets. The performance of MFMC could have been further improved.

While these classifiers have demonstrated varying degrees of success, most studies apply them independently, without exploring hybrid or ensemble strategies that could potentially enhance performance. Furthermore, end-to-end deep learning pipelines remain underexplored, with many of these models still relying on traditional classifiers for final decision-making. For real-time SER applications, integrating classifier fusion with advanced feature extraction or fusion techniques should be considered.

The next section reviews selected studies applying deep learning techniques for SER as presented below and summarized in Table 2.2 [refer to Appendix 2]. The review focuses on the key components that influence model performance and their applicability in real-world settings.

2.4.4 Datasets Utilised

The choice of dataset plays a pivotal role in shaping the performance and generalizability of SER models. Deep learning models, in particular, require diverse and sufficiently large datasets to extract emotionally relevant speech patterns. The reviewed works of the selected studies reveal extensive usage of four primary categories of datasets: acted, semi-natural, natural, and multilingual/culturally diverse corpora. A few studies also introduce multimodal and synthetic/augmented datasets to support model robustness. Acted datasets form the foundation of many SER studies due to their clarity, structured emotions, and ease of annotation. These datasets typically feature professional actors delivering scripted utterances corresponding to predefined emotional categories under studio conditions as discussed in Section 2.3.2.1. The Emo-DB (Berlin Emotional Speech Database) is one of the most frequently used datasets across reviewed works [35, 73, 82, 90, 91, 93, 95, 97, 98, 102, 103, 105]. Deep learning models evaluated on dataset Emo-DB have shown achieving high accuracy, including up to more than 94% with CNNs [90, 103], about 90% with CNN [93] and 95.42% with CNN-GRU-LSTM ensembles[73]. The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is widely employed in studies using spectrogram-based CNNs, hybrid networks, and attention-based architectures [35, 55, 73, 90, 91, 92, 94, 96, 98], due to its clean recordings, and diversity in speaker gender and emotion types. It is shown achieving more than 90% accuracy with CNN [73, 90, 94]. The SAVEE (Surrey Audio-Visual Expressed Emotion) dataset also features prominently reviewed in studies [35, 71, 73, 90, 94, 97, 98, 100, 101]. Other acted datasets include TESS [73, 106], CASIA [82, 99, 101]. TESS, while clear and high quality, includes scripted utterances performed by professional actors and therefore falls under the acted category. These datasets tend to deliver high classification performance due to their structured nature and low noise, but their exaggerated expressions may limit generalizability. Models trained solely on such datasets often fail to perform reliably in spontaneous or real-world environments. Semi-natural datasets aim to strike a balance between the clarity of acted emotions and the authenticity of spontaneous interactions as discussed in Section 2.3.2.2. They typically involve scripted dialogues or scenario-based improvisations in studio-like environments. Among these, the IEMOCAP dataset is the most widely used in deep learning

SER research [68, 71, 91, 92, 93, 96, 97, 101, 102, 103, 104, 105, 107, 108, 109, 110, 111, 112]. Because of its size, multimodality, and rich annotation (categorical and dimensional), IEMOCAP supports various architectures, including BiLSTM, CNN-LSTM, and transformer-based models. Performance on IEMOCAP varies with feature fusion and model type, ranging from 74.23% [102] to 79.8% [103]. Other semi-natural datasets include (Crowd-sourced Emotional Multimodal Actors Dataset) CREMA-D [73], MSP-IMPROV [110], and eNTERFACE'05 [97, 102], RML (Ryerson Multimedia Lab) [94]. These corpora allow researchers to evaluate model performance in conversational contexts with moderate emotional variation. As such, they tend to yield more reliable results for real-world SER applications than purely acted corpora, while avoiding the complexity of spontaneous data collection. Spontaneous datasets capture unplanned, naturally occurring emotional speech in diverse acoustic conditions, offering the highest realism but posing significant challenges for SER models. These datasets typically include background noise, emotion ambiguity, and speaker variability, making them ideal for robustness testing. However, they are limited in availability and pose potential copyright and privacy concerns as discussed in Section 2.3.2.3. Prominent spontaneous datasets include AFEW 5.0 [70, 72], BAUM-1s [70, 71, 72]. These datasets are often used in studies employing memory-augmented networks, attention-enhanced transformers, and multi-stream CNNs. While performance on spontaneous datasets is typically lower than on acted or semi-natural corpora, their use is critical for building robust and deployable SER systems. Another natural dataset FAU-Aibo Emotion corpus (FAU-AEC) [112]. which is made up of spontaneous and emotional German speech samples. It gave lower accuracy rate as well. Recognizing the need for cross-linguistic generalization, some studies incorporated datasets in multiple languages. These include SUBESCO, a Bengali emotional speech corpus [114], Emirati Emotional Speech Dataset-SUSAS [115],CASIA [99, 101], and CHEAVD, the Chinese emotional speech-visual datasets used for multimodal SER. These datasets help evaluate how emotion perception and expression vary across linguistic and cultural groups. For instance, [114] used CNNs on SUBESCO to demonstrate the model's capacity to recognize Bengali emotions. Most SER research is still centred around English or Western emotional corpora because of scarcity of availability of large, annotated emotional speech datasets in non-Western languages. To expand beyond acoustic features, some studies used multimodal datasets combining audio, visual, and physiological signals. These include DEAP (speech signals and physiological) [113], and the audiovisual/motion-capture modalities of

IEMOCAP (audio-visual and motion capture modalities) [68, 105, 107, 108, 109], CHEAVD (Chinese Audiovisual Dataset) [101]. Such datasets enable fusion-based models to integrate affective cues from multiple sources—such as speech, facial expression, and EEG—thereby enriching emotion representation. For example, [113] fused EDA, zEMG, and audio using graph neural networks and achieved better classification performance than unimodal baselines. Across the reviewed studies, acted datasets remain the most commonly used, likely due to their clean signals and easy availability. However, this results in inflated accuracies that may not reflect model performance in real-world conditions. Spontaneous datasets, while valid for real time scenarios, often result in lower accuracy due to speaker variability, emotional ambiguity, and background noise [70, 71, 72]. Semi-natural datasets emerged as practical middle grounds. They allow for emotional authenticity and contextual richness yet maintain a level of control suitable for reproducible evaluation [68, 71, 91, 92, 93, 96, 97, 101, 102, 103, 104, 105, 107, 108, 109, 110, 111, 112].

To improve the utility and generalizability of deep learning-based SER systems in online teaching, there should be more use of semi-natural and spontaneous speech datasets. English-language datasets remain essential for developing core SER models in global educational settings whereas multilingual datasets incorporating regional languages and dialects should support inclusive learning environments. Additionally, incorporating cross-corpus evaluation and combining multimodal and synthetic data generation techniques will help build reliable and scalable models.

2.4.5 Speech Features Used for Emotion Recognition

Effective feature extraction and selection are central to SER. Deep learning approaches aim to model emotional cues from complex and varied speech signals, either directly from raw audio or from extracted speech features. The reviewed works for the selected studied adopt a wide spectrum of feature engineering strategies—ranging from traditional handcrafted features to advanced learned embeddings. The studies show a wide range of low-level descriptors (LLDs), high-level features, deep representations, and handcrafted-prosodic features, often in combination, to boost accuracy. Feature types can be broadly categorised into traditional spectral-prosodic features, time-frequency representations, learned deep features, multimodal and fused feature sets, and end-to-end or attention-enhanced embeddings.

Despite the growing dominance of deep learning, many studies continue to use or supplement models with handcrafted low-level descriptors (LLDs) due to their interpretability and alignment with known speech production features. The most commonly used handcrafted feature are the MFCCs, which appears across a wide range of studies including [55, 90, 94, 98, 103, 104, 108, 111, 113, 114]. MFCCs were widely adopted either alone [55, 98, 103, 113, 114], or in combination with other acoustic features like Chroma, Mel-spectrogram and such [90, 94, 104, 108, 111, 114]. When MFCCs were used alone with mixed dataset types, it gave accuracy of 89.53%, 80.81%, 79.87%, 94.21%,83.13%, and 89.93% on DEAP, RAVDESS, IEMOCAP, EMODB, SAVEE, and RAVDESS respectively. Whereas a Bi-LSTM model using MFCCs with Chroma and Mel-spectrogram achieved 83.33% accuracy on a Bengali corpus [114]. Studies that employed MFCC with Chroma, Zero-Crossing Rate (ZCR), Root Mean Square (RMS) and spectrogram gave average accuracy of 99.46% for TESS, 95.42% for EMO-DB, 95.62% for RAVDESS, 93.22% for SAVEE, and 90.47% for CREMA-D datasets [73]. MFCCs with CNN achieved accuracies of 96.45%, 83.13%, and 89.93% on the EMO-DB, SAVEE, and RAVDESS datasets, respectively [98]. MFCCs and time-domain features processed through a 1D CNN yielded 96.6% (EMO-DB), 92.6% (SAVEE), and 91.4% (RAVDESS) [90]. Other models used MFCCs in lightweight CNNs [103] or with hybrid GMM-DNN classifiers [114], yielding competitive performance even under noisy conditions. MFCC with CNN gave accuracy of 79.87% for IEMOCAP and 94.21% for EMODB [103] and MFCC with sequential GMM-DNN classier performed accuracy shown to be 83.97% on SUSAS dataset. MFCCs are usually computed on short-time windows and statistically aggregated before being input into CNNs, LSTMs, or fully connected neural networks. While easy to compute and interpret, these features may not fully capture the complex temporal or emotional patterns found in real-world speech, especially when used in isolation.

Several studies employed time–frequency representations like Mel-spectrograms, log-Mel spectrograms, and cochleagram [71, 72, 91, 94, 100, 102, 106, 108, 111, 112]. These features allowed CNNs and LSTMs to capture both spectral and temporal variations in emotional speech. Dual-level and augmented spectrogram approaches were explored in [108] and [71]. Deep segment-level spectrogram features in [72] were combined with LSTM outputs but yielded lower accuracy (40.73%AFEW5.0 and 50.22% on BAUM-1), suggesting that such natural data though good for real-world use comes with lot of noise. Spectrograms in [106] were fed to a CNN-LSTM and implemented on FPGA, achieving 97.86% on TESS. [102] used spectrograms with a DCNN to reach 95.14% (EMO-DB), 74.23% (IEMOCAP), and 89.46% (eNTERFACE05). In [112], 3D spectrograms were processed with 2D CNNs, achieving 73.1%

WA and 66.3% UA on IEMOCAP. While highly informative, these features when used alone demand extensive preprocessing and are computationally intensive, posing challenges for real-time SER.

With the rise of deep learning, there has been a shift toward automatically learning features from raw data. Some studies bypassed handcrafting by learning features directly from raw audio [70, 71, 92, 93, 105, 110]. Multi-resolution filters and dilated CNNs were used achieving 60.23% [110] and 73% [93] on IEMOCAP. In [92], ConvLSTM layers captured spatiotemporal cues, reaching 80% accuracy on RAVDESS. CNN models trained on raw speech and augmented spectrograms in [71] obtained up to 97.19% on SAVEE and 94.09% on IEMOCAP and 53.98% on BAUM-1s with 1D CNN and 96.85%, 88.80%, and 48.67% on SAVEE, IEMOCAP, and the BAUM-1s respectively on 2D CNN. Raw audio processing removes the need for feature engineering and can capture nuanced, speaker-specific cues. However, performance can suffer when training data is limited or acoustically inconsistent.

To enhance representation learning, many studies employ feature fusion strategies. Intra-modal fusion involves combining different types of speech-derived features. For example, [35] proposed a two-stream CNN architecture that processes spectrum, and spectrogram features independently before fusing them through concatenation. Similarly, [104] apply multi-head attention mechanisms to fuse outputs from parallel feature branches, improving emotional context capture and fusion of acoustic features [73] . Cross-modal fusion was also explored in datasets that include physiological or visual channels. For instance, [113] fused physiological features like zEMG and PPG with audio features using Deep Belief Networks, while More complex models in [70] and [72] fused 1D, 2D, and 3D CNNs to model hierarchical features, though accuracy varied depending on dataset quality from real-world datasets like BAUM-1s and AFEW5.0. [101] fused handcrafted and DNN features via local attention and RNNs, reporting 72.3% (IEMOCAP), 81.8% (SAVEE), and 63.3% (CASIA). Although fusion generally enhances performance, especially in noisy or ambiguous contexts, it also adds model complexity. [68] fused DNN, CNN, and RNN outputs derived from Frame-level low-level descriptors (LLDs), segment level Mel-spectrograms (MS), and utterance-level outputs of high-level statistical functions (HSFs) features, achieving weighted accuracy 57.1% and unweighted accuracy 58.3% for semi-natural dataset – IEMOCAP. To overcome data scarcity, some models leveraged features from pre-trained models like wav2vec 2.0 [96, 104]. For instance, [96] combined wav2vec features with a DNN and achieved 66.3% on IEMOCAP and 77.5% on

RAVDESS. In [104], MFCCs, spectrograms, and wav2vec embeddings were merged using a co-attention mechanism, achieving 71.05% accuracy. These models offer a strong baseline even with limited training data, especially in multilingual or low-resource scenarios, however they must be properly fine-tuned and evaluated across multiple datasets. Although feature selection and dimensionality reduction are less commonly explored in deep learning-based SER, a few notable exceptions exist. An INCA algorithm is employed for optimal feature selection, eliminating redundancy and discrepancies from the fused features [35]. This approach achieved 95%, 82%, and 85% recognition rates on the EMO-DB, SAVEE, and RAVDESS datasets. Attention-based feature weighting, as used in [73], allows models to focus on the most emotionally informative time frames. A cascaded spatiotemporal-frequential attention network in [97] achieved up to 83.3% (EMO-DB) and 80.47% (IEMOCAP). In another case, [107] applied hierarchical multi-task learning to progressively refine coarse features into fine-grained emotion-aware representations. Some works [98, 108] also suggest the potential of feature pruning and regularization techniques such as dropout or weight decay to improve generalization. However, explicit use of classical dimensionality reduction methods like Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA) remains rare, even though they could reduce computational load and improve interpretability. Unique strategies were introduced in [95, 99, 109]. [99] used deep bottleneck features with SVM and hybrid classifiers, achieving 75.83% but struggled to differentiate emotions like fear and sadness. [95] applied few-shot learning with Siamese Networks and log-Mel spectrograms on EmoDB, effective for some emotions like anger but lacking language independence. [109] explored stacked transformer layers in an end-to-end setting, improving performance by 20% on IEMOCAP. These novel systems push the envelope of SER architecture but often lack cross-corpus validation.

While various feature extraction strategies have shown high performance on benchmark datasets, several challenges remain. Handcrafted features are limited in expressiveness, deep-learned features need lot of data, and fusion models raise complexity. Multi-scale and multi-stream architectures that integrate both handcrafted and deep embeddings across temporal and spectral domains must be explored further. Feature interpretability should be prioritized using techniques such as saliency mapping or attention visualization. Intelligent, automated feature selection techniques such as evolutionary algorithms or differentiable neural architecture search can optimize representation pipelines.

2.4.6 Classifier(s) Employed: Performance and Limitations

In speech emotion recognition (SER), the classification component is responsible for accurately mapping extracted features to discrete emotional states. As speech signals are inherently complex, capturing both spectral and temporal dynamics, the choice of classifier significantly affects system performance, robustness, and generalizability. The reviewed studies under deep learning approach adopted a broad spectrum of classifier types—including shallow neural networks, convolutional and recurrent models, attention-based frameworks, and sophisticated hybrids.

Feedforward neural networks and statistical classifiers are commonly used as baselines in SER. In [113], the authors combined a Deep Belief Network (DBN) with a fuzzy genetic SVM (FGSVM), achieving 89.53% accuracy on the DEAP dataset. Another study employed SVM, DNN-SVM, and DNN–decision tree hybrid classifiers and achieved 75.83% accuracy on the Chinese Academy Sciences Emotional corpus, although it struggled to differentiate closely related emotions such as fear and sadness [99]. These architectures are generally lightweight and straightforward to implement but are limited in their ability to model the sequential nature of speech without additional temporal modeling.

Convolutional Neural Networks (CNNs) have become dominant in SER, particularly for learning localized features from time-frequency representations. CNNs trained on MFCC and time-domain features in [90] reached accuracy of 96.6% on EMO-DB, 92.6% on SAVEE, and 91.4% on RAVDESS. A deep multi-branch with improved CNN model (SCAR-NET) is introduced in [98] extract spectral, temporal, and spectral–temporal correlation features, achieving up to 96.45% on EMO-DB, 83.13% on SAVEE and 89.93% on RAVDESS. CNNs were also successfully applied with spectrogram inputs in [102], which yielded an accuracy of 95.14%, 74.23%, and 89.46% for datasets Emo-DB, IEMOCAP, and eNTERFACE05 respectively. A CNN model in [103], trained with MFCC features, achieved 94.21% on EMO-DB and 79.87% on IEMOCAP. In [112], a combination of techniques was introduced to identify subtle discriminative and relevant features in speech. The proposed approach uses an architecture called Deep Stride Convolutional Neural Networks (DSCNNs), which is a variant of CNN designed to boost computational speed by utilizing fewer convolutional layers while retaining accuracy. The DSCNN model achieved a prediction accuracy of 87.8%, outperforming the traditional CNN, which reached 79.4% on SAVEE dataset. A one-dimensional dilated convolutional neural network (DCNN) was presented to deal with the gap pertaining to real-time speech processing in SER systems [93] . The model uses a multi learning strategy to extract spatially salient emotional features in parallel and get long-term contextual dependencies from speech signals. A residual block with a skip connection (RBSC) module was used to get the long-term contextual dependencies in the input features. The model was evaluated on the IEMOCAP and EMO-DB datasets and obtained high recognition rates of 73% and 90%, respectively [112]. Finally, a fusion layer, a Parallel Convolutional Network with Squeeze-and-Excitation (PCNSE) was used, which integrates parallel convolutional layers (PCN) with a Squeeze-and-Excitation Network (Senet). It was proposed that relationships from 3D spectrograms across time steps and frequencies be captured. Log-Mel spectrograms with deltas and delta–deltas were used as input. Furthermore, a self-attention Residual Dilated Network (SADRN) with CTC as a classification block for SER was utilized. The effectiveness of the approach has been evaluated with the IEMOCAP and FAU-AEC datasets which did not give very high accuracy. However, these models are often limited in handling long-term temporal relationships. These results affirm CNNs' ability to handle static input representations effectively. However, standard CNNs are often inadequate for capturing long-term dependencies in emotional speech, which unfolds over time.

To overcome the limitations in temporal modeling, several studies proposed hybrid architectures that combine 1D, 2D, and 3D CNN layers, enabling comprehensive modeling of speech features across different dimensions. In [105], 1D and 2D CNN-LSTM combinations were evaluated on both EmoDB and IEMOCAP, with the 2D CNN-LSTM reaching 95.89% and 89.16% respectively—substantially outperforming earlier DBN and CNN baselines. Similarly, in [71], a pair of models using 1D CNN (Model A) and 2D CNN (Model B), each followed by LSTM layers, were tested on raw and augmented Mel spectrograms, achieving accuracies of 97.19%, 94.09%, and 53.98% on the SAVEE, IEMOCAP, and BAUM-1s datasets respectively. Another study proposed a 3D CNN guided by attention and connected to LSTM layers [94], resulting in accuracy improvements for the SAVEE, RAVDESS, and RML datasets. In [70], the authors fused representations learned independently by 1D, 2D, and 3D CNNs, improving classification on challenging datasets such as AFEW5.0 (35.77%) and BAUM-1 (44.06%).These results demonstrate that hybrid CNN architectures are particularly powerful in handling complex emotional dynamics, although they require extensive computation and memory resources. To address this, hybrid architectures have emerged that combine convolutional models across different dimensions.

Recurrent neural networks (RNNs) and their variants—including Long Short-Term Memory (LSTM), Bidirectional LSTM (Bi-LSTM), and Gated Recurrent Units (GRU)—have also been widely adopted to capture temporal structures in speech. A conventional LSTM model in [55] achieved 80.81% on RAVDESS when trained on MFCC features, while a Bi-LSTM using a richer feature set (MFCC, Chroma, Mel spectrogram) attained 83.33% on SUBESCO, a Bengali speech with seven distinct emotional states [114]. In [108], a dual-level model is presented for emotion prediction by combining MFCC features, and Mel-spectrograms. The MFCC features were processed using a standard LSTM, while the two Mel-spectrograms were processed concurrently through a novel Dual-Sequence LSTM (DS-LSTM). Using the IEMOCAP dataset, the model achieved an average weighted accuracy of 72.7% and an unweighted accuracy of 73.3%. This work introduces DS-LSTM, an architecture capable of processing dual Mel-spectrograms simultaneously, yielding substantial improvements over baseline and state-of-the-art unimodal models and reaching performance levels comparable to multimodal models. This demonstrates that unimodal models, relying solely on audio signals, still hold untapped potential. In [91], discriminative and salient spectrogram features were used with CNN and Bi-LSTM to attain 77% accuracy on RAVDESS and 72.25% on IEMOCAP. The use of ConvLSTM layers in [92] enabled modeling of both spatial and temporal interactions, yielding 80% accuracy on RAVDESS and 75% accuracy on IEMOCAP. A dual-stream LSTM architecture that ingested two Mel-spectrograms simultaneously improved the state of the art on IEMOCAP, with unweighted accuracy reaching 73.3% [108]. These models demonstrate that LSTM-based classifiers are especially effective in temporal modeling, though they can be slow to train and risk overfitting with limited data.

Attention mechanisms have been increasingly integrated into SER systems to selectively emphasize emotionally relevant speech segments. In [97], a novel approach was proposed: a Spatiotemporal and Frequential Cascaded Attention Network coupled with Large-Margin Learning. The Spatiotemporal attention mechanism discerns emotional regions within extended speech spectrograms, while the Frequential attention component captures emotional features based on frequency distribution within these regions. The Cascaded attention mechanism aids the neural network in progressively extracting effective emotion features from the extended spectrogram. Furthermore, Large-Margin Learning is employed during training to enhance intra-class compactness and amplify inter-class distances. Accuracies of 80.47%, 83.30%, 75.80%, and 56.50% were shown on IEMOCAP, EMO-DB, ENTERFACE05, and SAVEE datasets. A novel dual-level architecture named dual attention-based bidirectional long short-term memory networks (dual attention-BLSTM) for speech emotion recognition was proposed in [111]. It improved the unweighted accuracy on IEMOCAP to 70.29%, a 2.89% increase over the baseline. A co-attention mechanism that fused MFCC, spectrogram, and wav2vec features achieved 71.05% accuracy on IEMOCAP dataset [104], highlighting the benefit of integrating multiple modalities. Meanwhile, a transformer-based classifier using stacked transformer layers (STLs) in [109] achieved a 20% improvement over conventional CNNs on IEMOCAP. Despite their benefits, attention-based models are often more computationally expensive and prone to overfitting on smaller datasets.

Ensemble and fusion models also played a significant role in recent SER advances. In [73] three hybrid architectures—CNN-FCN, LSTM-FCN, and GRU-FCN—were developed and tested across five datasets, yielding up to 99.46%, 95.42%, 95.62%, 93.22%, and 90.47% for the TESS, EMO-DB, RAVDESS, SAVEE, and CREMA-D datasets. Such a work could reduce training time for individual models for the ensemble prediction. Another ensemble approach combining CNN, DNN, and RNN classifiers in [68] introduces a confidence-based fusion method comprising three multi-task learning-based sub-classifiers: a DNN with utterance-level high-level statistical features (HSFs), a CNN with multiple segment-level mid-level statistics (MS), and an RNN with multiple frame-level LLDs. The proposed fusion method effectively combined the recognition strengths of the three sub-classifiers, achieving a weighted accuracy (WA) of 57.1% and an unweighted accuracy (UA) of 58.3% for IEMOCAP, which was higher than each individual classifier. The work in [35] introduces an accurate and automated SER system based on a two-stream CNN architecture with optimal feature selection. The system processes the spectrum and spectrogram of speech signals, extracting high-level discriminative features through 2D and 1D-CNN architectures. An INCA algorithm is employed for optimal feature selection, eliminating redundancy and discrepancies from the fused features. This approach achieved 95%, 82%, and 85% recognition rates on the EMO-DB, SAVEE, and RAVDESS datasets. In [72], a multiscale deep convolutional long short-term memory framework for spontaneous speech emotion recognition was proposed. Deep segment-level features based on spectrograms were learned using a deep convolutional neural network. Later, a deep LSTM model was used to learn the segment-level CNN features. Finally, the CNN with the LSTM at multiple lengths of segment-level spectrograms were combined to recognize emotions. This was conducted on AFEW5.0 and 50.22% on BAUM-1s datasets. An accuracy of

40.73% was achieved on AFEW5.0 dataset. Work done in [101] proposes a multi-feature fusion and multilingual speech emotion recognition algorithm based on an RNN with an enhanced local attention mechanism. This proposed work extracted handcrafted and deep automatic features from existing Chinese and English speech emotion data. The diverse features extracted individually were fused and then integrated with the fused features across different languages. The proposed model gave an accuracy of 72.3%, 81.8%, 55.7%, and 63.3% on the IEMOCAP, SAVEE, CHEAVD, and CASIA datasets, respectively. In [110], raw speech fed directly into a CNN-LSTM model only achieved 60.23% on IEMOCAP and MSP-IMPROV, suggesting raw feature input requires stronger context modeling. These ensemble systems typically improve classification accuracy by leveraging diverse model strengths, though they introduce architectural and optimization complexity.

Finally, some studies proposed more novel or hardware-aware classifiers. A hybrid GMM–DNN system in [114] outperformed conventional classifiers in noisy environments, achieving 83.97% accuracy for SUSAS dataset. Siamese Neural Networks in [95] were explored for few-shot emotion classification, showing strong class-level accuracy for emotions like anger, sadness and neutrality for EMODB dataset. In [106], a real-time speech emotion recognition system was designed to classify emotions in real-time, enhancing the human-computer interaction experience. The system employs a CNN-LSTM neural network model, which was deployed on a Zynq FPGA board from Xilinx after conducting simulation experiments using the TESS speech emotion dataset. By reading speech data features from an SD card, the system successfully classified emotions and achieved a recognition accuracy of 97.86%, demonstrating its feasibility. A hierarchical CNN–RNN attention model using 3D log-Mel spectrograms in [107] obtained an F1 score of 0.4673, showcasing the potential for multi-scale emotional modeling. In the initial phase, 3D data was used as input to train on broader emotion categories, and the outputs were then utilized in a second phase focused on more specific emotions. This approach, tested on the IEMOCAP corpus, demonstrated a notable improvement over the standard models. Furthermore, pre-trained wav2vec 2.0 embeddings used in [96] showed promising results with a DNN classifier, yielding 66.3% on IEMOCAP and 77.5% on RAVDESS.

A wide range of classifiers have been explored in the above review, including feedforward neural networks, statistical models, CNNs, RNNs (LSTM, Bi-LSTM, GRU), attention-based frameworks, and advanced hybrid architectures. Shallow classifiers and single deep models such as DBNs, SVMs, and standard CNNs or LSTMs offer simplicity and reasonable accuracy but often struggle with modeling the complex temporal and spectral dynamics of emotional speech, particularly in real-world settings. CNNs excel at learning localized spectral features but are limited in capturing long-term dependencies, while RNNs effectively model temporal sequences but are prone to overfitting and slow training. Attention-based models enhance emotional relevance extraction but require large datasets and high computation. Hybrid architectures that combine CNNs with LSTMs or integrate 1D, 2D, and 3D CNNs demonstrate superior performance by capturing multi-scale, spatiotemporal patterns, though they demand greater computational resources. Ensemble models further improve robustness by fusing outputs from multiple classifiers, achieving higher accuracy and generalizability. Limitations across studies include reliance on small or acted datasets, lack of cross-corpus testing, and minimal consideration of semi natural or spontaneous speech. Focus should be on hybrid or ensemble approaches that are lightweight and interpretable, to support deployment in diverse, real-time contexts. In online teaching environments, such models can enhance accessibility by providing accurate emotion feedback to late-deafened educators, including those without hearing impairments, thereby supporting responsive and inclusive instruction.

The above section reviews the literature on SER for emotion detection. Key articles were selected based on their use of machine learning and deep learning classifiers over time. The review focused on the datasets used, features extracted, classifiers chosen, and the accuracy of trained SER models in emotion recognition. Together, these factors are essential for their potential for a real-world application. Recent literature highlighted that deep learning, a subset of machine learning, is increasingly preferred due to its multi-layered structure and efficiency, as indicated in works like [35, 41, 42, 55, 68, 73, 90, 91, 92, 93, 99,100, 107, 108, 109, 110, 113, 116, 117]. Although these studies have demonstrated high accuracy in emotion recognition, limitations exist. Many models were evaluated on limited datasets without extensive testing on diverse databases [35, 55, 68, 91, 92, 95, 99, 100, 104, 107, 108, 109,111,113]. Testing models across varied datasets enhances reliability and robustness [73, 103,113, 118]. Also, handcrafted LLDs alone fall short of accurate SER, necessitating deep learning to automatically extract high-level features, or "deep features" from data. These deep features capture complex patterns, enabling more effective emotion detection [40, 113, 119, 120, 121]. Integrating acoustic and deep learning features forms a hybrid vector that leverages traditional audio characteristics and learned representations, enhancing classification accuracy [35, 68, 70, 71, 84, 85, 94, 101, 104]. Additionally, combining classifiers by pairing deep learning classifiers or fusing traditional and deep learning methods often improves emotion detection [73, 91, 110]. However, some studies, such as [55, 81, 82, 83, 84, 88, 89, 97, 103], did not employ classifier ensembles, leading to lower accuracy rates. Moreover, certain methods with improved accuracy are either time-consuming [87] or computationally complex [55, 91]. Thus, the discussion highlights a need for a robust algorithm that enhances SER performance and accuracy. A hybrid, multi-branch, or ensemble classifier approach [32, 35, 70, 71, 72, 73, 90, 91, 99, 106, 110, 118] could represent a step forward in emotion recognition systems, especially if tested on diverse and semi-natural speech datasets. There is little focus on user-centred evaluations, especially for assistive or accessible technologies. Few systems are evaluated under noisy conditions or tested longitudinally. Real-time feedback is rarely incorporated. Despite technological advancements, most SER systems remain confined to laboratory settings with minimal transition into real-world applications. Studies must integrate usability testing, explore real-world deployments, and align with accessibility needs, particularly in educational and assistive technology domains. This would make integrating the resulting model into a real-world application viable. It is these issues that the current research aims to address. In the methodology chapter, solutions for addressing these issues have been outlined.

2.5 Human-Computer Interaction (HCI) for Late-deafened Educators

2.5.1 Overview

HCI aims to create a harmonious interaction between the user, the machine, and the system's services [23]. The purpose of a system's design can be understood by looking at the functions it offers. These functions are the actions or services the system provides. These are valuable only if the users can use them effectively and efficiently. The usability of a system refers to the extent to which it can be easily and adequately used by users to achieve their goals. The systems' overall effectiveness is realized when functionality and usability are well-balanced [23]. HCI's main emphasis is examining interfaces and how users interact with them. As HCI research has advanced, the interaction between humans and machines has evolved significantly. New types of interfaces have continuously emerged, ranging from command line interfaces to graphical user interfaces, voice user interfaces, and other smart adaptive interfaces, all aimed at enhancing the quality of interaction [122].

In addition to delivering high-quality and optimal services, HCI design aims for a universal approach that benefits everyone [24]. This encompasses diverse user groups, particularly individuals with disabilities, such as significant speech, hearing, or motor impairments, who often rely on computers as alternative communication devices. HCI designers intend to create solutions that promote the social integration of individuals with disabilities [24]. In this research as well, the aim is to provide solutions for late-deafened educators, who are often overlooked in technology design. The HCI research seeks to offer solutions to support their teaching activities by addressing their specific needs. First, facts and issues of the hearing impaired are presented, and then a review of the HCI solutions provided for the late-deafened.

According to a report from the World Health Organization (WHO), as per a fact sheet updated in February 2024, “By 2050, nearly 2.5 billion people are projected to have some degree of hearing loss, and at least 700 million will require hearing rehabilitation” [123]. The degree of hearing loss can be broadly categorized based on available audiograms and classified according to the WHO standards, 2019 [212]: Normal hearing (25 dB or better), mild impairment (26–40 dB), moderate impairment (41–60 dB), severe impairment (61–80 dB), and profound impairment (81 dB or higher). These hearing thresholds were further updated in the WHO Report on Hearing, 2024 [210]. “dB” stands for decibels, a unit of measurement of intensity of sound [210]. Hearing loss may affect one or both ears. It can arise from various causes, including congenital defects, early childhood hearing loss, chronic middle ear infections, noise exposure, ageing, or certain medications permanently damaging the inner ear [123, 131].

People born with hearing impairments often experience social isolation and loneliness. At the same time, those who acquire hearing loss or deafness later in life can also suffer significantly, feeling frustrated as their communication abilities decline, impacting their work and social lives. Nearly one in three older adults is affected by disabling hearing loss, and over one billion young adults face a heightened risk of hearing impairment due to unsafe listening practices [123].

2.5.2 Challenges of Hearing-Impaired People in Working Environment/ Workplace

Hearing plays a crucial role in the work life of any individual, enabling effective communication and enhancing job performance. Hearing loss can present several challenges, particularly in communication. It can lead to stress and anxiety for affected individuals [214]. The study by Svinndal et al. demonstrates how hearing impairment of any degree impacts work-life and highlights the benefits of providing support to such individuals [209] . Employees with hearing impairments frequently face considerable challenges in the workplace, such as barriers in group interactions and limited awareness of suitable workplace accommodations. It is a challenging condition that can negatively affect work participation [209, 215]. Provisions must be made for their inclusion and care. Adequate services and professional support are essential to assist such stakeholders in the workplace [216]. Despite all the challenges the hearing-impaired professionals face, they are often committed to maintaining their job competency and doing whatever it takes to succeed [209].

Many software applications have been built, but it is imperative to consider the needs of the hearing impaired before developing them. The idea is that what works for the hearing impaired will often work for everyone, but not necessarily the other way around. As more AI-driven applications are developed, prioritizing the needs of users with disabilities in the design of HCI applications will encourage the creation of inclusive systems for all. Considering the feedback and experiences of the hearing impaired will help build applications that support their independence and accessibility [130]. Although hearing-impaired individuals use computers the same way as those with normal hearing, user interfaces play a particularly crucial role [217].

2.5.3 Late-Deafened Educators as a Subset of Hearing Impairment Individuals

Hearing impairment is a broad term encompassing partial or complete loss of hearing in one or both ears, ranging from mild to profound levels. It can be congenital (present from birth) or acquired later in life due to factors such as age, illness, noise exposure, or injury [131, 209]. Late-deafened individuals, a subset of those with hearing impairment, specifically refer to people who lose their hearing after developing speech and language skills, typically in adulthood. Unlike those who are born deaf or lose hearing in early childhood, late-deafened individuals often face unique challenges, such as adapting to communication barriers in previously familiar environments and coping with the psychological impact of losing a previously functional sense [18]. The term “hearing-impaired educators” is also used in this work to represent late-deafened educators, as it facilitates the inclusion of the degree of hearing loss (in decibels) described above in Section 2.6.1.

Late deafened educators develop hearing loss or any form of hearing impairment later in their life, having acquired language and speech skills by then. This hearing loss can lead to mental fatigue, stress, and psychological strain, further affecting their teaching and relationships with students, as communication difficulties can alleviate confusion [29]. Many times, educators evade revealing their hearing loss issues with a fear of social stigma or with the fear that their academic career will be negatively impacted. As per Smith and Andrews, there has been a concerning rise in faculty with hearing issues [30]. This research aims to provide solutions for the late-deafened educators on issues related to effective computing when teaching online.

Student responses in a class help the educator know of the students’ understanding, learning and involvement in the ongoing class [16]. These responses carry emotions expressed through body language, facial representations, or voice. Similarly, educators gauge emotions through students’ facial and verbal responses in an online class. But often, this becomes challenging as the camera of the learners’ device is switched off. Verbal responses are heard, but gauging the emotion is not conclusive. Even hearing the student’s responses is difficult for educators with hearing issues. Hence, for late deafened educators, gauging a student’s emotion is much more difficult as compared to their colleagues without hearing impairments. This lack of a definite understanding of a student’s emotion in online classes indicates an emotional deficit in online teaching and learning [3], which can be solved by emotion recognition through affective computing.

Speech-based interaction between a computer and a human is a key area within HCI systems, focusing on information derived from various speech signals [23]. Additionally, affective computing, a branch of HCI, automatically recognises emotions through programming based on facial or audio information [5]. Developing such applications can facilitate the extraction and visualization of emotions in online education, allowing teaching methods to be adjusted according to detected emotional states. A search query on 'late-deafened educators' yielded limited results, indicating that this area is underexplored in the literature.

The following review examines the few available HCI solutions addressing late-deafened educators as presented below and summarized in Table 2.3 [refer to Appendix 3].

2.6 Related work on HCI Solutions for Late-deafened Educators

2.6.1 Challenges Faced by Hearing-Impaired Educators

Several studies have explored the challenges faced by educators with hearing impairments, highlighting gaps in existing technological solutions and inclusive practices. A personal narrative presented in [29] examines the characteristics of age-related hearing loss and the challenges experienced by senior professors with hearing impairments at universities and colleges. The researcher suggests strategies for hearing-impaired individuals to effectively adapt to age-related hearing loss, drawing on examples from her personal experiences as a senior faculty member. The article sheds light on the significant issues facing educators with hearing impairments through these insights. While the work by Tidwell [29] provides a rich, first-person account of the psychosocial and pedagogical challenges faced by senior faculty with hearing loss, the study lacks a technological perspective on adapting to modern teaching environments such as online education.

Efforts to support disabled individuals in academia are often student centric. As noted in [30], most academic institutions focus primarily on creating policies that support comfort and opportunities for students with disabilities. However, limited attention is directed toward faculty members who are deaf or hard of hearing (DHH) or experience late-onset deafness. This lack of focus risks side-lining these faculty members from essential benefits and opportunities, a concerning trend given their increasing numbers. This work presents recommendations to support DHH faculty, including workplace accommodations, funding, and preferred services. Additionally, it advocates for providing assistive technologies, such as specialized software and smartphone applications, to enhance communication. These recommendations include voice-to-text translation apps, instant face-to-face text communication with a full-sized keyboard, self-developed video-conferencing tools, and wearable devices like hearing-aid-compatible microphones. Since DHH faculty often engage in online teaching, the study stresses the need for technology to facilitate a smooth transition between in-person and virtual classes. However, it does not specifically address issues or challenges that DHH faculty may encounter in the online teaching environment.

The COVID-19 pandemic exposed accessibility gaps for hearing-impaired individuals, as discussed in [124]. The study by Deshmukh et al. assesses issues such as limited access to information, difficulties with social distancing, and the impact of mask-wearing. Among the strategies proposed to address these obstacles, integrating supportive technology is essential for promoting inclusivity. Systems such as sign language interpreters and telemedicine are suggested as effective support tools for the hearing-impaired. The authors advocate for combining technology, healthcare, empathy, and support for those with hearing disabilities. However, their work could cover more ground, particularly regarding the educational challenges faced by hearing-impaired individuals and potential solutions tailored to their needs in this field. The study lacked targeted strategies for supporting them in remote or online teaching environments. These studies [29, 30, 124] collectively highlight the institutional and psychosocial barriers faced by hearing-impaired educators, while revealing a persistent gap in technology-focused solutions for online teaching.

2.6.2 Assistive Technologies and HCI Approaches

The issue of digital inclusion is addressed in [125]. The study shows there has been significant digitization due to the centralization of the Internet in the development of education, commerce, leisure, government, communication, and health services. To promote e-inclusion, advancements in organizational, economic, societal, and technical aspects are necessary. This article aimed to bridge this gap by adapting interfacing devices, a crucial point of digital disparity. A platform was designed and tested with users having various disabilities in collaboration with numerous associations and support groups. This user-friendly platform, based on TV interfaces like gesture-controlled remotes and special keyboards, significantly improved accessibility. A 15-day pilot study in Spain involved seven individuals interacting with the platform, revealing it as useful, convenient, and accessible. Although not statistically significant, the positive feedback emphasis the platform’s potential to enhance independence and reduce the need for third-person assistance. However, its short pilot duration and absence of pedagogical context make its applicability for online teaching among late-deafened educators uncertain, but the article signifies the need to build inclusive systems.

Interactive educational platforms designed in [126] incorporated features such as sign language and speech buttons and were tested with students of mixed abilities. An HCI approach was proposed for an innovative e-learning interface to accommodate students with diverse visual and hearing needs. The interface included prototypes featuring an interaction panel with functionalities such as Button Voice, Tab Voice, Text Voice, and sign language support. The interactive features were tested by students with and without visual and hearing impairments and evaluated by the teachers. A mixed-methods approach was employed, combining quantitative measurements of student satisfaction with qualitative interviews exploring teacher perspectives on implementation challenges. The findings reveal high satisfaction levels with the new features across all student groups, regardless of their abilities. Teachers emphasized the importance of support from academic organizations, including technical, financial, and human resources, to ensure the successful implementation of such innovative solutions. Teacher feedback highlighted the importance of institutional support for inclusive tools, which resonates with the needs of late-deafened educators.

Automatic Speech Recognition (ASR) systems are evaluated in [127], where design improvements are suggested to enhance usability for deaf or hard-of-hearing users. ASR can facilitate smoother communication for DHH individuals, particularly in small-group meetings with hearing colleagues or in classroom settings. The research focuses on the HCI design of ASR captioning systems, identifying improvements to enhance usability for DHH users. This study represents a step toward creating more accessible ASR technology, considering the challenges DHH individuals face with current systems and aiming to redesign and refine it for a better user experience. While the focus is on user interface refinements, the study opens avenues for integrating ASR with emotion recognition to support educators' understanding of student feedback. From late-deafened educators' perspective, ASR technology can benefit online classes, and incorporating emotion recognition capabilities could further enhance its value by enabling a better understanding of student engagement. While HCI-driven solutions [125, 126, 127] show promise in promoting digital accessibility, their educational applicability remains limited without targeted adaptation for the needs of late-deafened educators.

2.6.3 Educational and Communication Support Technologies for HCI

Studies such as [128, 129], and [130] highlight the growing role of web and mobile-based assistive technologies as viable alternatives to traditional communication methods for the hearing-impaired, particularly in educational contexts. Baglama et al explore how technologies aimed at improving language and speaking skills—such as speech-to-text tools and web-based applications—serve as functional alternatives to sign language [128] . Their work advocates for the development of more mobile applications tailored to different diagnosis levels of hearing impairment and calls for in-service training to improve teachers' technological competence.

Amandeep and Williamjeet categorize assistive technologies into Assistive Listening Devices (ALD), Augmentative and Alternative Communication (AAC), and alert systems, emphasizing tools like speech enhancement, real-time transcription, and sign language learning applications [129]. These tools are found to be effective in enhancing communication and educational participation among hearing-impaired learners and trainers. However, both studies primarily focus on learners, leaving a gap in addressing the needs of educators—particularly in online teaching environments. Kim et al. provide a broader perspective by evaluating the usability and effectiveness of communication-assisting mobile applications, incorporating feedback from hearing-impaired users and experts [130]. They emphasize the importance of inclusive, user-centred design—particularly interfaces that foster social support and emotional connection—and advocate for the integration of AI-driven assistive tools that are responsive to users lived experiences. These insights offer valuable directions for the development of next-generation accessible technologies. While these studies [128, 129, 130] showcase mobile and assistive communication technologies for hearing-impaired users, they fall short of addressing how such tools can be adapted to support hearing-impaired educators—particularly in real-time, emotion-sensitive online teaching contexts.

2.6.4 Emerging Technologies and Future Directions

Reviews on haptic [131] and VR [132] technologies demonstrate innovation in sensory substitution and immersive training. Flores Ramones et al. conducted a comprehensive review on haptic devices, which are assistive technologies for individuals with hearing disabilities [131]. These devices provide tactile feedback through the sense of touch, allowing users with limited hearing to rely on sensory substitutes for vital information. The review focuses on the hardware and technical features of haptic devices, emphasizing variations in the number of stimulation zones, the areas where stimulation is applied, and the intended purpose of the devices. The findings from the review highlight substantial room for further enhancements, suggesting numerous avenues for future research. As the global population of individuals with hearing impairments rises, developing and refining these technologies becomes even more critical. Also, A hybrid approach that combines deep learning with graph theory to efficiently recognize sign language gestures, leveraging advancements in artificial intelligence was presented [133]. The developed system in this work was utilized for recognizing and categorizing letters in American Sign Language (ASL) by leveraging Convolutional Neural Networks (CNNs). The success of this system was attributed to the adoption of best practices in data preparation, augmentation, model architecture design, optimization, and deployment. It effectively addresses key challenges in ASL recognition, such as the variability in hand gestures and the limited availability of training data [133].

Reviews on haptic [131] and VR [132] technologies demonstrate innovation in sensory substitution and immersive training. However, these tools are often suited for learners or clinical use rather than real-time online teaching contexts. VR's isolating design may further deter older users, such as senior educators. Similarly, the ASL gesture recognition system in [133] addresses communication gaps but does not extend to teaching environments.

HCI solutions play a vital role in addressing the daily key challenges people with hearing impairments face. [134]. From assistive devices like hearing aids and captioning systems to advanced tools such as SER. HCI innovations offer the potential to improve accessibility and inclusivity significantly. HCI research plays a crucial role in developing technologies that can benefit all, especially the late-deafened educators. HCI research can contribute to more effective teaching and learning experiences by focusing on the specific needs of late-deafened individuals. The challenges faced by late-deafened educators in using online teaching systems are underrepresented in HCI literature. Studies indicate that while universities primarily focus on supporting disabled students through targeted policies, there is limited provision for faculty with disabilities, especially those who are hard of hearing, in terms of policy, technology, and application support [30]. Consequently, these educators may feel isolated, negatively impacting their job satisfaction and performance [129]. For late-deafened educators, conducting virtual classes can be particularly difficult. As discussed in Section 1.4, real-time student feedback during class is crucial for enhancing teaching quality [119]. However, if the educators have hearing difficulties, student responses may go unheard, affecting the overall teaching and learning process. As society aims to increase inclusivity across all domains, supporting late-deafened educators in teaching and training activities becomes essential. Such efforts will drive the creation of technologies that are more effective and deeply aligned with the authentic needs of all users [130]. This research seeks to address these gaps and contribute an appropriate solution.

2.7 Conclusion – Gleaning from the Review of SER and HCI Literature

The ability of educators without hearing impairments to discern emotions from students' vocal feedback during online classes enables them to adapt their teaching approaches for better engagement and understanding. In contrast, late-deafened educators face significant challenges in interpreting this feedback due to their hearing limitations. Existing online teaching platforms and HCI systems are primarily designed with the assumption that educators can hear, neglecting the needs of those with hearing impairments. This lack of accessible features in HCI tools prevents late-deafened educators from effectively gauging students' emotional states, such as happiness, sadness, and fear, thereby creating a critical gap in inclusivity and equitable access to teaching resources. Addressing this gap requires an innovative solution that bridges the accessibility divide. The proposed SER system, built on deep learning models, offers a potential remedy by enabling real-time recognition of student emotions from speech, irrespective of an educator's hearing ability. For this solution to be impactful, it must be designed to operate efficiently and accurately across diverse datasets. This research highlights the need for HCI systems to support late-deafened educators in online teaching environments. By addressing these overlooked challenges, the study contributes to an inclusive system. The next methodology chapter outlines the steps to develop and evaluate the proposed SER system, addressing the identified HCI gaps.

Chapter 3

3. Methodology

3.1 Introduction

In Chapter 2, the limitations of existing SER systems and the HCI tools that cater to the needs of late-deafened educators in online teaching scenarios were identified. To address these challenges, this study proposes a real-world SER system that not only detects emotions accurately in real time but also incorporates HCI principles to ensure usability for late-deafened educators. This system aims to bridge the gap by enabling educators to interpret student emotions effectively during online teaching. Late-deafened educators can adapt their teaching approaches as needed based on the detected emotions, thereby enhancing student engagement.

3.2 Research Approach

This study adopts a multi-phase research approach grounded in both technical innovation and user-centred design as a response to the limitations identified and opportunities highlighted in the previous chapter. As discussed in the introduction, to address the identified gap, the following research questions were articulated to guide the study, as outlined earlier in Section 1.6:

- RQ1 - What are efficient and effective pre-processing and feature engineering approaches for a deep learning-based real-world speech emotion recognition (SER) system?
- RQ2 - In the context of deep learning, specifically convolutional neural networks (CNNs), what is an effective hybridization approach for combining 1D, 2D, and 3D convolutional layers to improve accuracy whilst maintaining efficiency?
- RQ3 - How can emotional insights from an SER system be shown in a way that educators, especially late deafened, find it easy to understand and use during online classes?
- RQ4 - What is the perceived impact of an SER system built upon RQ1-RQ3 on educators with hearing impairment and educators without hearing impairment regarding outcome and experience?

To address these research questions systematically, the diagram shown in figure 3.1 presents the overall framework for the research study. This research framework serves as a structured plan or conceptual guide outlining how the research study progresses from system design to interface development and subsequent evaluation. Each stage has been designed to directly align with the study's objectives, ensuring both the technical goals and the needs of the end users are met. This approach emphasizes the integration of SER capabilities with HCI considerations, facilitating an inclusive and intuitive user experience.

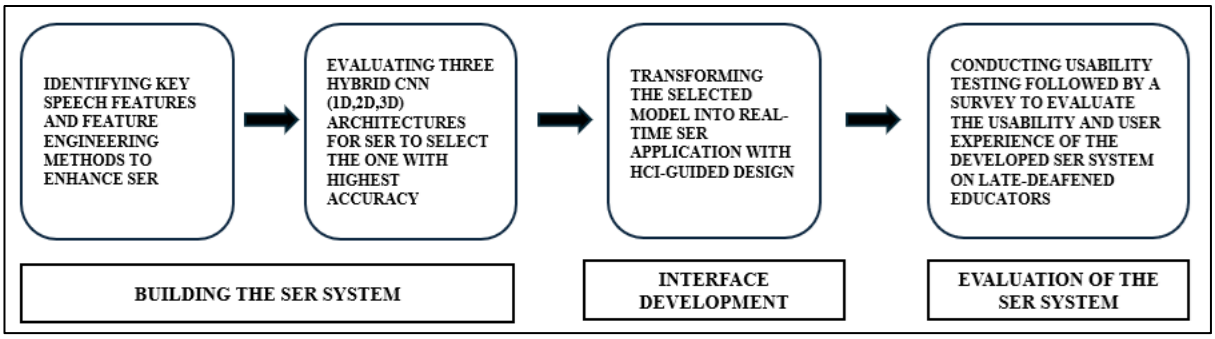

IDENTIFYING KEY SPEECH FEATURES AND FEATURE ENGINEERING METHODS TO ENHANCE SER
EVALUATING THREE HYBRID CNN (1D,2D,3D) ARCHITECTURES FOR SER TO SELECT THE ONE WITH HIGHEST ACCURACY
TRANSFORMING THE SELECTED MODEL INTO REAL-TIME SER APPLICATION WITH HCI-GUIDED DESIGN
CONDUCTING USABILITY TESTING FOLLOWED BY A SURVEY TO EVALUATE THE USABILITY AND USER EXPERIENCE OF THE DEVELOPED SER SYSTEM ON LATE-DEAFENED EDUCATORS
BUILDING THE SER SYSTEM
INTERFACE DEVELOPMENT
EVALUATION OF THE SER SYSTEM

Figure 3.1. Flow diagram for overall framework of the Research Study

This research framework aligns with the following research objectives, as outlined in Section 1.7:

- RO1 - To identify the key speech features and feature engineering approaches that enhance the efficiency and accuracy of emotion recognition in a deep learning-based real-world SER system.
- RO2 - To compare and assess the effectiveness of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers through three selected fusion techniques, identifying the approach with the highest accuracy and efficiency and adapting it to build a real-world SER application.
- RO3 - To design and develop a graphical user interface (GUI) that integrates SER-derived emotional feedback and displays emotions to late-deafened educators accurately and in real-time.
- RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency and perceived impact on late deafened educators for an effective engagement in online teaching.

Table 3.1 gives the mapping between the Research Objective and each phase of the Research framework as defined in Figure 3.1.

Table 3.1: Research Framework and Objective Mapping

Framework Stage	Description	Linked Research Objective
Identifying Key Speech Features and Feature Engineering Methods to Enhance Ser	Focus on extracting and selecting relevant speech features to improve SER system performance.	RO1

Evaluating Three Hybrid CNN (1d,2d,3d) Architectures for Ser to select the One with Highest Accuracy	Compare three hybrid models using three different fusion techniques of combining 1D, 2D and 3D CNN to determine the most effective architecture for emotion classification.	RO2
Transforming the selected Model into Real-Time SER Application with HCI-Guided Design	Integrate the model with highest accuracy into a GUI that detects and displays the student emotions to the late-deafened educator in real-time	RO3
Conducting Usability Testing followed by a Survey to evaluate the Usability and User Experience of the developed SER system on Late-Deafened Educators	Evaluate the usability and user satisfaction of the developed system with late-deafened educators through structured testing and feedback.	RO4

Building upon this framework; the following sections outline the methodology adopted for each stage. The methodology combines the technical approach with user-centric design to ensure the resulting emotion detected from the developed SER system is not only accurate but also accessible and practical for late-deafened educators. Accordingly, this chapter is organized to reflect the research process, beginning with an understanding of user needs, followed by the design, development, and evaluation of the system. It begins by describing the methodology for conducting the preliminary study, followed by the approaches for building the SER system. Subsequently, it outlines the methods used for interface development, guided by HCI principles to create an intuitive and inclusive user experience. The chapter concludes with a methodology adapted for a usability study to evaluate the effectiveness of the SER system, focusing on its practical application and user satisfaction among late-deafened educators.

The chapter is organized as follows:

1. **Preliminary Study**: This section describes the methods used to understand the needs and perspectives of late-deafened educators.
2. **Building the SER System**: This section compares the methodologies of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers using three selected fusion techniques to identify the one achieving the highest speech emotion recognition accuracy.
3. **Interface Development**: This section describes the process of transforming the selected SER system, from the three compared, into a functional real-time application for detecting and displaying emotions. It also details the design and development of the user interface, ensuring seamless integration with the backend.
4. **Evaluation of the SER system**: This section describes the methods used to evaluate the usability and user experience of the developed SER system in supporting late-deafened educators during online teaching.

Since the focus of this research study is to support effective user interaction, particularly for late-deafened educators, the methodology adopted integrates HCI principles. The goal of HCI research is to enhance the design and usability of technology by understanding how people interact with computer systems. Ultimately, the purpose is what these systems can do and how their functions contribute to achieving the system's intended goals. A system's functionality is defined by its actions or services, which should be efficiently accessible to users. Usability, in the context of a system's functionality, refers to how users can effectively and efficiently utilize the system to accomplish specific goals [23].

While SER systems provide powerful tools for emotion detection, their successful design and integration into real-world settings rely on adhering to HCI principles. HCI emphasizes the creation of user-friendly and accessible systems, while the Software Development Life Cycle (SDLC) offers a structured framework for building them. Integrating HCI principles into the SDLC ensures that usability and user needs are prioritized throughout development. SDLC generally outlines key activities at each stage of software development, including planning, analysis, design, coding, testing, and maintenance [135]. There are many SDLC models like Waterfall, Agile and such [26]. For this study a hybrid SDLC model has been adopted, combining principles from both the Incremental and Iterative approaches for HCI development process [26, 136]. The incremental component of the model is reflected in the way the system was constructed in modular stages, with each major component developed in sequence. As illustrated in Figure 3.2, the development process was divided into clearly defined stages: understanding user requirements, building the SER system, designing the user interface, evaluating the integrated system, and reporting outcomes. This approach allowed for manageable complexity by phase wise planning, development and deployment. Simultaneously, the iterative nature of the model was integrated. Throughout the process, user feedback was gathered for continuous refinement. The insights gathered during the preliminary study informed both the requirement of the SER module and the design of the user interface. Also, Pilot phase was conducted prior to the main usability testing to evaluate initial system functionality and interface clarity; findings from this pilot usability testing contributed to early refinements. By emphasizing feedback and iterative adjustments, the development process aligned with the research goals, facilitating a system that met its intended objectives. By adopting this hybrid SDLC model, the study ensured that both incremental progress and iterative feedback loops allowed the system to evolve in alignment with real-world constraints and user needs, making it suitable for practical deployment in online teaching environments. This approach maintains a structured yet adaptive workflow, prioritizing customer satisfaction and enabling faster development cycles. By combining the incremental and iterative approaches with HCI principles, the development process for the SER system focuses on making it efficient, effective, and tailored to the specific needs of late-deafened educators.

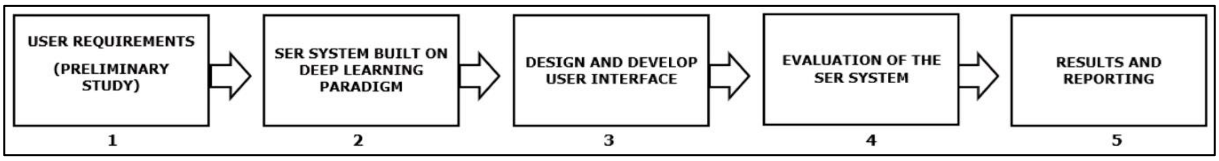

USER REQUIREMENTS (PRELIMINARY STUDY)
1
SER SYSTEM BUILT ON DEEP LEARNING PARADIGM
2
DESIGN AND DEVELOP USER INTERFACE
3
EVALUATION OF THE SER SYSTEM
4
RESULTS AND REPORTING
5

Figure 3.2 Stages of the HCI research adapted for the development of the SER system

These stages can be outlined as follows:

1. User Requirements (Preliminary Study):

User research is conducted through surveys, interviews, or observations to understand user needs and goals. For this research, the first step involved gathering information about the users who would interact with the system and understand the context in which it would be used.

As part of the preliminary study, an initial discussion was held with two late-deafened educators to explore the challenges they face in detecting student emotions during online classes and the importance of understanding student emotions during teaching. This discussion also aimed to identify how an automated system for detecting student emotions through speech could address these challenges and benefit educators.

Following this, a survey was conducted as part of the preliminary study. The survey targeted educators with hearing impairments (late-deafened) and those without hearing impairments to identify their requirements and the challenges of online teaching. The focus was on understanding the need for detecting student emotions, the difficulties in doing so during online classes, and the impact of student emotions on student engagement.

The findings from the preliminary study were crucial for ensuring that the proposed SER system effectively addressed the identified user needs.

2. SER System Built on Deep Learning Paradigm

The findings from the preliminary study highlighted the need for and benefits of an automatic SER system to assist late-deafened educators in accurately identifying student emotions during online classes. The SER system was developed after comparing three hybrid CNN architectures (a deep learning approach), with the model achieving the highest emotion recognition accuracy being selected.

3. **Design and Development of the User Interface:**

In this step, the developed SER system was integrated with a user-friendly GUI for real-time evaluation. A user-friendly web interface was developed as the system's front end, with a deliberate focus on simplicity and intuitiveness. The design prioritized the detection, visualization, and impact of the identified emotions to accurately present the detected student emotions in real-time.

4. **Evaluation of the SER System:**

The GUI-integrated SER system was tested in real-time with educators with hearing impairments (late-deafened) and without hearing impairments to evaluate the effectiveness, efficiency and perceived impact. As part of the Iterative approach of the hybrid SDLC, pilot users evaluated the system first. Feedback from the pilot users was used to make improvements in the system. This was followed by an evaluation of the system by the educators with and without hearing impairments.

5. **Results and Reporting:**

After the evaluation of the SER system via the usability study, the results were analysed and reported. The findings demonstrated a positive impact of the SER system on teaching effectiveness and efficiency.

Given this, the following sections first detail the methodology for the preliminary study, followed by the methodologies for developing the SER system, designing and developing the interface, and evaluating the SER system. These correspond to stages 1, 2, 3, and 4 in the SDLC diagram (refer Figure 3.2). The next section specifically describes the methodology for conducting the preliminary study.

3.3 Preliminary Study

The preliminary study is the first stage of the HCI research, as illustrated in Figure 3.2. This phase focuses on gathering user requirements to assess the necessity and scope of the proposed SER system. It also describes the methods used to conduct the preliminary study, providing direction for designing the SER system tailored for late-deafened educators. The preliminary user requirement study process is shown in Figure 3.3.

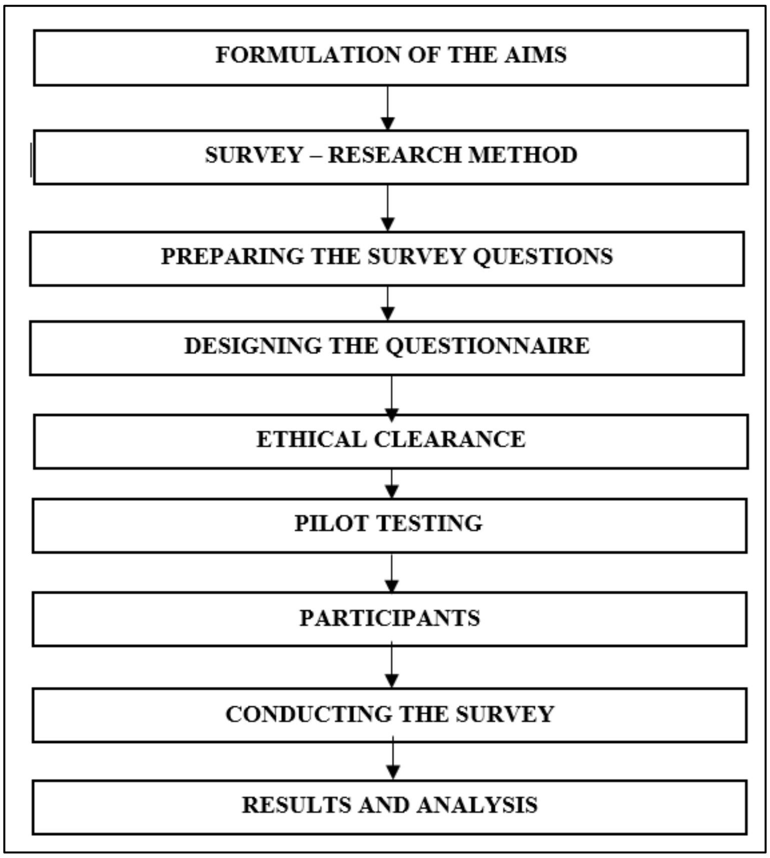


Figure 3.3. Process of the preliminary user requirement study[1]

As per the flow, the first task would be setting up the aims of this preliminary study, which are as follows -

1) Investigate the perspective of educators (with and without hearing impairment) on the necessity of detecting student emotions to enhance teaching outcomes for effective teaching.
2) Understand the challenges faced by educators (with and without hearing impairment) in discerning student emotions during online teaching.
3) Identify the benefits of having an emotion-detection system (SER) for detecting student emotions.

As seen above, the aims of the preliminary study for this research are clearly outlined. In light of this, a survey was conducted to gather information on the challenges educators with hearing impairments (late-deafened) and without hearing impairments face in discerning student emotions during online teaching sessions. The focus was also on exploring the need to detect student emotions for adjusting teaching approaches and the benefits of having a system that

[1] *Wherever the term educators with hearing impairment is used, it refers to late-deafened educators, defined in Section 2.6.4.*

automatically detects students' emotions. An analysis of the survey results indicated the potential benefits of implementing an emotion-detection system via speech, which was the intention of this research work. Quantitative and qualitative research methods have been used for the analysis of these survey results, which have been detailed in Chapter 4. The next section describes details of the method adopted for the preliminary study.

3.3.1 Survey

A survey, a widely used method in HCI research [137], was conducted to gather insights from educators. Surveys allow data collection without direct researcher interference, providing valuable perspectives on key issues. Surveys enable data collection without the researchers' interference, offering valuable insights from a sample group on key issues [138, 139]. The participants represented a sample group from a larger population [140]. The educators in question are both – (1). with hearing impairment (late-deafened) and (2). those without hearing impairment.

3.3.2 Survey Questions

For the preliminary study, a 21-question questionnaire was designed [refer to Appendix 5]. The initial section consisted of single-option questions to gather demographic information about the respondents. These included details such as gender, age, teaching experience, area of expertise, level of students taught, and whether they had formal training in education.

The next set of questions focused on the respondents' hearing conditions. These questions aimed to identify whether participants had normal hearing or any form of hearing impairment and the severity of the impairment. One question specifically asked if participants could hear student feedback or required a hearing aid in cases of severe hearing impairment. An open-ended question in this section sought insights into the challenges educators with hearing impairments face when hearing student feedback.

Subsequent questions explored educators' use of online teaching systems. Respondents were asked how many online classes they had conducted in the past two years and which tools they used for these sessions. These questions included both single-option and open-ended formats.

The final questions addressed student feedback and the recognition of emotions, covering f2f (physical) and online teaching contexts. Respondents were asked whether emotions conveyed through students' vocal feedback helped them understand student engagement and adjust their teaching strategies accordingly. This aspect was particularly emphasized for online classes, where physical interaction is absent. Additional questions investigated the potential benefits of an automated system that detects underlying emotions (such as happy, sad, and neutral) from students' vocal feedback and displays these emotions through images. The final questions assessed how understanding students' emotions during online teaching influenced educators' ability to adapt their teaching approaches.

The questionnaire included a mix of closed and open-ended questions. Some closed-ended questions utilized a 5-point Likert scale to measure responses and produce ordinal data, a common approach in survey research [138].

3.3.3 Designing the Questionnaire

After formulating the questions, the next step was to structure the questions into a coherent format and an organized questionnaire. During the questionnaire development, each question formed was ensured to align with and contribute to answering the aims of this preliminary study laid out earlier. The questionnaire [refer to Appendix 5] was divided into four sections to collect comprehensive information from the respondents.

- Section A gathered demographic details and personal profiles.
- Section B focused on questions linked to the respondent's hearing condition
- Section C explored how respondents interacted with online systems.
- Section D addressed questions related to student feedback and their emotion recognition in online systems.

This structured approach provided a thorough understanding of the respondents' backgrounds, experiences, and perspectives on the topic at hand.

3.3.4 Obtaining Ethical Clearance

Obtaining approval from an ethics committee is mandatory to ensure that the study adheres to established ethical standards. As this research involved human subjects and a socially significant group for the survey, the questionnaire received the necessary ethical clearance [refer to Appendix 4].

3.3.5 Pilot Testing

Before administering the questionnaire to the target respondents, a pilot test was conducted with a small group of individuals. These pilot testers assessed the questions' clarity, relevance, and effectiveness in achieving the study's objectives [142], enabling any issues to be addressed before the preliminary study [138]. Pilot test participants belong to the target population [138]. Hence, an academic and a corporate trainer were selected for this study. The academic participant is highly experienced, with over 30 years of service at a prestigious university. The corporate trainer, with 20-plus years of experience—including 10 years in academia before transitioning to a project manager and corporate trainer role in an IT company—is well-versed in pedagogy and system design. Their valuable feedback and suggestions were incorporated into the questionnaire before distributing it to the recruited participants for the preliminary study.

3.3.6 Participants

The next step in the preliminary study was identifying suitable participants. According to Goode and Hatt [140], a sample refers to a subset of individuals or participants selected from a larger population for research or testing purposes. Sampling involves selecting participants to draw inferences about the broader population [138, 140]. A participant represents the research focus and is part of a defined group of individuals. Selecting the right participants is crucial to ensure the generalizability of the findings. Appropriate sampling techniques are vital to obtain a representative sample from the identified population.

For this study, selecting respondents—educators with and without hearing impairments—was purposeful rather than random. Participants were recruited using a purposive sampling approach. As defined by Adolph Jenson [140], "A purposive selection denotes the method of selecting a number of groups of units in such a way that selected groups together yield as nearly as possible the same average or proportion as the totality with respect to those characteristics which are already a matter of statistical knowledge." In purposive sampling, participants are chosen based on specific criteria to ensure their relevance to the study [140, 141, 143]. This approach helps identify respondents with the required background and experience to provide meaningful responses, focusing on a well-defined population of interest.

The criteria for participant selection in this preliminary study included years of teaching experience, experience with online teaching, and hearing condition. Specifically, participants were required to have a minimum of four years of teaching experience; additionally, all participants were required to have experience teaching online (mostly all educators taught online during the COVID-19 pandemic, as noted by UNESCO IESALC, 2020 [8]), and information about their hearing condition. This purposeful approach ensured that the sample frame was designed to achieve adequate representation, including educators with and without hearing impairments.

Further, in the case of educators with hearing impairment, a snowballing method was used, where connections were established with the participants based on referrals. Snowball sampling is a technique in which one approaches individuals from the population to help identify participants who could be suitable for the ongoing study [138, 142, 143]. This approach was used as educators were required to deal with specific problems of late-onset deafness. As Lazar et al. [144] noted, research involving users with disabilities often accepts smaller sample sizes, typically 5–10 participants with a specific disability. This is due to the difficulty of identifying participants who meet all inclusion criteria, such as employment, education, or technical expertise. Consequently, studies focusing on users with disabilities generally have smaller sample sizes compared to those involving participants without disabilities. However, a small number of participants is often sufficient to identify usability issues specific to users with disabilities, as demonstrated in prior research [145, 146, 147].

The recruitment process, using the snowballing method, was initiated by the researcher, with friends and colleagues being contacted to request introductions to educators in their network who had hearing impairments. These initial contacts referred additional potential participants, creating a chain of referrals through which the sample size was expanded. This approach was particularly effective for reaching a niche group of participants, such as educators with varying degrees of deafness, who might not have been easily identifiable through conventional recruitment methods. 10 participants with late-onset deafness, ranging from mild to profound, and 23 participants who were educators with normal hearing/no hearing impairments were successfully recruited for this study.

3.3.7 Procedure for Conducting Preliminary Study

The structured questionnaire was administered through an online survey, which is the research tool used to collect data from the survey. The questionnaire was administered through Google Forms, with the link, consent form, and participant information sheet distributed via email to potential respondents.

3.3.8 Analysis and Findings

The results and analysis from the preliminary study have been presented and elaborated in Chapter 4. Quantitative and qualitative methods were employed in this preliminary study to gather insights from educators, both with and without hearing impairments, on how they receive student feedback and discern emotions in that feedback during online sessions. The quantitative and qualitative findings from the survey, based on both open-ended and close-ended questions have been discussed in detail in Chapter 4. The next section outlines the methodology for the SER system.

3.4 Speech Emotion Recognition System

This section describes the next stage (Stage 2) of the HCI research adapted for developing the SER system, as seen in Figure 3.2. As discussed in Section 1.4.2, understanding students' emotions during online classes can be addressed through a reliable SER system capable of accurately identifying underlying emotions. An SER module offers significant benefits for late-deafened educators teaching online and all educators. The significance of SER lies in the inherent challenge machines face in identifying emotions from audio speech. Chapter 2, Section 2.3, details that developing an SER system involves essential steps [44]. The process begins with collecting datasets of audio recordings labelled with their corresponding emotional categories. Next, the audio data is pre-processed, and relevant features are extracted for analysis. These features are then fed into deep learning classifiers and trained on the labelled data to recognize patterns associated with different emotions. Finally, the trained model detects and categorizes emotions such as happiness, sadness, and others based on the learned patterns.

As concluded in Section 2.4, deep learning classifiers outperform traditional methods in detecting a wide range of emotions, making them well-suited for this study. This work used CNNs, a deep learning technique, to develop the SER system. Specifically, multidimensional CNNs (1D, 2D, and 3D CNNs) were employed for deep feature extraction and emotion classification.

The next sections detail each step in developing the SER system for accurate emotion detection. The discussion begins with the first step in the process, where the datasets used for this research have been described.

3.4.1 Semi-Natural and Acted Datasets

Among the three types of datasets—acted, semi-natural, and natural—discussed earlier in Section 2.3.2 of the SER components, two semi-natural and three acted datasets were utilized. Although semi-natural datasets are simulated, they are closer to real-world conditions, making them suitable for building real-time applications [19, 20, 43]. Specifically, the IEMOCAP and DEMoS semi-natural datasets have been used for this study. Additionally, three acted datasets—TESS, RAVDESS, and EMODB—were selected. Despite being acted, these datasets are well-suited for machine learning and deep learning models and have been frequently utilized for testing [46]. The rationale for using a variety of datasets was that it ensures the model's reliability and robustness for its usage in real-world application development [35, 56, 73, 103, 118]. The five different datasets used to train the model are described below.

1. Interactive Emotional Dyadic Motion Capture Database (IEMOCAP)

IEMOCAP is an English language dataset. This semi-natural dataset captures emotions considered naturally expressed rather than entirely simulated. These emotions are comparatively close to reality and thus are well-suited for developing practical applications. The dataset contains approximately 1,150 utterances of clean subset with selected clear emotion labels spoken by five male and five female speakers [43, 44]. The original set of emotions in the database included anger, happiness, sadness, and frustration. Later, four additional categories—disgust, fear, excitement, and surprise—were added. The dataset was also annotated with continuous dimensions [43]. This licensed dataset is well known and extensively used in traditional machine-learning models and is increasingly popular in experiments involving convolutional neural networks [46, 148].

2. Database of Elicited Mood in Speech (DEMoS)

DEMoS is an Italian language dataset. It consists of seven emotional states—six primary ones (anger, sadness, happiness, fear, surprise, and disgust), along with an additional emotion, guilt. It contains approximately 9,697 speech samples, recorded by forty-five males and twenty-three females. These are not acted emotions, instead, the speech is elicited using Mood Induction Procedures (MIP) to provoke genuine emotional responses. A three-step process was followed before the dataset was made available for usage. First, a mood induction procedure was applied to evoke specific emotions. Second, a combination of the alexithymia test and self and external assessments was used to select typical samples. Finally, machine learning techniques assessed how the emotion typicality impact algorithm performance [149]. The DEMoS dataset is distributed under a license for use.

3. Toronto emotional speech set (TESS)

TESS is an English language dataset. It contains 2,800 audio recordings representing the emotions of anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. These emotions were portrayed by two female actors, aged sixty and twenty. Although the students did not participate in recording the speech samples, they were involved in labelling the emotions. A group of 56 undergraduate students identified the emotional content of each sentence and only sentences with a labelling confidence above 66% were retained in the final dataset [43]. The dataset is frequently used in SER research due to its emotional clarity and balance across categories. It is often used for testing systems developed with machine learning and deep learning models [46, 150].

4. Ryerson Audio-Visual *Database* of Emotional Speech and Song (RAVDESS)

RAVDESS is a multimodal dataset of facial and vocal expressions in North American English. It consists of 1,440 speech audio files representing eight emotions: calm, neutral, happy, sad, angry, fearful, surprised, and disgusted. It features twenty-four professional actors, evenly split between twelve male and twelve female, all recording their utterances in a neutral North American accent. Each emotion is recorded twice—once in normal speech and once in a singing voice—providing two distinct intensities. This dual format makes it unique, earning it the title of a database of emotional speech and song [43, 44]. Due to its comprehensive nature, the dataset is popularly used in machine learning and deep learning models [46].

5. Berlin Emotional *Database* (EMO-DB)

EMO-DB is a German language dataset. This publicly available emotional dataset contains 700 sentences in German, capturing emotions such as anger, boredom, anxiety, happiness, sadness, disgust, and neutrality. The recordings were done by ten speakers [43]. To maintain naturalness, the speakers recorded commonly used or everyday sentences from memory rather than reading from a script. Since trained actors often express emotions in an exaggerated manner, performers for EMO-DB were recruited through a newspaper advertisement rather than relying on professionals [46, 151]. Each recruited person expressed the seven mentioned emotions across approximately 10 sentences. It serves as a reliable benchmark for evaluating classification models and emotion recognition techniques using machine learning and deep learning approaches [44, 151]. Table 3.2 shows a comparative summary of the SER datasets used for this study.

3.4.2 Feature Selection and Extraction

Following the selection of appropriate datasets, the next crucial step is feature selection and extraction from the speech samples. The choice of features significantly impacts the classifier’s accuracy and overall performance. Traditional classifiers rely on manually engineered acoustic speech features known as LLDs. These include prosodic features such as pitch, energy, and duration; spectral features like Mel-frequency cepstral coefficients (MFCCs) and Linear Predictive Coding (LPC); voice quality features like jitter and shimmer; and TEO-based features

Table 3.2: Comparative Summary of SER Datasets Used for This Study

Dataset	**Language**	**Type of dataset**	**Emotions Covered No, of Emotions**	**No. of Speakers**	**Total Samples**	**Notes**
IEMOCAP	English	Semi-Natural	9	10	1150 for selected four emotions and about 12,000 for 9 emotions	The dataset comprises dyadic voice recordings in English, capturing performers engaged in both scripted and spontaneous interactions [34].
DEMoS	Italian	Semi-Natural	8	68	9,500	The 68 participants (23 females, 45 males) were all students from an engineering Faculty. DEMoS is particularly suitable for studies involving young adults or university students. Emotions are elicited through realistic prompts, not theatrically acted, offering higher ecological validity [149].
TESS	English	Acted	7	2	2,800	Students did not participate in recording the speech samples; however, they were involved in labelling the emotions.
RAVDESS	English	Acted	8	24	1,440 speech samples only	The actors are young adults, between 20-40 years old [68].
EMO-DB	German	Acted	**7**	10	535	To encourage natural speech, familiar everyday sentences were chosen, allowing actors to deliver them from memory rather than reading from a script [46].

such as TEO-FM-VAR. These features have been discussed in detail in Section 2.3.3. However, the handcrafted LLDs alone are insufficient for effectively determining emotions in SER [152, 153]. Therefore, deep learning techniques automatically extract deep features from the data without manual intervention. These algorithms capture high-level features, called deep features, from the low-level information of the input data [56, 120, 121]. For this work, a combination of acoustic LLDs and deep features have been employed to enhance the performance of SER. A feature fusion approach integrating temporal and spectral speech features with deep features has been used to achieve an efficient SER system.

As demonstrated in [35, 57, 73, 90, 154], a classifier performs better when using a carefully selected or optimized set of features from a speech signal. For this research work, acoustic features such as Zero Crossing Rate (ZCR), Root Mean Square (RMS), Chroma STFT, MFCC, and Mel-Spectrograms were extracted from the datasets and sent as input into deep the network architecture. From these LLDs, deep features were extracted, producing a feature vector used for final predictions.

The acoustic features can be categorized into temporal and spectral features. Temporal features are linked to time-varying properties, and spectral features are linked to frequency-domain properties of speech. The time-varying characteristics of speech are speaking rate, pause duration, intonation, and such. They are related to the timing of events in a speech and hence can be used to differentiate between different types of speech. On the other hand, spectral features relate to the frequency domain properties of the speech signal and how they change over time. Feature fusion combines the strengths of multiple features [57]. These features help to distinguish between different phonemes and help identify the speaker's age, gender, and accent [56, 120, 155]. When utilized with deep learning techniques, a mix of local and global speech features and discriminative features help enhance the recognition rate in the case of real-time SER applications using different speech datasets [60, 156]. The different features selected for this work have been described below.

3.4.2.1 Description of the Selected Acoustic Features

1. ZCR

This is a widely used audio feature in speech and audio-related applications. By analysing the temporal characteristics of speech, ZCR proves valuable for various speech-processing tasks. It assists in detecting voice activity within speech segments by distinguishing between voiced, unvoiced, and silent sections in a speech frame. It measures the rate at which the waveform crosses the zero axis, indicating sign changes within a given period [56, 120]. This means the number of times a signal changes its sign from positive to negative and vice versa, divided by the length of the frame. For a discrete-time signal x[n] of length N, the Zero-Crossing Rate can be represented by a formula as below:-

$$\text{ZCR} = \frac{1}{2N} \sum_{n=1}^{N} |sgn(x[n]) - sgn(x[n-1])| \tag{3.1}$$

where sgn(x) is a sign function and $\text{sgn(x)} = \begin{cases} 1\ if\ x > 0 \\ 0\ if\ x = 0 \\ -1\ if\ x < 0 \end{cases}$

and N is the total number of samples in the analysed signal segment [120].

2. RMS

This feature describes the volume or loudness of a speech signal, which is a crucial feature of the human auditory system. Since it measures the overall loudness, it is commonly used in speech-related applications. To avoid clipping or distortion, RMS helps maintain the overall loudness of an audio signal within a certain range [120, 121]. It is measured by getting the square root of the sum of the mean squares of the amplitudes of the sound samples. For a discrete-time signal x[n] over a window of N samples, RMS can be represented by a formula as below:-

$$\text{RMS} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (x[n])^2} \tag{3.2}$$

$x[n]$ is the amplitude of the speech signal at the nth sample. N is the total number of samples in the segment or window of the analysed speech signal.

3. Chroma-STFT

This refers to a type of spectrogram based on the chromatic scale. It is also known as a chromogram. It is derived using the audio signal’s logarithmic Short-Time Fourier Transform (STFT). Chroma features effectively capture the pitch and tonality of speech, which are closely linked to emotional expression. These features can differentiate the pitch class profiles of various audio signals, aiding in the identification of pitch and overall signal structure. This makes them valuable for speech-processing applications such as speaker identification, emotion recognition, and speech-to-text transcription [157, 120, 121].

4. Mel Spectrograms

Mel Spectrograms are an effective feature in applications linked to speech emotion recognition, providing a time-frequency audio signal representation. They are computed by applying the Short-Time Fourier Transform (STFT) to the speech signal and mapping the resulting frequencies onto the Mel scale, aligning with how humans perceive sound. This feature captures pitch, intensity, and energy changes over time—attributes often linked to emotional expression. For instance, anger may be characterized by higher energy and pitch, while sadness could result in lower intensity and a flatter pitch contour [73, 157]. A Mel spectrogram represents the power spectrum of an audio signal over time, with frequency bands scaled according to the Mel scale. It provides a detailed visualization of how energy is distributed across frequencies at each time frame. This rich time-frequency representation can capture changes in tone, pitch, and energy, making it valuable for analysing the prosodic features of speech that often convey emotion, such as intonation and rhythm. MFCCs are derived from the Mel spectrogram but undergo additional processing. They are described as the next feature. After generating the Mel spectrogram, the logarithmic Mel filterbank energies are transformed using a Discrete Cosine Transform (DCT), which results in a compact set of coefficients (MFCCs) [158, 159]. Mel Spectrograms can be represented by the following formula [94]:-

$$\mathrm{M[m,t]} = \sum_{k=0}^{N-1} |X\ [k,t]\ |^2\ H_m(f[k]) \tag{3.3}$$

M[m,t] is the Mel spectrogram at Mel filter m and time frame t. X[k,t] represents the STFT of the speech signal at frequency bin k and time t. $\mathrm{H_m(f[k])}$ is the m$^{\mathrm{th}}$ Mel filter applied to the frequency bin f[k]. This formula captures the signal's power spectrum and maps it onto the Mel scale. By analysing the Mel spectrogram, speech emotion recognition systems can detect variations in pitch and tone that correspond to different emotional states.

5. MFCCs

Mel-frequency cepstral coefficients (MFCCs) are one of the most valued features in speech processing applications. Human-produced sounds are unique to each individual, shaped by the specific configuration of the vocal tract, including elements like the tongue and teeth. This shape plays a crucial role in determining an individual's voice, and the arrangement of these elements defines the phonemes produced. This shape is reflected in the short-term power spectrum envelope, as captured by MFCCs. MFCCs are derived from the short-term power spectrum of a sound signal, using a Mel-scale filter bank to extract frequency components, followed by a discrete cosine transform (DCT) to compute the cepstral coefficients. By capturing the spectral shape, MFCCs have become widely used in SER research and other speech-related applications [73, 116]. MFCC can be represented by the following formula [121]:

$$\text{MFCC[n]} = \sum_{m=0}^{M-1} (\log \left(\sum_{k=0}^{K-1} \left(\frac{|X[k]|^2}{N} . H_m[k] \right) \right) . \cos [\frac{\pi\, n}{M} (m + 0.5)]) \tag{3.4}$$

MFCC[n] represents coefficients computed to represent a frame’s spectral properties. Index n of the MFCC being computed usually ranges from 0 to the number of desired coefficients (e.g., 12 or 13). Lower indices represent the coarse structure of the spectrum, while higher indices capture finer details. k is an index that runs over the Mel filter bank outputs, which range from 1 to K, where K is the number of Mel filters (typically around 20-40) used. X[k] is the energy output of the kth Mel filter. It is obtained by first converting the signal’s power spectrum into the Mel scale using a filter bank and then summing the energy within each Mel filter. The logarithm of the Mel filter energy compresses the dynamic range of the filter bank energies, which is analogous to how human ears perceive sound intensity. Discrete Cosine Transform (DCT) basis functions are used to decorrelate the log Mel energies and convert them into cepstral coefficients. This helps compress the information into a small set of coefficients that effectively represent the spectral properties of the speech signal.

3.4.2.2 Reason for Selecting the Acoustic Features

The idea behind selecting these specific acoustic features (described above) for speech emotion recognition lies in their complementary strengths, each addressing different aspects of the speech signal that are crucial for capturing emotional nuances. Utilizing multiple audio features instead of relying on just one integrates diverse sound characteristics into a single training sample. This results in a more comprehensive representation of the audio, enhancing the performance of speech emotion recognition models [73, 111, 150, 155, 158]. Using a broad set of complementary features to improve emotional detection accuracy and generalization.

Among the selected speech features, ZCR differentiates voiced emotions like anger and happiness, which have higher sign changes, from unvoiced ones such as neutral and sadness, where ZCR decreases due to subdued articulation. RMS indicates loudness, with higher values in energetic emotions like anger and happiness, moderate levels in neutral speech, and lower RMS in sadness due to softer delivery. Chroma STFT highlights pitch and tonal shifts, showing dynamic variations for anger and happiness, flat monotones for sadness, stable intonation for neutral speech, and variable patterns for fear, depending on its expression. Mel Spectrograms, mapping frequencies onto a perceptual scale, reveal high-frequency energy for anger and happiness, smooth, subdued patterns for sadness, balanced contours for neutral speech, and fear oscillating between intense and restrained ranges. Finally, MFCCs capture vocal traits, with anger producing distinct patterns from forceful articulation, happiness reflecting varied intonation and rapid delivery, sadness showing slower, less articulate speech, neutral speech exhibiting typical spectral distribution, and fear alternating between high-pitched rapidity and hesitant tones. These features allow emotions to be grouped as high-energy (anger, happiness), low-energy (sadness, fear), or moderate (neutral), enabling clear visualization and computational analysis [27, 57, 120].

3.4.2.3 Extracting the Selected Acoustic Features for the Multi-dimensional CNNs

1. Extracting ZCR, RMS and Chroma STFT

These are one-dimensional features that provide information about the temporal and spectral characteristics of the speech signal. ZCR and RMS are time-domain features, while Chroma STFT is a spectral feature computed using a one-dimensional Fourier transform [121]. The Librosa library in Python is used to extract these features. The speech files are loaded with a specified sampling rate, typically 16,000 Hz, suitable for capturing human speech frequencies. The average value of each feature is calculated across the entire signal, resulting in separate arrays for ZCR, RMS, and Chroma STFT. These arrays are concatenated to form a single feature vector of 14 features, used as input for a 1D CNN.

2. Extracting MFCC's

MFCCs are widely used features in speech processing, particularly for speech emotion recognition tasks [73, 120, 121]. They are derived from the audio signal's spectrum. The dimensions of MFCCs typically consist of MFCC coefficients and the number of frames. While the number of coefficients is often 12 or 13, this can vary. A higher number of coefficients can capture more spectral detail. On the other hand, the number of frames is determined by the length of the speech signal and the selected frame size. The Librosa library in Python is used to extract these features. The 'mfcc' function from Librosa is used for extraction, with parameters like the number of coefficients (n_mfcc) and the minimum frequency (fmin) specified. For this work, the number of coefficients was set to 40, with a frame size of 150. This resulted in a 2D MFCC array with a shape of (40, 150), which can be represented as a 2D matrix for the 2D CNNs [160]. This 2D array is input to a 2D CNN, which utilizes 2D convolutional filters to extract features [121].

3. Extracting Mel spectrograms

Mel spectrograms are 2D representations of audio signals, with time on the x-axis and Mel frequency bands on the y-axis. They are computed using melspectrogram function from Librosa, which applies the Short-Time Fourier Transform (STFT) to map frequencies onto the Mel scale, mimicking the human ear's sensitivity to different frequencies. Parameters such as the number of Mel bands (n_mels), hop length, and sampling rate are specified during computation, and the resulting power-scale array is converted to a decibel (dB) scale to represent loudness. Padding is added to ensure consistent spectrogram length, and the 2D spectrogram is reshaped into a 3D tensor with dimensions such as (number of Mel bands, number of time frames per segment, total number of segments). In this study, the specified dimensions are (128, 100, 10), chosen to balance detail and computational efficiency. Using (128, 100, 10) strikes a balance between detail and computational efficiency. A higher number of Mel bands generally provides a more detailed representation of frequency content, though it also increases computational complexity. A segment length of 100-time frames captures meaningful temporal patterns, such as changes in pitch and energy, without redundancy or excessive computational overhead. Dividing the audio into 10 segments ensures the entire signal is analysed while preserving temporal variations without overwhelming the model. The formatted Mel spectrogram data allows the 3D CNN to analyse time and frequency variations, with 3D convolutional filters effectively capturing patterns across three-dimensional audio data [79].

3.4.3 Classification using CNN

As outlined in Section 2.3.5, deep learning architectures such as multilayer perceptrons (MLPs), CNNs and long short-term memory networks (LSTM-RNNs) have been widely employed in SER for their ability to learn high-level features [20, 36]. Among these, CNNs—originally popular in computer vision applications like image classification, face recognition, and object detection—have gained significant use in speech-related tasks, including speech processing, recognition, and emotion recognition [36, 57, 150]. Compared to other methods, CNNs require less preprocessing and have also demonstrated strong performance in real-world applications [161]. Their strength is automatically extracting discriminative features without human intervention, making them effective classifiers. For this work, CNNs have been used to develop the SER system, specifically multidimensional CNNs. 1D, 2D, and 3D CNNs were employed for deep feature extraction and subsequent classification.

CNNs

CNNs employ a series of filters on input data to extract high-level features, which are then used for classification. A typical CNN model consists of multiple convolutional layers that learn local features from input data, refining them through subsequent layers for classification or prediction

[20, 28, 80, 156, 161]. The architecture of a CNN broadly consists of three main components: convolutional layers, pooling layers, and fully connected layers. Convolutional layers apply a set number of convolution filters to the input image, while pooling layers reduce the dimensionality of the feature maps, thereby decreasing processing time and memory requirements. Fully connected layers extract global features from local feature maps and perform classification on these features. These layers are typically organized hierarchically, with multiple convolutional layers followed by pooling layers and fully connected layers toward the end.

A CNN is a hierarchical neural network where convolutional layers and pooling layers sequentially extract features, transforming the input image into higher-level representations, as shown in Figure 3.4. The initial layers capture basic features like pixels and edges, whereas deeper layers capture more complex local features. The final fully connected layer aggregates these local features into a global representation, which is then passed to a SoftMax classifier to generate class probabilities.

The convolutional layer is the most critical component. It comprises a set of convolutional filters, also known as kernels. A kernel is defined as a grid of discrete values, with each value referred to as a kernel weight. Initially, random numbers are assigned as the weights of the kernel at the start of the CNN training process. The input image, represented as an N-dimensional matrix, is convolved with these filters to produce the output feature map [102]. A convolutional layer repeatedly applies filters (kernels) to small regions of the input image, performing a dot product operation to produce a single value in the output feature map for each position. Each convolutional kernel creates a feature map where activation values indicate the presence of specific features, and multiple feature maps are generated within each convolutional layer [156].

The output of the filter is provided to another mathematical function called an activation function. The Rectified Linear Unit (ReLU) activation function is commonly used in neural networks to introduce non-linearity into the model. It works by setting all negative values in the input to zero while leaving positive values unchanged. This allows the network to learn complex patterns by breaking linearity, making the backpropagation of gradients more efficient (addressing the vanishing gradient problem) compared to many other activation functions (activation functions are described in Section 2.3.1). Between convolutional layers, pooling layers are applied to add robustness and reduce computation.

The pooling layer's role is to reduce the spatial dimensions of the convolved features, which decreases the memory and computational requirements for processing the data. This dimensionality reduction also helps extract key features that increase the model's relative invariance to position and rotation, maintaining the model's effectiveness during training. Pooling not only shortens training time but can also help prevent overfitting. Some of the most common forms of pooling include max pooling and average pooling. Max pooling is the most commonly used pooling technique, which retains the maximum value in a local neighbourhood and discards the rest. In many CNNs, the final portion of the network consists of fully connected layers. The networks fully connected (FC) portion consists of one or more fully connected feedforward layers, typically located at the network's deepest layers. It takes input from the output of the final pooling or convolutional layer, which is flattened before being passed to the first FC layer. Flattening involves converting the multidimensional output (a 3D matrix) into a vector. Adding one or more FC layers allows the model to learn nonlinear combinations of high-level features extracted by the convolutional layers. This is followed by a SoftMax layer to calculate the probability of each class. The success of CNN models depends on carefully selecting the kernel numbers, shapes, sizes, strides, and pooling neighbourhoods [28, 156, 162].

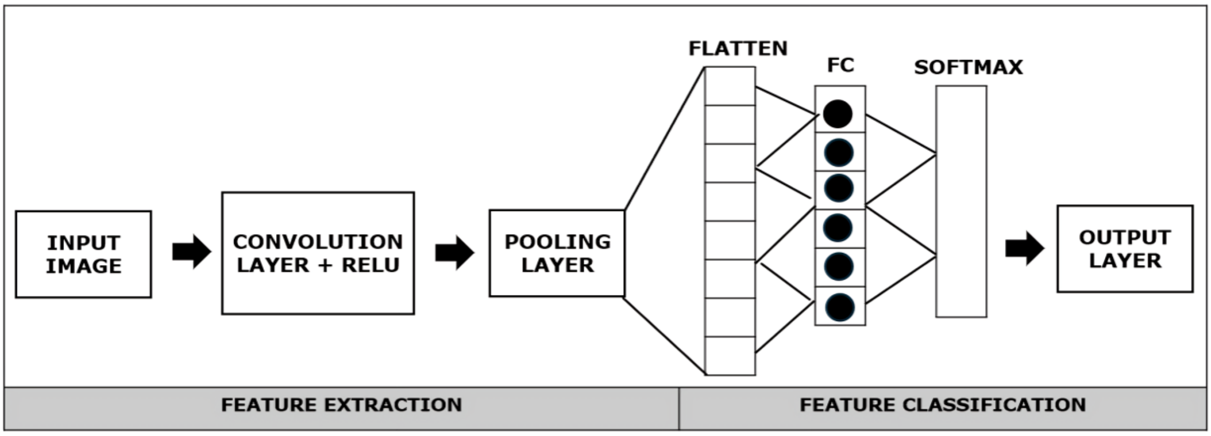


Figure 3.4: CNN architecture

Multi-Dimensional CNNs

Further, in this work, multi-dimensional CNNs were utilized. Driven by the success of hybrid models, which have shown promising results using multi-dimensional CNN approaches [35, 70, 71, 105], this study focuses on combining 1D, 2D, and 3D CNNs to capture different features of speech signals. This enables more accurate emotion classification by leveraging CNN fusion architectures for semi-natural and acted datasets. The deep features extracted from the 1D, 2D, and 3D CNNs possess complementary attributes, allowing them to be effectively integrated within a multi-dimensional architecture, yielding enhanced performance in speech emotion classification.

As per work done in [163, 164, 165, 166], multi-dimensional CNNs are designed to handle data in varying dimensions, such as one-dimensional sequences, two-dimensional matrices, or three-dimensional volumes, by tailoring their structure and operations to the specific dimensional characteristics of the input. A 1D CNN is designed for one-dimensional data, such as sequences or signals, applying filters along a single axis to capture temporal or sequential patterns. 2D CNNs operate on two-dimensional data, such as images or matrices, using filters to analyse spatial relationships across height and width. These are the most common CNNs, extensively used in image recognition and processing tasks. 3D CNNs, on the other hand, extend these operations to three dimensions, analysing data across height, width, and depth, making them suitable for volumetric data or temporal-spatial patterns in videos or medical imaging. With the rapid growth in data volumes, traditional models often struggle to balance classification accuracy and computational efficiency. To address these challenges, integrated CNNs have been introduced as a promising solution [166]. By combining or sequentially using 1D, 2D, and 3D operations, these networks can extract richer and more comprehensive features from datasets, enhancing performance in tasks ranging from signal analysis to multimodal applications.

In this work, a comparison and assessment of the effectiveness of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers using three selected fusion techniques is presented. Inspired by previous studies [35, 70, 71, 105, 163, 164, 165, 166] that successfully employed similar hybrid CNN approaches, the focus of this study is placed on three hybrid multi-dimensional CNN combinations—Model A (averaging), Model B (parallel merging), and Model C (sequential integration)—for speech emotion recognition (SER). These combinations are represented as straightforward yet distinct strategies for integrating multi-dimensional CNNs. These three models were selected to investigate the feasibility and effectiveness of hybrid multi-dimensional CNNs for SER and to prioritize the model that demonstrated the most promising performance. The model achieving the highest accuracy was identified as suitable for real-world SER applications. This selected model was subsequently refined and adapted for implementation in a practical SER system.

A hybrid model that combines 1D CNNs, 2D CNNs, and 3D CNNs can be used for SER by processing the speech signal in different ways and combining the results to make a final prediction. For example, the 1D CNN can be used to process the raw waveform, the 2D CNN can be used to process the spectrogram, and the 3D CNN can be used to process a sequence of spectrograms over time. The output of each network can be concatenated or fed into a fully connected layer for classification. Overall, the hybrid model combining different types of CNNs can improve the accuracy of SER by capturing different aspects of the speech signal, making it a promising approach for this task.

In the next section, the three hybrid architectures, each combining the three types of CNNs (1D, 2D, and 3D) using different fusion techniques, have been described.

3.4.4 Comparison of Three Selected Hybrid Architectures Integrating 1D, 2D, and 3D CNNs

This section compares three selected hybrid architectures to identify the one best suited for building our real-world SER system. The three hybrid architectures, combining the three types of CNNs (1D, 2D, and 3D), using different fusion techniques, are – (1). Averaging (Model A), (2). Parallel merging (Model B), and (3). Sequential integration (Model C). These models have been designed to leverage the fusion of acoustic features. The first two architectures (Models A and B) are multi-stream CNNs: Model A fuses streams by averaging decisions, while Model B concatenates the output from the multi-streams for further processing. In a multi-stream**,** multiple parallel CNN streams process different inputs or features separately before combining their outputs at a later stage. Such a multi-stream approach generally helps improve the model’s performance [167]. The third architecture (Model C) corresponds to a sequential combination of 3D, 2D, and 1D CNNs. Configurations of all three proposed models, averaging (Model A), parallel merging (Model B), and sequential integration (Model C), have been described below:-

3.4.4.1 Network Architectures of the Three Selected Models

1. Model A (hybrid average)

Model A (hybrid average) exemplifies the hybrid average approach. The speech features described in Section 3.4.2.1 were fed as input to the multi-dimensional CNNs. For this approach, 1D CNN processed one-dimensional sequences - ZCR, RMS, and Chroma-STFT features; 2D CNN handled two-dimensional data, which are MFCCs, which form a 2D representation of sound. 3D CNN works with three-dimensional data, often conceptualized as volumes. For this work, a 3D CNN processes mel spectrograms segmented into time frames, creating a 3D representation. For this method, 1D, 2D, and 3D CNNs were independently trained, and the corresponding speech features were fed into each model. The predictions from the three models were averaged to generate the final prediction. A Score-level fusion [168] method combined the predicted scores (probabilities) of the three classifiers, making the final prediction more robust. This technique aggregated the confidence scores of multiple classifiers for each class, leveraging the strengths of different models to improve classification performance. The score level fusion can be expressed as –

Averaging the Predictions (Score-Level Fusion): The average prediction is computed for each class by taking the mean of the confidence scores from the three models:

$$Y_{pred\_avg}(c) = \frac{1}{3}\sum_{i=1}^{3} y_{i\,(c)} \tag{3.5}$$

where $Y_{pred\_avg}(c)$ is the average score for class c and $y_{i\,(c)}$ is the predicted confidence score for class c from model i.

Argmax to Get Final Predicted Class: After averaging the scores, the argmax function is applied to select the class with the highest averaged score for each input:

$$C^{\wedge} = \text{argmax}\ (Y_{pred\_avg}\ (\text{c})) \tag{3.6}$$

Where C^ is the predicted class label.

Figure 3.5 shows the simplified architecture of the proposed hybrid average method (i.e., Model A).

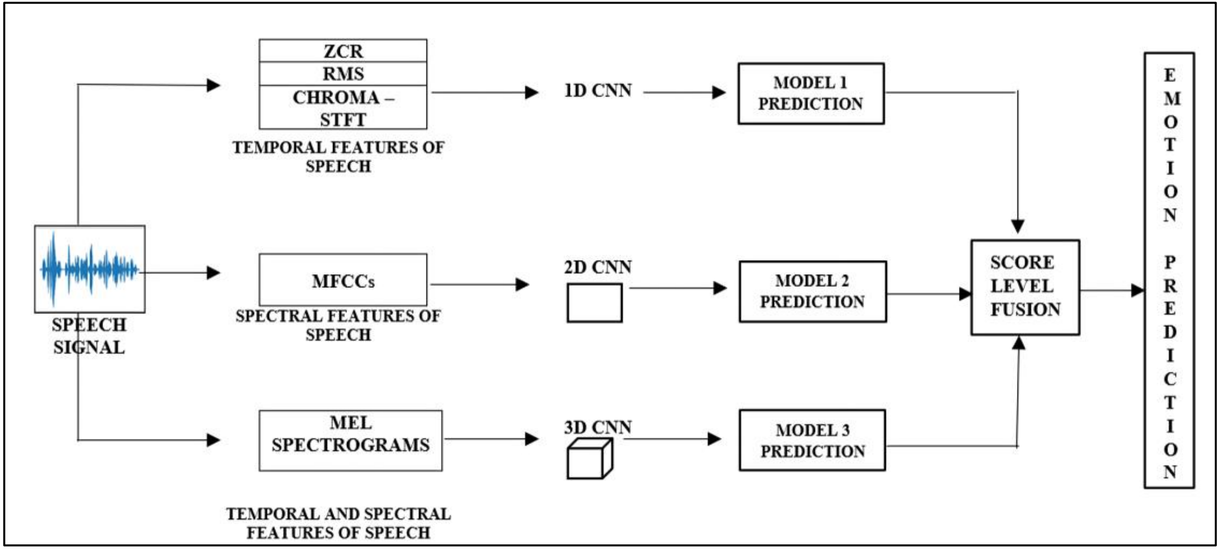


Figure 3.5: Model A (hybrid average)

2. Model B (hybrid merge)

Similar to Model A (hybrid average), Model B (hybrid merge) also uses separate CNNs for the different feature types. However, instead of averaging the predictions, it concatenates the outputs from the CNNs before feeding them into dense layers for further processing. In this approach as well, the 1D CNN processes one-dimensional sequences—ZCR, RMS, and

Chroma-STFT features; the 2D CNN handles two-dimensional data, specifically MFCCs; and the 3D CNN processes mel spectrograms segmented into time frames, creating a 3D representation. The outputs from the multi-stream CNNs (1D CNN, 2D CNN, and 3D CNN) are merged, and the result is passed through additional layers (Dense and SoftMax) for final classification [98]. The hybrid merge method combines multiple convolutional neural networks (CNNs) to process different types of input data, enhancing the model's ability to recognise patterns in diverse features. Figure 3.6 shows the overview architecture of the proposed hybrid merge method (i.e., Model B).

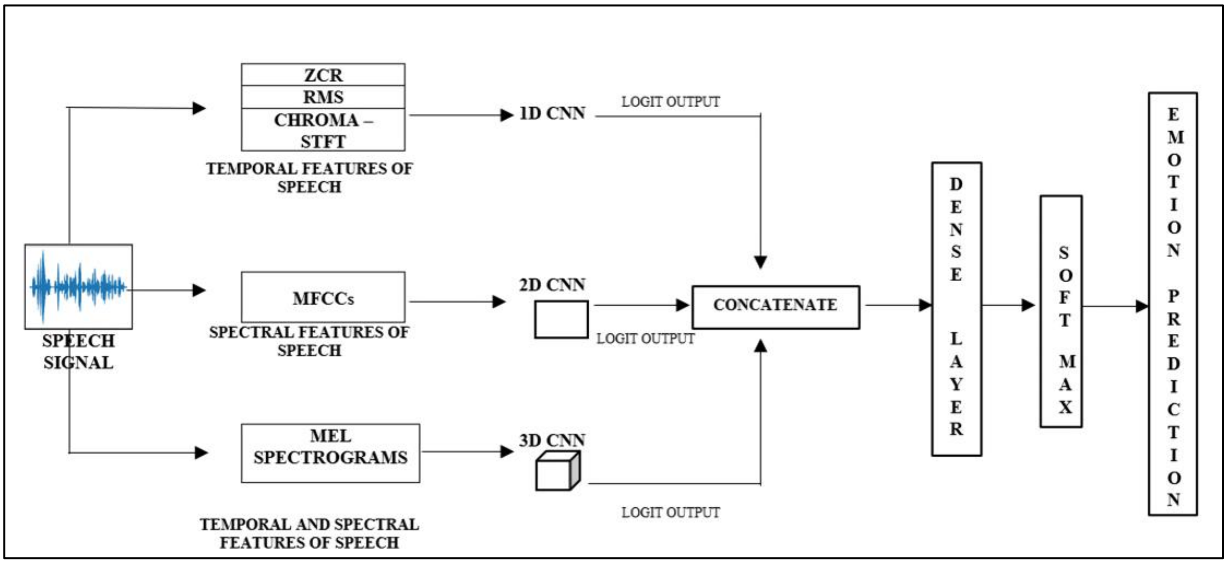


Figure 3.6: Model B (hybrid merge)

3. Model C (hybrid sequential)

In the hybrid sequential approach, all the speech features—ZCR, RMS, Chroma-STFT, MFCCs, and Mel spectrograms—are combined and fed as input data, which is passed through the individual CNN models in sequence: 3D, followed by 2D, and then 1D. The final output is then obtained for classification [105, 169]. The architecture consists of convolutional layers with progressively decreasing dimensions, culminating in dense layers for the final classification. Finally, the result is passed through a SoftMax classifier to generate predictions for different emotions. Figure 3.7 below shows the overview architecture of the proposed hybrid sequential method (Model C).

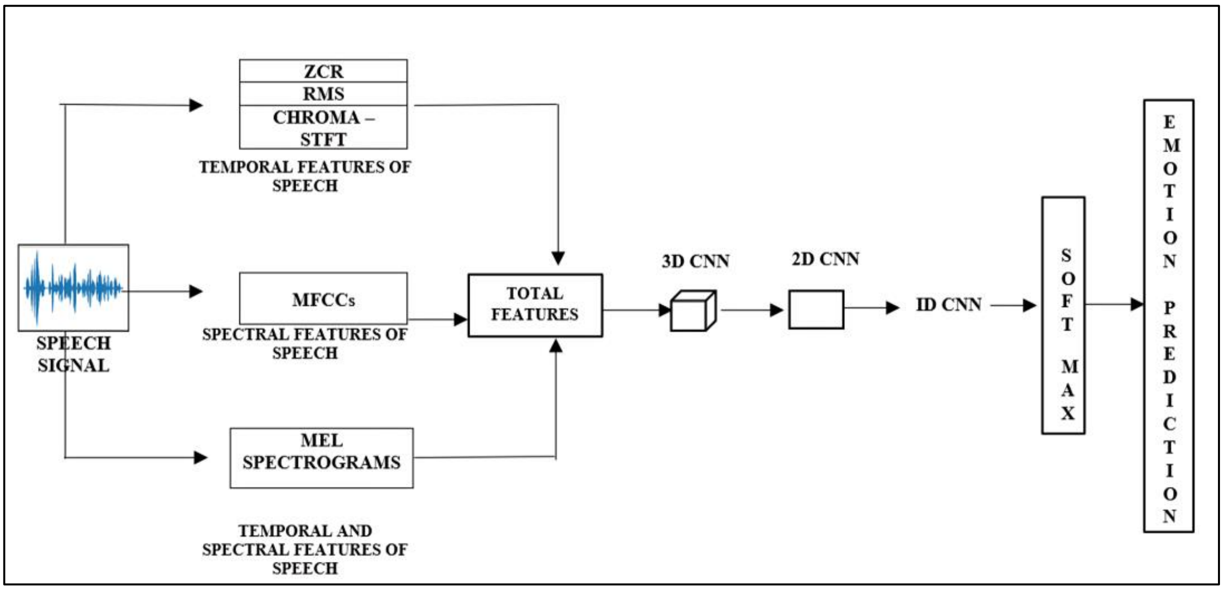


Figure 3.7: Model C (hybrid sequential)

The network architectures of the three selected models—Model A (averaging), Model B (hybrid merge), and Model C (hybrid sequential)—have been described. The speech features used in this work: Zero Crossing Rate (ZCR), Root Mean Square (RMS), Chroma STFT, MFCCs, and Mel spectrograms, were detailed in Section 3.4.2.1. The extraction of these features and their preparation as input for multidimensional CNNs were explained in Section 3.4.2.2. LLDs such as ZCR, RMS, Chroma-STFT, MFCCs, and Mel spectrograms are fundamental features extracted from audio signals. These serve as inputs for deep learning models to learn more abstract and high-level (deep) features. The process of learning deep features from LLDs allows the model to capture more complex patterns and relationships in the data, which can lead to improved classification performance in tasks of SER [35, 90, 160]. In the next section, the deep features are described learnt from LLDs and the configurations of the three selected models—Model A (averaging), Model B (hybrid merge), and Model C (hybrid sequential).

3.4.4.2 Learning Deep Features from LLDs and the Network Configuration of the Three Selected Hybrid Models

After extracting multiple speech features, the next step involved building deep features from these LLDs (low-level descriptors), which were used as input to multi-stream CNNs for Model A (averaging) and Model B (hybrid merge). For Model C (hybrid sequential), all the LLDs were combined and passed sequentially through individual CNN models, starting with 3D, followed by 2D, and finally 1D CNN layers. Before training the model, each selected dataset was split into 75% for training and 25% for testing. After splitting, the data was converted into NumPy arrays to ensure compatibility with machine learning models.

In the proposed CNN architectures, a minimal number of convolutional layers with small receptive fields were employed to effectively capture deep, salient, and discriminative features from speech spectrograms. This design not only enhances accuracy but also reduces computational complexity [60], as demonstrated by the experimental results presented in Section 4.3. In the next section, the network configurations of the three models have been described.

1. Model A (hybrid average)

a) 1D CNNs Extracting ZCR, RMS and Chroma STFT

The 1D CNN configuration processes time-series features extracted from speech files to classify emotions. The process begins with a speech file as input, from which three key 1D features are extracted: Zero Crossing Rate (ZCR), which indicates the rate at which the audio signal crosses the zero amplitude level and provides insights into the signal's frequency content and noisiness; Root Mean Square (RMS), representing the average energy or loudness of the signal over time and capturing intensity variations; and Chroma Short-Time Fourier Transform (STFT), which captures the spectral energy distribution over 12 chroma bands, offering information about the pitch content and harmonic characteristics. Combined, these features result in a 14-dimensional feature vector for each time step in the audio signal.

These features are input into a 1D CNN, designed to learn patterns and relationships within the time-series data. The architecture consists of two Conv1D layers. It starts with a 1D convolutional layer (Conv1D), which applies 64 filters of size 5 to extract local features from the input sequence. The activation function used is ReLU (Rectified Linear Unit), which introduces non-linearity by zeroing out negative values while allowing positive values to pass through. The layer uses 'same' padding to maintain the input dimensionality and has a stride of 1. This is followed by a max-pooling layer (MaxPooling1D), which reduces the spatial dimensions by taking the maximum value in each pooling window, with a pool size and stride of 1. A batch normalization layer is then applied to normalize the activations and stabilize the training process, followed by a dropout layer with a rate of 0.2 to reduce overfitting by randomly deactivating a fraction of the neurons. The second convolutional block follows a similar structure, with a Conv1D layer containing 32 filters and ReLU activation, again paired with max-pooling, batch normalization, and dropout. The convolutional outputs are then flattened using a Flatten layer to convert the multi-dimensional data into a 1D vector suitable for fully connected layers. The dense layers include 128 and 64 units, each using ReLU activation to learn complex representations. A dropout layer with a rate of 0.5 is added between them to mitigate overfitting further. Finally, the output layer consists of 5 units with a SoftMax activation function, which converts the outputs into probabilities for multi-class classification. The model is compiled with the Adam optimizer for efficient gradient-based optimization, sparse categorical cross-entropy as the loss function for handling integer-labelled outputs, and accuracy as the evaluation metric.

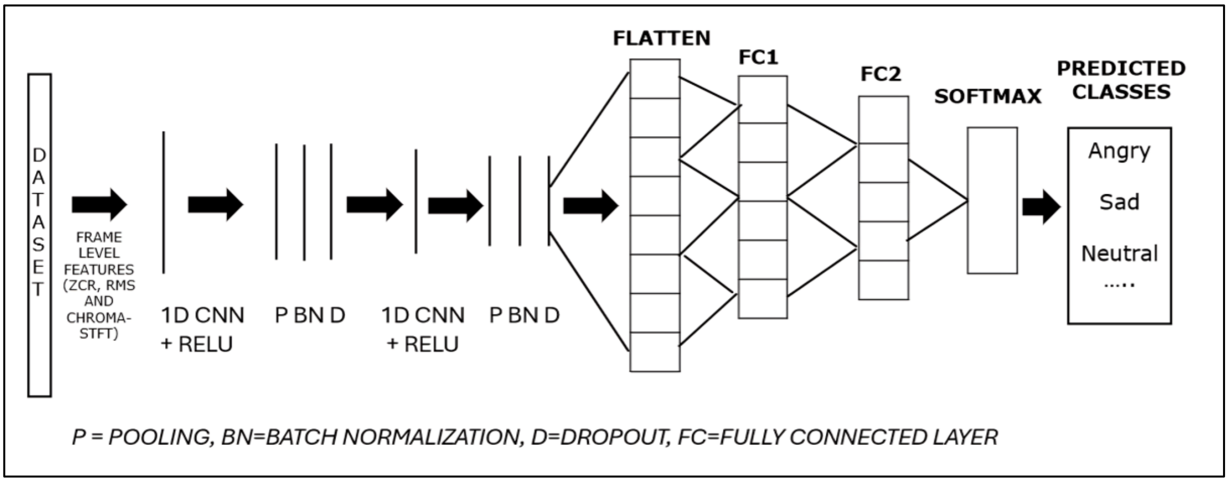

DATASET
FRAME LEVEL FEATURES (ZCR, RMS AND CHROMA-STFT)
1D CNN + RELU
P BN D
1D CNN + RELU
P BN D
FLATTEN
FC1
FC2
SOFTMAX
PREDICTED CLASSES
Angry
Sad
Neutral
.....
P = POOLING, BN=BATCH NORMALIZATION, D=DROPOUT, FC=FULLY CONNECTED LAYER

Figure 3.8. Block Diagram of the 1D CNN architecture

b) 2D CNNs Extracting MFCC's

The 2D CNN configuration processes Mel-Frequency Cepstral Coefficients (MFCCs) extracted from speech files to classify emotions. MFCCs, representing the spectral envelope of the audio signal, provide a compressed representation of its frequency content. These 2D features are fed into a 2D CNN designed to learn spatial patterns for emotion recognition. In this work, the number of coefficients was set to 40 and the frame size to 150, resulting in a 2D MFCC array with a shape of (40, 150). This 2D array is treated as a matrix and serves as input to the 2D CNN. As 2D CNNs are designed to process two-dimensional data, the input consists of a 3D tensor representing the MFCC features extracted from the audio signals. The input shape for the 2D CNN is defined as (40, 150, 1), where the dimensions represent the features extracted from the audio signal. The first dimension, 40, corresponds to the number of MFCC coefficients, capturing the frequency features of the audio. The second dimension, 150, represents the time frames, providing the temporal resolution of the signal. The final dimension, 1, indicates a single channel, meaning that each (coefficient, time) pair contains a single intensity value corresponding to the MFCC feature. This structure allows the 2D CNN to process the input as a single-channel matrix of audio features.

The configuration begins with a 2D convolutional layer (Conv2D) that applies 64 filters of size 5x5 to extract spatial features from the input data. The layer uses ReLU (Rectified Linear Unit) as the activation function, which introduces non-linearity by setting negative values to zero while allowing positive values to pass through. A stride of 2x2 is used to reduce the spatial dimensions of the feature maps. This is followed by a max-pooling layer (MaxPool2D) with a pool size of 2, which further reduces the spatial dimensions by selecting the maximum value within each pooling window. Batch normalization is then applied to stabilize the training process by normalizing the activations, and a dropout layer with a rate of 0.2 is added to reduce overfitting by randomly deactivating a fraction of the neurons. The second convolutional block follows a similar structure, with a Conv2D layer containing 32 filters of size 4x4 and ReLU activation, again paired with max-pooling, batch normalization, and dropout. After the convolutional blocks, the outputs are flattened using a Flatten layer to prepare them for fully connected layers. The dense layers include 128 and 64 units, both with ReLU activation to learn complex, high-level representations. A dropout layer with a rate of 0.5 is added between the dense layers to further prevent overfitting. The final output layer consists of 5 units with a SoftMax activation function, which converts the outputs into probabilities suitable for multi-class classification.

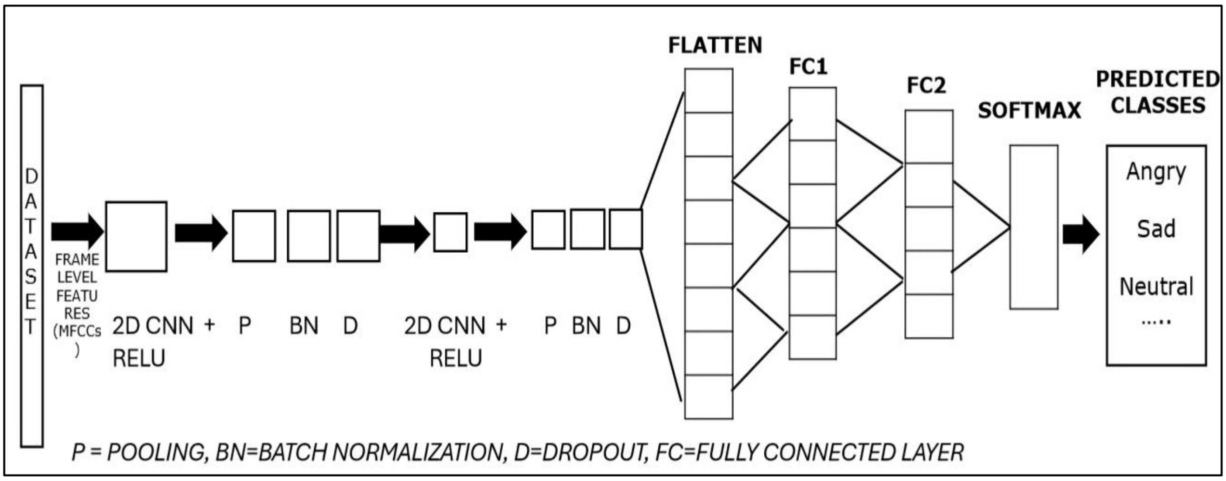

DATASET
FRAME LEVEL FEATURES (MFCCs)
2D CNN + RELU
P
BN
D
2D CNN + RELU
P
BN
D
FLATTEN
FC1
FC2
SOFTMAX
PREDICTED CLASSES
Angry
Sad
Neutral
.....
P = POOLING, BN=BATCH NORMALIZATION, D=DROPOUT, FC=FULLY CONNECTED LAYER

Figure 3.9: Block Diagram of the 2D CNN architecture

c) 3D CNNs extracting Mel spectrogram

Mel Spectrograms are used to process three-dimensional data. These networks are particularly useful when working with video data. They can also be applied to SER by using the Mel spectrogram as a 3D input volume [79]. A 3D convolutional neural network (CNN) is used for this work, designed to handle data with spatial and temporal dimensions, such as video or audio features with time, frequency, and channel components. In this case, the input data is structured as a 4D tensor with the shape (mels, time frames, segments, channel). Here, mels (128) represents the number of Mel-frequency coefficients, time frames (100) correspond to the number of time frames in the signal, segments (10) refers to how the audio data is divided into smaller segments, and channel (1) indicates that there is only one feature per (mel, time frame, segment). This structure allows the 3D CNN to process the input as a spatiotemporal representation of audio, where the model captures both temporal and frequency-based patterns.

The configuration starts with a 3D convolutional layer (Conv3D) that applies 32 filters of size 3x3x3 to extract spatiotemporal features from the input data. The activation function used is ReLU (Rectified Linear Unit), which allows positive values to pass through while setting negative values to zero. The input shape is defined based on the training data. This is followed by a 3D max-pooling layer (MaxPooling3D) with a pool size of 2x2x2 to reduce the spatial and temporal dimensions. A dropout layer with a rate of 0.5 is added to reduce overfitting by randomly deactivating a portion of the neurons. The second convolutional block includes another Conv3D layer with 64 filters of size 3x3x3 and ReLU activation. This is followed by another 3D max-pooling layer with the same pool size (2x2x2) to further down-sample the feature maps. After the convolutional blocks, the outputs are flattened using a Flatten layer to prepare the data for the dense layers. The dense layers consist of 128 units with ReLU activation to learn high-level representations from the extracted features. A dropout layer with a rate of 0.5 is applied to prevent overfitting. Finally, the output layer contains 5 units with a SoftMax activation function, converting the outputs into probabilities for multi-class classification.

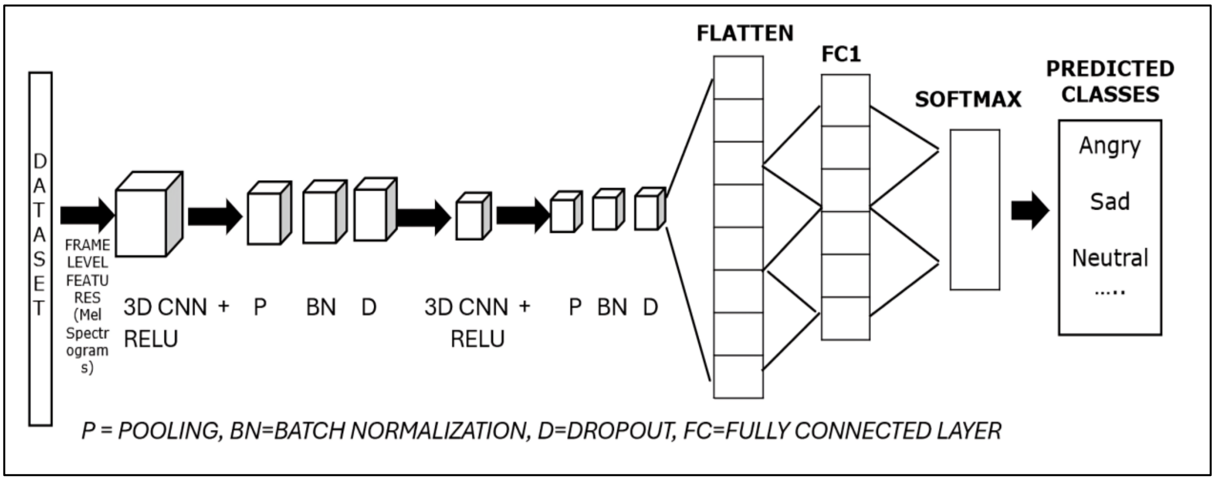

DATASET
FRAME LEVEL FEATURES (Mel Spectrograms)
3D CNN + RELU
P
BN
D
3D CNN + RELU
P
BN
D
FLATTEN
FC1
SOFTMAX
PREDICTED CLASSES
Angry
Sad
Neutral
.....
P = POOLING, BN=BATCH NORMALIZATION, D=DROPOUT, FC=FULLY CONNECTED LAYER

Figure 3.10 Block Diagram of the 3D CNN architecture

d) Score-level fusion of the multi-stream CNNs

The three multi-stream CNNs, namely, 1D CNN, 2D CNN and 3D CNN, are independently trained with their corresponding speech features, resulting in three independent CNN models. Predictions on the test dataset are made and stored as arrays for the three models, respectively. Predictions from the three models are averaged to generate the final prediction. A Score-level fusion [168] method combines the predicted scores (probabilities) of the three classifiers, making the final prediction more robust. This technique aggregated the confidence scores of multiple classifiers for each class, leveraging the strengths of different models to improve classification performance. Here, the score-level fusion on the SoftMax probabilities from multiple classifiers was performed.

2. Model B (hybrid merge)

Similar to Model A (hybrid average), Model B (hybrid merge) also uses separate CNNs for the different feature types. However, instead of averaging the predictions, it concatenates the outputs from the CNNs before feeding them into dense layers for further processing. In this approach as well, the 1D CNN processes one-dimensional sequences—ZCR, RMS, and Chroma-STFT features; the 2D CNN handles two-dimensional data, specifically MFCCs; and the 3D CNN processes mel spectrograms segmented into time frames, creating a 3D representation. The 1D CNN processes 1D audio features with two convolutional layers using 256 and 64 filters of size 3, applying the ReLU activation function and L2 regularization to avoid overfitting. Each convolutional layer is followed by max-pooling to reduce dimensionality, batch normalization for stability, and dropout (0.5) to improve generalization. The 2D CNN processes MFCC features with two convolutional layers using 256 and 64 filters of kernel sizes (5x5) and (4x4), respectively, with strides (2,2) for down sampling. Similar to the 1D stream, ReLU activation, max-pooling, batch normalization, and dropout (0.5) are applied sequentially. The 3D CNN processes Mel spectrogram features with two 3D convolutional layers, each using 256 and 64 filters with a kernel size of (3x3x3). Max-pooling is performed with a pool size of (2x2x2) to reduce spatial and temporal dimensions, followed by dropout (0.5) and batch normalization. The outputs from the three CNN streams are flattened and concatenated to merge the extracted features. The combined features pass through a dense (fully connected) layer with 128 units and ReLU activation, followed by another dropout layer (0.5) to reduce overfitting. The final output layer uses the SoftMax activation function to classify the input into 5 emotion classes. The model takes three separate inputs (1D, 2D, and 3D feature tensors) and produces a single output. It is compiled using the Adam optimizer with a loss function of sparse categorical cross-entropy and tracks accuracy as the evaluation metric. This architecture integrates multi-scale features from different dimensions to enhance emotion recognition performance.

3. Model C (hybrid sequential)

This model employs a hybrid sequential deep learning architecture integrating 3D, 2D, and 1D CNNs to process audio features. All the features discussed in Section 3.4.2—ZCR, RMS, Chroma-STFT, MFCCs, and Mel Spectrograms—are combined and fed as input data. Specifically, ZCR contributes 1 feature, RMS results in 1 feature, Chroma-STFT provides 12 features, MFCCs (which capture the spectral envelope of the audio signal and compress its frequency content) contribute 40 features, and the Mel Spectrogram (representing the frequency content over time using a Mel scale) adds 128 features. This results in 182 local features extracted from the audio file. These features are combined and used as input to the model, as illustrated in Figure 3.11. The model begins with a 3D CNN layer, which processes the input shape of (182, 1, 1, 1) using 32 filters with a kernel size of (5,5,5) and ReLU activation. A 3D max-pooling operation with a pool size of (5,5,5) and a stride of 2 follows, reducing the spatial dimensions while retaining the most relevant features. The output is then reshaped to remove unnecessary dimensions, making it compatible as input for the subsequent 2D CNN layer. The 2D CNN layer applies 32 filters with a kernel size of (5,5) and ReLU activation to the reshaped input. A 2D max-pooling operation with a pool size of (5,5) and a stride of 2 further reduces the feature map dimensions. A dropout layer with a rate of 0.2 is applied to ease overfitting. The output of the 2D CNN layer is then reshaped to serve as input for the 1D CNN layer. The 1D CNN layer applies 16 filters with a kernel size of 5 and ReLU activation, extracting relevant sequential features. A 1D max-pooling operation with a pool size of 5 and a stride of 2 further down samples the data. After the convolutional layers, the output is flattened and passed through a dense layer with 16 neurons and ReLU activation to combine high-level features. A dropout layer with a rate of 0.3 is added for regularization. Finally, the model concludes with a SoftMax output layer containing 5 units, corresponding to the five emotion classes. This hybrid architecture effectively leverages hierarchical feature extraction across 3D, 2D, and 1D dimensions to process and classify audio features, ensuring robust performance for emotion recognition tasks.

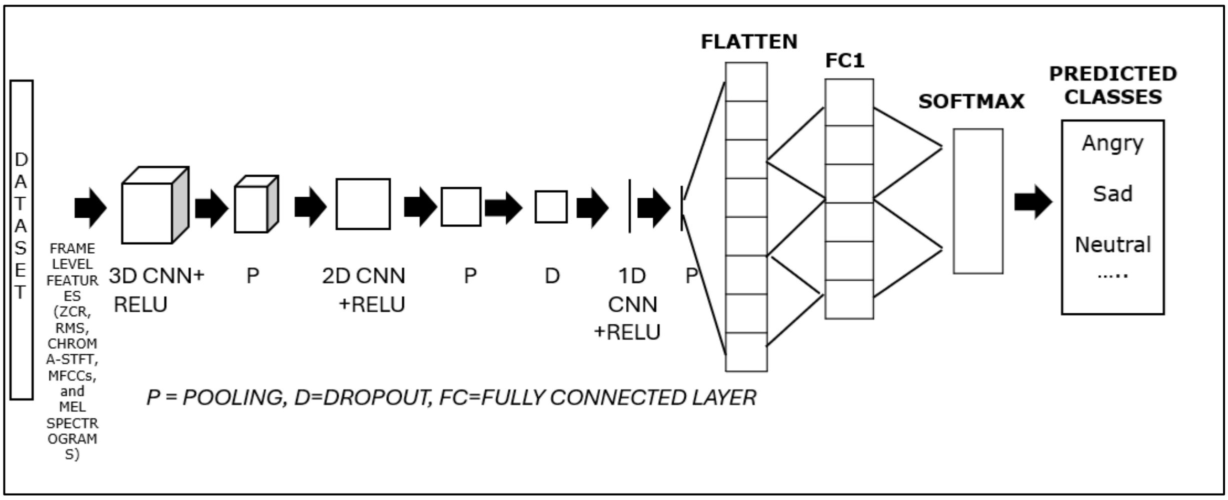

DATASET
FRAME LEVEL FEATURES (ZCR, RMS, CHROMA-STFT, MFCCs, and MEL SPECTROGRAMS)
3D CNN+ RELU
P
2D CNN +RELU
P
D
1D CNN +RELU
P
FLATTEN
FC1
SOFTMAX
PREDICTED CLASSES
Angry
Sad
Neutral
.....
P = POOLING, D=DROPOUT, FC=FULLY CONNECTED LAYER

Figure 3.11: Block Diagram of the Model C (hybrid sequential) architecture

In the proposed CNN architectures, a minimal number of convolutional layers have been utilized, with small receptive fields to effectively capture deep, salient, and discriminative features from speech spectrograms. This approach enhances accuracy while reducing computational complexity, as demonstrated by our experimental results [60].

In the above section, the network configuration of the three hybrid architectures, Model A (averaging), Model B (parallel merging), and Model C (sequential integration), was presented. These models were designed to leverage the fusion of acoustic features, each combining three types of CNNs (1D, 2D, and 3D) using different fusion techniques. The first two architectures, Models A and B are multi-stream CNNs. Model A fuses streams by averaging decisions, while Model B concatenates the outputs from the multi-streams for further processing. The third architecture, Model C, combines 3D, 2D, and 1D CNNs sequentially.

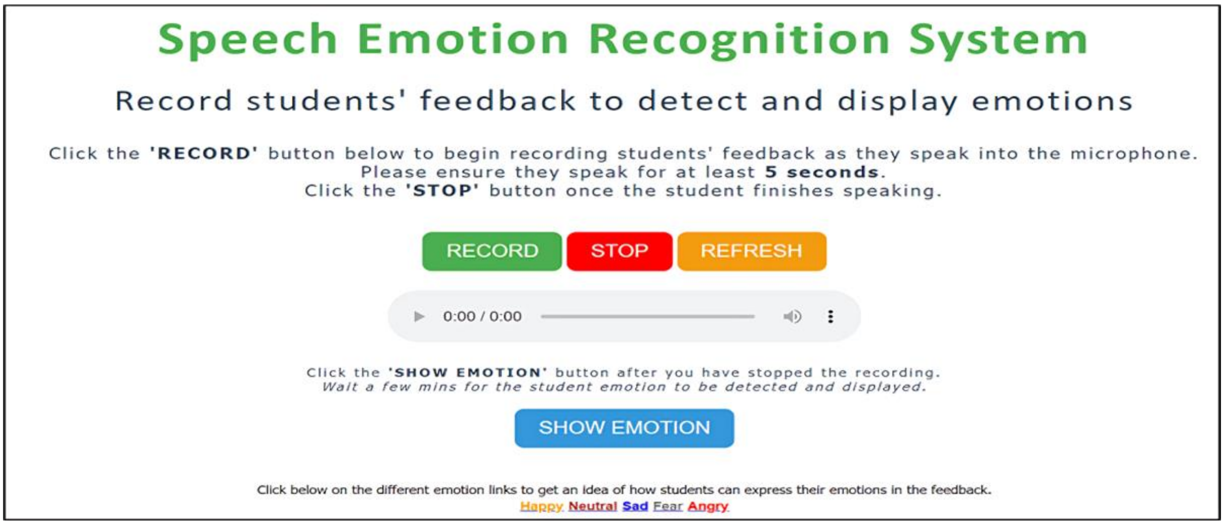
Section 3.4 of the methodology focused on identifying and extracting key speech features, referred to as low-level descriptors (LLDs)—ZCR, RMS, Chroma-STFT, Mel Spectrograms, and MFCCs. Each of these features has complementary strengths, capturing distinct acoustic characteristics of the speech signal that are essential for identifying emotional nuances. By leveraging multiple audio features rather than relying on a single feature, the methodology achieves a more comprehensive representation of the audio signal, thereby enhancing the performance of speech emotion recognition models. Following the feature extraction process, the methodology outlined how the LLDs were used as inputs for multi-stream CNN models and how the deep features were learned. In Model A (averaging) and Model B (hybrid merge), the extracted features were processed in parallel. Conversely, Model C (hybrid sequential) combined all LLDs and passed them sequentially through individual CNN layers, starting with 3D CNN, followed by 2D CNN, and finally, 1D CNN layers. Using a broad set of complementary features to improve emotional detection accuracy and generalization concludes the achievement of the research objective RO1.

In Chapter 4, the results from all three hybrid architectures are compared and presented. The experiments were conducted based on the methodology described in Section 3.4.4 for the three hybrid classifiers. The experiments demonstrated that Model A (hybrid average) achieved the highest accuracy among the tested models, as detailed in Chapter 4. Consequently, Model A (hybrid average) was selected for integration into a real-time application with a suitable graphical user interface. The following section focuses on the methodology for transforming the SER system (Model A – hybrid average) into a functional, real-time application capable of detecting and displaying emotions instantly. It also details the design and development of the user interface for the SER system, ensuring seamless integration with the backend.

3.5 Design and Development of the Graphical User Interface for SER

The methodology of three hybrid architectures—hybrid average, hybrid merging, and hybrid sequential was described in Section 3.4.4. Chapter 4 details the experimental comparison of these architectures and the selection of the best-suited model for building the proposed real-world SER system. Among the three, the architecture with the highest accuracy was chosen to transform into a functional, real-time application capable of instantly detecting and displaying emotions. This section outlines the design and development of the user interface for the selected SER system, ensuring seamless integration with the backend.

This research also focused on facilitating interaction between the developed real-world SER system and educators, both with and without hearing impairments. The goal was to verify the system's ability to detect emotions in real time accurately. Additionally, the study aimed to investigate educators' perceptions of the SER system's efficiency, its effectiveness in recognizing student emotions, and how educators respond to the system's output, among other factors. To achieve this, the SER system was designed to accurately display the detected student emotions in real-time. A user-friendly web interface was developed as the front end for the SER system. The interface was intentionally kept simple and intuitive, prioritizing the detection, display, and impact of the detected emotions over complexity. The web page interface of the SER application is shown in Figure 3.12.



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

Click the **'RECORD'** button below to begin recording students' feedback as they speak into the microphone.
Please ensure they speak for at least **5 seconds**.
Click the **'STOP'** button once the student finishes speaking.

Click the **'SHOW EMOTION'** button after you have stopped the recording.
Wait a few mins for the student emotion to be detected and displayed.

Click below on the different emotion links to get an idea of how students can express their emotions in the feedback.
Happy Neutral Sad Fear Angry

Figure 3.12: Web interface for the SER application

The SER neural network operated in the backend, while live audio (student verbal feedback) was received through the web interface, which served as the front end. As shown in Figure 3.11, the web interface offered features for capturing student audio feedback and displaying the detected emotion. In developing the GUI for the SER system, Nielsen's Usability Heuristics recognized guidelines for creating user-friendly and user-centred designs [63] were applied. These principles guided the development process to ensure a seamless and intuitive user experience. Table 3.3 lists the user interface components of the SER system along with their corresponding functionalities.

Table 3.3: User Interface components and their functionality

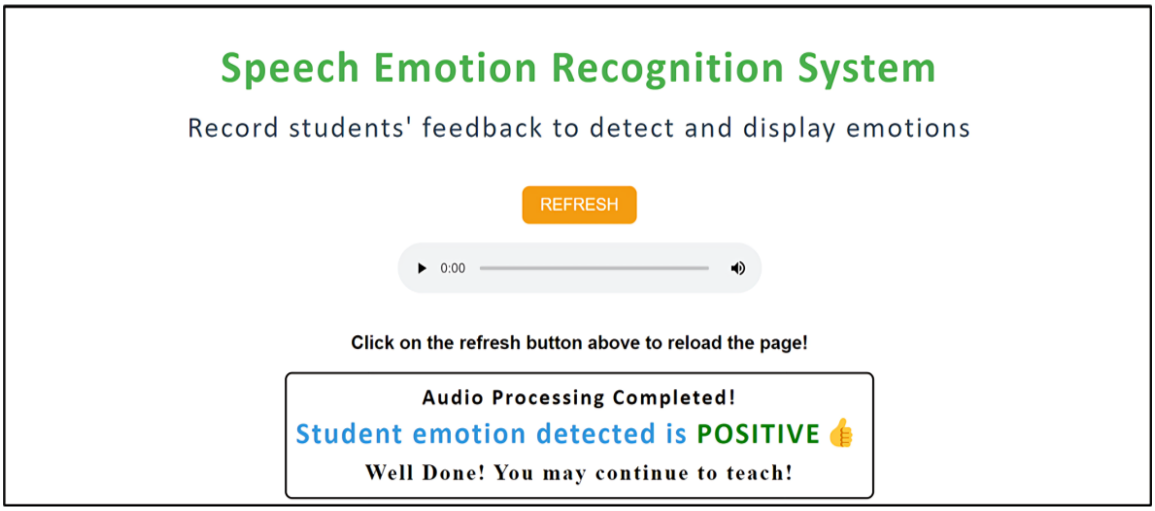
SNo.	Feature	HTML element type	Functionality
1.	Record	Button	This button initiates the recording process. When the student is ready to provide verbal feedback, the educator clicks the 'RECORD' button to capture the student's speech.
2.	Stop	Button	Once the student has finished speaking, the educator clicks on the 'STOP' button to end the recording. This ensures that the system accurately captures the complete verbal feedback.
3.	Refresh	Button	The 'REFRESH' button reloads the page. This can reset the interface and clear any previous recordings or displayed information, providing a fresh start for new recordings.
4.	Play	Audio	This control allows educators to repeat the recorded audio as often as needed. It enables them to review the student's verbal feedback.
5.	Show emotion	Button	After recording, the educator can click the 'SHOW EMOTION' button to see the displayed student's emotions based

			on their verbal feedback. This feature processes the recording and provides an emotional assessment from the backend.

As seen in Figure 3.12, a simple interface was created with three buttons, ‘RECORD’, ‘STOP’ and ‘REFRESH’, meant to start the recording, stop the recording and reload the web page, respectively. The functionality of each component of the interface is explained in Table 3.3. The ‘SHOW EMOTION’ button is meant to be clicked after the student’s verbal feedback has been captured to see the emotion discerned by the system. These features collectively enhance the functionality of the speech emotion recognition system, making it easy to record, review, and display students’ emotions discerned from their verbal feedback.

In designing the GUI for the SER system, Nielsen’s 10 usability heuristics were implemented to ensure a user-friendly and effective experience for late-deafened educators. They are discussed below:

1. The principle of visibility and system state was prioritized by ensuring the application has a visible title page and clearly defined options (like record, stop and such). The system status is constantly communicated, ensuring educators know what the system is doing, whether detecting emotions from student feedback or displaying the results. All actions are directly visualized.
2. The connection between the system and the real world was considered by designing icons and language familiar to the users. Information is presented logically, and icons are chosen to resemble everyday objects, making the system more intuitive. Each icon performs the expected action, and phrases are familiar to the educator, ensuring clarity, as seen in Figures 3.13 and 3.14. Figure 3.13 shows the display of the message as “Student emotion detected is POSITIVE” when the SER system detects the positive emotion of the student, whereas Figure 3.14 shows the display of the message as “Student emotion detected is NEGATIVE” when the positive emotion of student is detected.



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

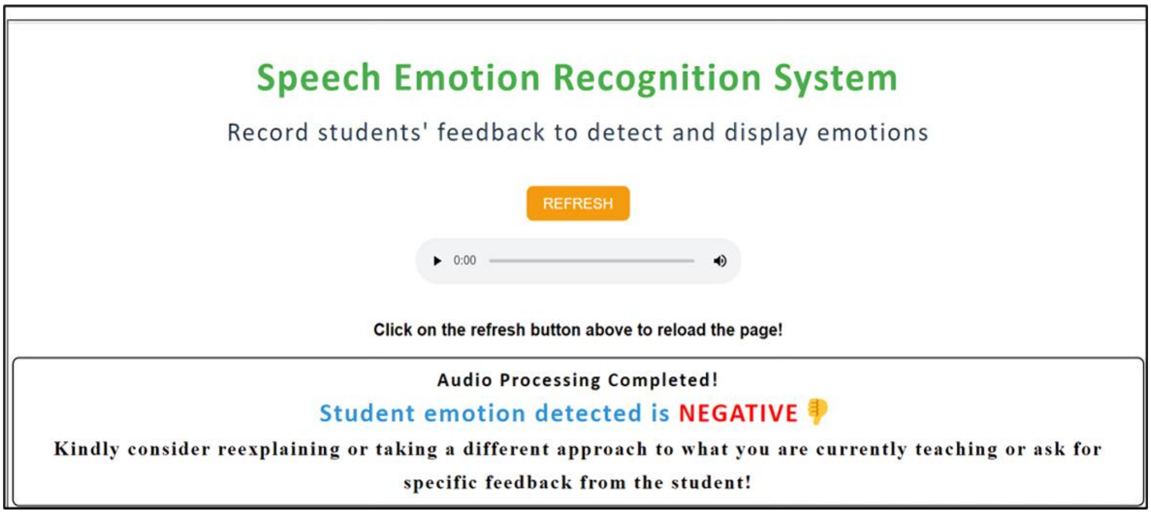
Click on the refresh button above to reload the page!

Audio Processing Completed!

Student emotion detected is POSITIVE 👍

Well Done! You may continue to teach!

Figure 3.13. Detection and display of positive emotion via the web interface



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

Click on the refresh button above to reload the page!

Audio Processing Completed!

Student emotion detected is NEGATIVE 👎

Kindly consider reexplaining or taking a different approach to what you are currently teaching or ask for specific feedback from the student!

Figure 3.14. Detection and display of negative emotion via the web interface

3. Regarding user control and freedom, the interface has only one web page, the home page (Figure 3.12). After every display of the detected student emotion, the educator can click the ‘refresh’ button to load the home page. The ability to quickly revert to previous states enhances the user experience.
4. Consistency and standards were maintained by labels having the same meaning across the interface, and information was displayed consistently across pages. Standard navigation elements, such as buttons, are used to avoid confusion.
5. To support recognition rather than memory, the interface is designed to be intuitive, with information organized logically.
6. For flexibility and efficiency of use, educators are provided with easy-to-follow visual cues and accessible interfaces, making the GUI efficient.

7. In case of ambiguous input received, the system will display an error message stating "Error processing the audio" and prompt the user to re-record the input.
8. For error prevention, the system minimizes errors through clear instructions provided on the interface, guiding users to ensure proper input and system usage.
9. The interface follows aesthetic and minimalist design principles, with concise, accurate, and well-organized information.
10. Help and documentation were made available via the guidelines provided to the educators before the usability testing.

In addition, the system used fonts of adequate size and high contrast to the background colour images, which were carefully chosen to ensure readability. The autonomy of the user is prioritized, with system statuses updated in real time. Educators were empowered to make personalized decisions based on the emotions displayed by students' vocal feedback. By focusing on Nielsen's heuristics principles, the GUI for the SER system is designed to be efficient, intuitive, and accessible, ensuring that the late-deafened can adapt to the system, which will seamlessly support their teaching.

This concludes the section on the methodology for the 'Design and Development of the User Interface', which is stage 3 of the process of the HCI research used for SER, as shown in Figure 3.2, has been completed. The methodology for evaluating the usability and user experience of the developed SER system in supporting late-deafened educators during online teaching is detailed in the next section.

3.6 Evaluation of the SER system

Section 3.3 describes the methodology for conducting the preliminary study for this research, followed by the methodology for building the SER system in Section 3.4. This section focuses on the methodology for conducting the usability study to evaluate the effectiveness, efficiency, and perceived impact of the SER system. It outlines the approach to assess usability and user experience of the SER system that can support late-deafened educators during online teaching. Figure 3.2(referenced in Section 3.3) illustrates the stages of the HCI research process, adapted for developing the SER system using a hybrid SDLC model principles from both the Incremental and Iterative approaches. Stage 4 involves evaluating the GUI-integrated SER system in real-time with late-deafened educators and those without hearing impairments. This evaluation measures the system's effectiveness, efficiency, and perceived impact. This section details the methodology employed for the usability study, emphasizing its role in evaluating the SER system's real-world application.

3.6.1 Usability and User Experience (UX) Evaluation

As per Lewis et al. [170], "Usability and User Experience (UX) are important concepts in the design and evaluation of products or systems intended for human use". The usability of a system refers to the ease with which a user can interact with it, with minimum effort, ensuring it is efficient, effective, and satisfying. It includes user interface design, accessibility, and overall user experience [109]. As outlined in ISO 9241-11 [145, 171], three dimensions are identified: effectiveness, which gauges the extent to which users accomplish their objectives with the system; efficiency, which scrutinizes the resources employed to fulfil user objectives; and satisfaction, representing the user's perception of the system's usability [172, 173, 174]. Usability testing is the process or a specific method to evaluate and improve the system's usability by observing real users interacting [170, 172]. This testing aims to identify any usability issues, gather qualitative and quantitative data, and take feedback leading to recommendations for enhancements of the system [175]. As per Lazar et al. [176], in their book *Research methods in human-computer interaction*, "Methods utilized as part of usability testing include surveys to measure user satisfaction".

In the context of the proposed SER system, the usability testing was conducted with educators with and without hearing impairments to determine if the system effectively addressed their challenges in discerning student emotions during online classes. This process involved identifying the educators fulfilling certain criteria, conducting system testing, and measuring usability and user experience through questionnaires. This was followed by an analysis to assess the impact of the SER system on effective engagement with students and to identify any further suggestions from educators.

Aims of Evaluation of the SER system:

1) To assess the effectiveness and efficiency of the developed SER system.
2) To obtain feedback on the perceived impact of the SER system on educators with and without hearing impairment [2].

[2] *Wherever the term educators with hearing impairment is used, it refers to late-deafened educators, defined in Section 2.6.4.*

In pursuit of achieving these aims of measuring usability and evaluating the user experience for the SER system, the research methodology and methods adopted to achieve them have been presented in the next section.

3.6.2 Research Methods for the Evaluation of the SER system

The complete usability study design is thoroughly examined in this section. Quantitative research collects and uses numerical values to describe a phenomenon. It analyses the collected data with mathematical methods. In contrast, qualitative research collects non-numerical data, such as textual data, to gain deep insights into a specific phenomenon. It analyses the collective data with interpretive methods to discover new insights. Quantitative research focuses on objectivity and measurable outcomes, whereas qualitative research focuses on context and the subjective experiences of individuals [177].

Both qualitative and quantitative methods were used for the evaluation of the SER system. The idea behind using these methodologies was to gain a comprehensive understanding of the usability and user experience of the SER system involving late-deafened educators. The target respondents, educators with and without hearing impairment, were first selected to achieve this. They were asked to test the system after providing guidelines regarding the system and system testing. After the testing was done, feedback was requested by asking them to complete the online survey (For the questionnaire, refer to Appendix 7).

The survey was administered to the educators to collect structured quantitative data on various aspects, such as demographic information (age, gender, teaching experience and hearing condition). It also employed qualitative methods to capture in-depth insights from educators on assessing students' emotions to ascertain their engagement in online classes. The survey consisted of closed and open-ended questions designed to gather feedback on the effectiveness and efficiency and to understand the user experience after usability testing. Participants were asked to express their views on the SER system's impact on teaching and to provide suggestions for possible improvements. A 5-point Likert scale and single-select questions were framed to quantify the user responses. For qualitative data, open-ended questions were included to allow participants to provide detailed responses in text boxes.

For the evaluation of the SER system, the quantitative data were analysed using descriptive statistics, specifically through frequency counts. Frequency refers to the count or the number of times a particular value or a category occurs in a dataset. It provides information about data distribution by indicating how frequently each value or category appears [140, 141, 143]. This distribution can be represented in tabular or any graphical form. Qualitative data were analysed using content analysis and sentiment analysis. Content analysis is a widely recognized method for data analysis of textual data [178]. It involves systematically examining the responses to identify and interpret patterns or themes. It allows researchers to quantify the presence of certain words, themes, or concepts in qualitative data and provides a systematic approach to understanding open-ended responses [179, 180]. Sentiment analysis, also known as opinion mining, focuses on understanding opinions. It is a branch of Natural Language Processing (NLP), part of computer science and artificial intelligence, dealing specifically with the interaction between human language and computers [181]. It involves collecting and examining users' views or opinions about a product, subject, or system [182, 183]. Sentiment analysis encompasses both subjectivity and polarity. It is a technique used to examine the polarity of the text to assess whether data expresses positive, negative, or neutral sentiments [184]. Further, a lexicon-based analysis was conducted. Lexicons are collections of tokens, where each token is assigned, a predefined score indicating its neutral, positive, or negative nature. These scores typically range from +1 for positive, 0 for neutral, and -1 for negative. This approach provides a numerical representation of sentiment useful for aggregating and analysing feedback [182, 184, 185]. Table 3.4 gives an overview of the tools used for the usability testing.

Table 3.4: Tools used for usability testing

Research parameters	**Usability Testing categorization**	**Purpose**	**Research Method/s used**	**Test tools used for analysis**
Demographics	User Data	To select suitable respondents	Quantitative	Descriptive analysis
Usability	Effectiveness and Efficiency of the SER system	To understand if the SER system is easy to use and intuitive.	Quantitative And Qualitative	Descriptive analysis, Content analysis and Sentiment analysis
User experience	Impact of SER on teaching	To get user feedback on the value, satisfaction and impact of SER system.	Quantitative And Qualitative	Descriptive analysis, Content analysis and Sentiment analysis

In the next section, complete details on the research methodology adopted for the usability study have been presented.

3.6.3 Data and data collection

Usability testing was conducted with selected participants to evaluate the usability and user experience of the SER system. Data was collected through a survey/questionnaire following the testing session. During the testing session, notes were also taken by the author as an observer. First, the data collection process is examined via the survey tool, followed by other information related to the survey.

Formulating the Survey Questions, Designing the Questionnaire, Obtaining Ethical Clearance, and Pilot Testing

1. **Formulating the Survey questions**:

As seen in Section 3.3.1, Surveys enable data collection without the researcher's interference, offering valuable insights from a sample group on key issues [138, 139]. A twenty-five-question questionnaire (Appendix 4) was constructed. The initial questions were single-option questions, defined to gather demographic information about the respondents. These included details about the respondents' gender, age, length, area of teaching, the level of students being taught, their hearing condition, and their usage of online systems. Subsequently, questions were formulated to ascertain the respondents' hearing condition, determining if they could hear verbal feedback from students during online classes and discern underlying emotions such as positive or negative. These questions were both closed and open-ended. A set of 5.0 Likert scale questions were used for some of the closed-ended questions. Likert-type questions are a common format used in surveys. They help measure participants' responses to questions and produce ordinal data [138]. One such set of questions was used to reveal whether the user understood the purpose of the SER system. A combination of closed and open-ended questions was used to ascertain the effectiveness, efficiency, and impact of the SER system. To evaluate effectiveness, several 5-point Likert scale questions focused on the interface, the visual representation of detected student emotions, and system features. Open-ended quantitative questions were included to gather comprehensive feedback on the efficiency of the SER system. These questions aimed to determine if the system accurately detected and displayed real-time student emotions, identify misinterpreted emotions, and gather suggestions for enhancing system efficiency and effectiveness.

Finally, closed and open-ended questions were included to gather feedback on the impact and improvement of the developed SER system in online education. These questions explored how detected negative feedback could influence teaching approaches, how positive student emotions could reinforce confidence in ongoing classes, and whether respondents would recommend using the system during online classes. The final questions assessed whether student emotion recognition through this system would influence perceptions of student engagement or understanding during online classes and invited additional comments or feedback. Refer to Appendix 4 for the questionnaire.

2. Designing the Questionnaire:

After formulating the questions, it was essential to structure the questions constructed above and format the survey into a coherent and organised questionnaire. An online survey was the research method used to gather feedback on various aspects of the system's usability and user experience from the respondents. The structured questionnaire was the research tool used to collect data from the survey.

The questionnaire was structured into five sections to gather comprehensive information from the respondents after completing the usability testing of the SER system. During the questionnaire development, each question discussed above aligned with a specific variable and contributed to answering the usability study's aims.

- Section A focused on the respondents' demographics and profiles,
- Section B focused on questions related to the respondents' hearing condition,
- Section C was dedicated to questions to understand how the respondent uses online systems and the significance of feedback during online sessions,
- Section D focused on questions related to the effectiveness and efficiency (or evaluation) of the developed SER system.
- Section E questions related to feedback on the impact and improvement of the developed SER system in online education.

3. Obtaining Ethical Clearance:

Approval from an ethics committee is mandatory to ensure the study complies with the required ethical standards. Since this work deals with the survey's human subjects and social needs groups, the questionnaire received the required ethical clearance (refer to Appendix 3).

4. Pilot Testing:

Pilot testing was conducted with a small group of individuals to evaluate the survey's clarity, relevance, and effectiveness in capturing the intended aims of the usability study [138]. Participants were selected from the target population to ensure meaningful feedback. The pilot testing addressed any issues in the survey questionnaire before the main study [142]. An academic and a corporate trainer were chosen to align with the requirement that pilot participants belong to the target population. The academic participant, a seasoned professional with over 30 years of service at a prestigious university, brought valuable insights from extensive teaching experience. The corporate trainer, with over 20 years of experience—10 of which were in academia before transitioning to a project manager and corporate trainer role in an IT company—contributed expertise in pedagogy and system design. Their feedback and suggestions were incorporated into the questionnaire before being distributed to the recruited participants. Additionally, the SER system was demonstrated to gather their perspectives on the interface's design and the appropriateness of the displayed messages following emotion detection. Specific comments regarding the user interface, such as font size, button background colour, and overall look and feel, were also integrated into the system before testing it with the recruited participants.

3.6.4 Participants

As per Goode and Hatt [140], a sample refers to a subset of individuals or participants selected from a larger population for research or testing purposes. Sampling is selecting these participants to draw inferences about the broader population [138, 140]. A participant is the research focus and part of a collection of individuals. Identifying the correct participants from the population is essential to ensure the generalizability of findings. Depending on the purpose and limitations of the study, there can be a few or many participants. Appropriate sampling techniques are essential for selecting representative samples from the identified population.

As seen in the preliminary study (Section 3.3.6), for system evaluation as well, purposive sampling was used to recruit participants based on specific criteria [140, 141, 143]. This method ensured that respondents had the necessary background to provide relevant answers. Participants were selected from a well-defined population of interest. Besides the capacity to participate in the designed study, they should also be willing to participate. The selection of participants for the study was based on teaching experience with a minimum of four years' experience in online teaching (every educator has taught online during the COVID-19 pandemic, UNSECO IESALC, 2020) [8]), teaching students at higher education or university and specifics about their hearing condition. The selection was purposeful and not arbitrary. The design of the sample frame ensured the required representation. This covered educators with or without hearing impairment.

Further, in the case of educators with hearing impairment, a snowballing method was also used, where connections were established with the participants based on referrals. Snowball sampling is a technique in which you approach individuals from the population to help identify participants who could be suitable for the ongoing study [138, 142, 143]. This approach was used as educators with specific problems of late-onset deafness were required. The snowballing method to recruit late-deafened educators has been outlined in Chapter 3, Section 3.3.6.

Finding and recruiting qualified participants (samples) is challenging. As Lazar et al. [144] suggest in their book, research involving users with disabilities typically accepts a sample size of 5–10 participants with a specific disability. This is because it may be hard to find participants with disabilities who meet all the inclusion criteria, such as employment, education, technical expertise, and such. Research focusing on users with disabilities often faces difficulty in finding participants who meet all requirements. As a result, studies in this area generally have smaller sample sizes than those examining users without disabilities. Also, a small number of participants with disabilities is sufficient to uncover usability issues related to them, as seen in studies [145, 146, 147]. 10 participants/respondents experiencing late-onset deafness, ranging from mild to profound, and 10 participants/respondents who were educators with normal hearing or without hearing impairment were recruited. As per purposive sampling, the type and number of participants are well balanced to get valid feedback as they had the required experience.

Additionally, 6 educators were selected as Pilot Users for testing the SER system in the pilot phase. System Testing by pilot users is treated very seriously and conducted exactly like the procedure planned for the experiment [142]. Testing the systems with pilot users is a preliminary trial where a small group of users or pilot users interact with the developed system to identify usability issues, gather initial feedback, and assess the overall user experience [186]. This helps refine the system's design and functionality before the broader implementation of usability testing with the actual respondents [186]. Out of the six, one of the pilot users was an educator with a hearing impairment. The selection of the educators as pilot users was done using purposive sampling and snowballing techniques. They evaluated the system and provided feedback via the online questionnaire.

3.6.5 Procedure for Conducting System Testing in Pilot and Main Phase

In this section, the procedure adopted for conducting the usability testing of the SER system for the pilot and the main phase is shown in Figure 3.14 (a) and 3.14 (b) respectively. Testing was divided into two phases: the Pilot phase and the Main phase. In the pilot phase, usability testing was conducted with the pilot users, the selected six educators, as discussed in Section 3.6.4. In the main phase, usability testing was conducted with selected final respondents who are educators with and without hearing impairment.

As mentioned above, the selection process of participants (educators with and without hearing impairment) for the SER system's pilot testing was the same as the respondents for the final testing. They came from similar backgrounds and were aware of real-world usage scenarios. They quickly adapted to the simulated online teaching environment with the SER system. Their interactions with the system provided valuable insights into how it would perform in practical settings. Based on their experience with the SER system via testing, they provide constructive feedback on possible improvements in system features, effectiveness and efficiency, and the system's perceived impact on adjusting teaching approaches. The system's quality and reliability can be improved by resolving these issues early.

1) Pilot Phase

As discussed in Section 3.6.4 above, six pilot users were selected to assess the SER system. 5 were educators without hearing impairments, and 1 was an educator with a hearing impairment. Each pilot user was also chosen through purposive sampling based on specific criteria like teaching experience, familiarity with online teaching, hearing condition and such. Each pilot user had to engage 2 of his/her students to participate in the testing. Before commencing the tests, comprehensive guidelines were shared with both educators and students. Guidelines had complete information about the SER system, testing procedures, and expectations during the usability testing session (refer to Appendices 5 and 6). Since the testing was conducted online remotely, a Zoom link was created and shared with the pilot user and their students (arranged by the educator). Before the commencement of the testing session, the presentation was given to the educators and their students. They were informed about the system's purpose, usage, expectations, and outcomes. A demonstration to explain the flow of the SER system was also provided.

The SER systems' website link was shared with the educator. The SER system's function was to capture students' verbal feedback and display the detected underlying emotion. The educator was briefed on how to use the SER system (website) with the ongoing Zoom call. Zoom was used as the teaching tool to create a simulated teaching environment, with the educator on one side of the online Zoom call and the students on the other side. Each learner/student had to provide real-time verbal feedback online during the Zoom call via their microphone. Essentially, four utterances with varying emotions were required. Students were asked to express themselves as happy, sad, scared, or angry while giving verbal feedback. This was necessary since it was a simulated, not an actual online class. Also, having the system detect and display the student's emotions four times, rather than just once, would help educators understand the system flow and its purpose more clearly. Each time the student was ready to give feedback via their microphone, the educator had to click the 'Record' button on the SER website. Once the student completed their verbal feedback, the educator needed to click on the 'Stop' button. This was how the students' verbal feedback was captured. Following this, the educator needed to click on the 'show emotion' button to see what emotion (positive or negative) the system had detected and displayed. This complete flow has been described in Chapter 4, Section 4.4. Once educators understood the flow, they tested the SER system with the students present during the system testing. After the testing, they gave feedback in an online questionnaire via a Google Form link (Refer Appendix 7). The author also documented all the observations in an Excel sheet as an observer. The pilot testing was conducted with 6 pilot users (educators). Every testing session was recorded. Based on the feedback captured in the questionnaire and observations in the Excel sheet, changes were made to the SER system before proceeding with the usability testing with the selected sample, respondents with and without hearing impairment.

2) Main Phase

As discussed in Section 3.6.4, a purposive sampling approach was used to recruit 20 educators for usability testing of the updated SER system. Out of the 20, 10 were educators with hearing impairment issues ranging from mild to profound. The remaining 10 were educators without hearing impairments. The usability testing was conducted exactly as it was done in the case of pilot users, as described in Phase 1. The author captured observations as an observer in an Excel sheet. Every testing session was recorded.

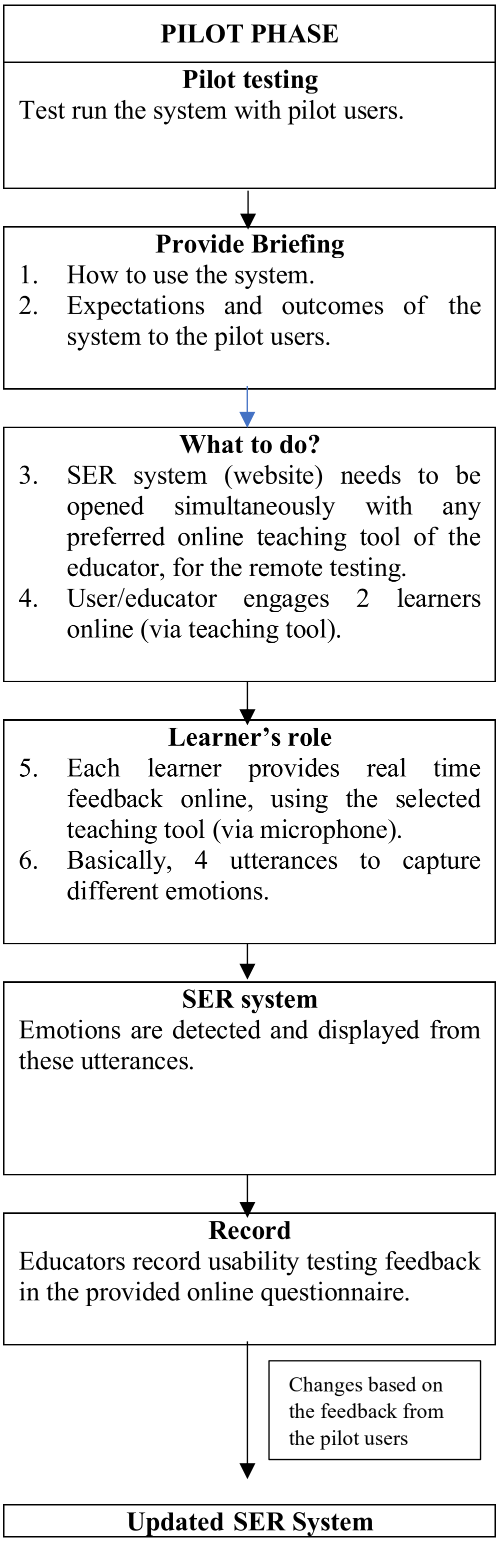

PILOT PHASE
Pilot testing
Test run the system with pilot users.
Provide Briefing
1. How to use the system.
2. Expectations and outcomes of the system to the pilot users.
What to do?
3. SER system (website) needs to be opened simultaneously with any preferred online teaching tool of the educator, for the remote testing.
4. User/educator engages 2 learners online (via teaching tool).
Learner's role
5. Each learner provides real time feedback online, using the selected teaching tool (via microphone).
6. Basically, 4 utterances to capture different emotions.
SER system
Emotions are detected and displayed from these utterances.
Record
Educators record usability testing feedback in the provided online questionnaire.
Changes based on the feedback from the pilot users
Updated SER System

Figure 3.15 (a). Flow diagram of the procedure for conducting usability testing of the SER system (Pilot Phase)

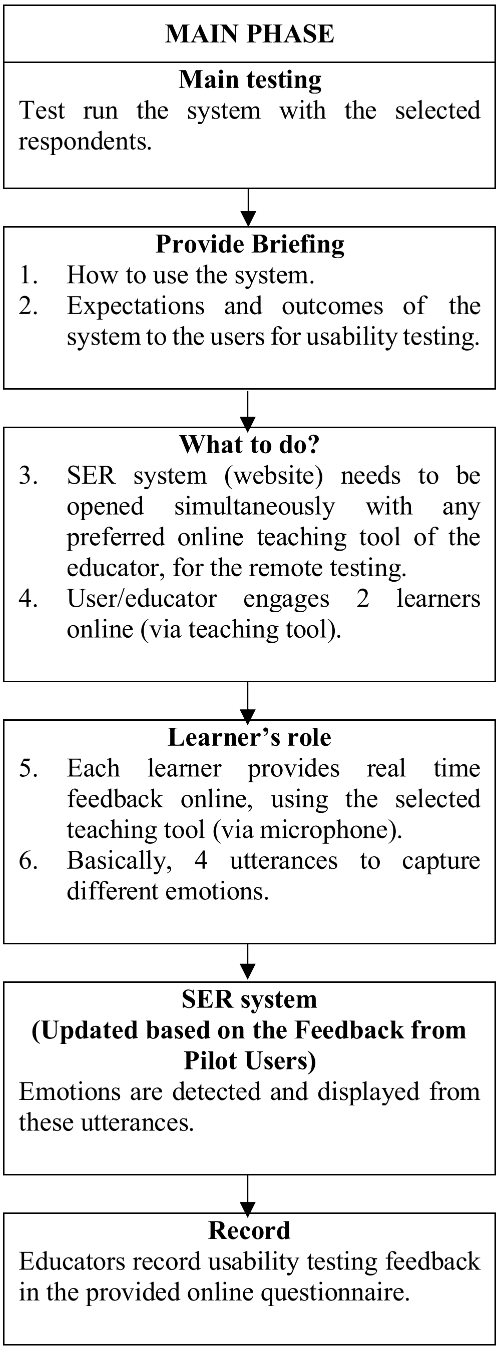

MAIN PHASE
Main testing
Test run the system with the selected respondents.
Provide Briefing
1. How to use the system.
2. Expectations and outcomes of the system to the users for usability testing.
What to do?
3. SER system (website) needs to be opened simultaneously with any preferred online teaching tool of the educator, for the remote testing.
4. User/educator engages 2 learners online (via teaching tool).
Learner's role
5. Each learner provides real time feedback online, using the selected teaching tool (via microphone).
6. Basically, 4 utterances to capture different emotions.
SER system
(Updated based on the Feedback from Pilot Users)
Emotions are detected and displayed from these utterances.
Record
Educators record usability testing feedback in the provided online questionnaire.

Figure 3.15 (b). Flow diagram of the procedure for conducting usability testing of the SER system (Main Phase)

3.6.6 Mimicking Emotions by Students/Learners

The previous section, Section 3.6.5, stated that students needed to mimic emotions. Generally, during an actual online class, educators keep taking feedback intermittently during the class. This is to gauge the students' comprehension of the ongoing class. Among many, one of the ways to take feedback is to select a student randomly and ask them to give verbal feedback via their microphone. The student's response is generally based on the educators' explanation of the topic at hand. They will express their emotions naturally through verbal responses, providing valuable insight into their understanding. If they grasp what the educator is teaching, their voice will naturally exude happiness or remain neutral. Conversely, if they're struggling to understand, their tone may convey sadness or frustration. These emotions will come out organically in an online class.

However, as discussed in Section 3.6.4, usability testing was conducted with a selected group of educators. A simulated teaching environment via a Zoom call was created, and the selected teaching tool was used for the purpose of testing. The educator and the two students engaged by them participated in the call. Since it was a simulated online class, students' emotions when giving verbal feedback might not be expressed naturally. Hence, they were asked to mimic the emotions. They had to give verbal feedback four times. Two times sounding positive (happy or neutral emotion) and twice negative (sad, angry or fearful emotion). Guidelines were given to them on how emotions can be enacted. The purpose of asking the student to give feedback with enacted emotions more than once was to allow the educator to understand how the system works, how the detected emotions are being displayed, and also to observe if detected emotions were being displayed accurately.

Guidelines for expressing these emotions were provided based on references from [187, 188, 189]. Among other parameters, pitch was important in conveying emotion. Sounds with greater energy at higher frequencies produce a higher pitch, and vice versa [189]. To express happiness, students had to slightly increase pitch, speak in a lively and upbeat tone, maintain a faster pace, and incorporate laughter or smiles. A moderate pitch, steady pace, and clear, composed tone are recommended for neutral or calm emotions. To convey sadness or the feeling of being upset, the student had to lower the pitch, use a softer tone, allow pauses, and adopt a more subdued delivery. To express fear, speaking with a higher pitch, increasing the pace with moments of hesitation, and using a shaky or trembling tone can convey urgency. To show anger, student had to raise the pitch, speak louder and more forcefully, slightly increase the pace, and use assertive, direct language. Based on these instructions, students attempted to enact the emotions while providing verbal feedback.

3.7 Consideration of Real-World Online Teaching Conditions

Although the system was evaluated in a simulated environment, the setup closely mirrored real online classroom conditions, with educators and students interacting live from separate locations via the video conferencing platform Zoom. The methodology accounted for several real-world constraints typically encountered during online teaching. Audio data was primarily collected using the university network during academic hours, though home internet was also used occasionally. This reflected the range of network conditions found in remote learning. Participants used various microphone types, including laptop-built-ins and external headsets. Additionally, background noise and occasional overlapping speech were naturally present in the recordings.

To approximate actual online teaching environments, real-world variables were incorporated into the data collection and testing process, as summarized in Table 3.5.

Table 3.5: Real-World Conditions Simulated During SER System Evaluation

Factor	Details
Network Conditions	The simulated online class was conducted over both university and home internet connections. Most recordings occurred during regular university hours. This exposed the system to realistic bandwidth, latency, and stability variations.
Microphone Types	Participants used built-in laptop microphones, wired headsets, and external USB microphones. This introduced natural variation in audio quality, helping to assess the system's responsiveness across hardware setups.
Background Noise and Overlapping Speech	Background sounds (e.g., fans, typing, surrounding voices) and occasional overlapping speech were intentionally retained to reflect realistic classroom audio conditions.
Recording Platforms	The video conferencing platform Zoom, commonly used in virtual classrooms, was used.

These methodological choices ensured that the SER system was evaluated under conditions closely aligned with its intended real-world use—online learning environments involving diverse setups, variable audio quality, and natural speech variability.

3.8 Ethical Consideration

Ethical approval for this study was obtained prior to data collection, and all research procedures involving human participants were conducted in accordance with the University of Nottingham Code of Research Conduct and Research Ethics. The study ensured voluntary participation, informed consent, data protection, and anonymity for all participants.

Student Participants

All student participants were above 18 years of age as per the inclusion criteria. They took part in the study voluntarily. Before participating in the system testing, they were provided with a Participant Information Sheet and Consent Form via email and through the Google Forms platform. These documents clearly outlined the purpose of the study, the students' role in providing verbal feedback during the usability testing, their right to withdraw at any time, and how their data would be handled.

Digital consent was obtained prior to their involvement. No personally identifiable information was collected from students, and all responses were anonymized. Verbal feedback from students was not recorded or stored in any database; it was used solely for emotion detection by the SER system in real-time.

The handling of student data complied with the University of Nottingham's Royal Charter, the UK General Data Protection Regulation (GDPR), the Malaysian Personal Data Protection Act 2010, and the University's Code of Research Conduct and Research Ethics. Although students were not the primary focus of the research, their role in simulating an online teaching environment was essential. Full ethical safeguards were applied to ensure their rights and privacy were protected. The informed consent form provided adequate information on the respondent's rights on anonymity, confidentiality, and the right of withdrawal.

Educator Participants

Educators—both with and without hearing impairments—were the primary participants in the usability testing. Each educator received a detailed Participant Information Sheet and Consent Form outlining the study's purpose, their role in evaluating the SER system, the duration of participation, and their right to withdraw at any time. These documents also explained that participation was voluntary and that there were no risks or payments involved.

Consent was obtained digitally before participation. No identifiable information was collected, and all responses were anonymized and stored securely. They were further informed that the data collected via survey will be stored and analysed using Microsoft Excel. However, this was not the personal or sensitive data. It was only the perceptual data, which was presented as an aggregate and not in relation to specific individuals. The informed consent form provided adequate information on the respondent's rights on anonymity, confidentiality, and the right of withdrawal.

The collection and processing of data from the educator followed the same legal and institutional frameworks mentioned earlier. The research was approved by the Science and Engineering Research Ethics Committee (SEREC), University of Nottingham Malaysia, under Application Identification Number: AV200124.

3.9 Development Environment

Here, the programming environment is described. Python is an open-source language under an OSI-approved license, which is free to use and distribute, even for commercial purposes. Python's flexibility across platforms and its extensive libraries makes it a popular choice for application development. Its main advantages include simpler code, dynamic data handling, and a wide range of libraries suited for modern programming [190, 191, 192]. For this application, Pandas, Librosa, Scikit-Learn and TensorFlow as they are essential for the development requirements.

1. Pandas Library

Pandas is an open-source data analysis and manipulation library built for Python. It provides easy-to-use data structures, like DataFrames, and functions for efficiently handling structured data, such as numerical tables and time series. Pandas is widely used for tasks such as cleaning, transforming, and analysing data, making it a powerful tool for everything from simple data processing to complex statistical analyses [193].

2. Librosa Library

Librosa is an open-source Python library designed for audio analysis. Librosa gives much functionality, like loading the audio files, generating spectrograms, visualizing speech signals, and extracting desired speech features. With Librosa, one can easily integrate with other Python libraries, and hence, it's very popular in processing applications like speech recognition, speech emotion recognition, music classification and such. It is heavily used in sound processing and machine-learning applications involving audio data [150, 194].

3. Scikit-Learn Library

Scikit-learn, often referred to as sklearn, is an extensively used open-source machine-learning library for Python. It is popular for developing machine learning models as it provides simple and efficient data mining and analysis tools. It supports supervised and unsupervised learning algorithms, such as classification, regression, clustering, and dimensionality reduction. It also includes tools for model evaluation, data preprocessing, and cross-validation. Based on other scientific Python libraries like NumPy, SciPy, and Matplotlib, scikit-learn is well-suited for many applications [191, 195].

4. TensorFlow

TensorFlow is an open-source deep learning framework developed by Google that is designed to build and deploy an end-to-end platform for machine learning. It provides a flexible platform for constructing and training neural networks. It enables users to work with various tasks, from image recognition and natural language processing to time-series forecasting and reinforcement learning. TensorFlow supports low-level operations, giving developers granular control over model architecture and high-level APIs like Keras for easier, more intuitive model building. It can run on various hardware setups, including CPUs, GPUs, and TPUs, making it highly scalable for large-scale machine-learning projects [191, 196, 197].

The Integrated Development and Learning Environment (IDLE) chosen for writing the code is Google Colab (Collaboratory). It is a cloud-based platform that allows users to write and execute Python code in an interactive environment, similar to Jupyter Notebooks. It provides free access to computational resources, including Central Processing Units (CPUs), Graphics Processing Units (GPUs), and Tensor Processing Units (TPUs), making it especially popular for machine learning and data science projects that require high-performance computing. Google Colab supports a wide range of Python libraries and is integrated with Google Drive, allowing users to save and share their notebooks easily. Since it runs in the cloud, no setup or installation is required, and it enables seamless collaboration by allowing multiple users to work on the same notebook simultaneously.

3.10 Conclusion

This concludes the methodology chapter, where the preliminary study methodologies were described to understand the needs and perspectives of late-deafened educators. This was followed by methodologies for building the SER System techniques of three selected hybrid models that were compared to identify the one which could give the highest speech emotion recognition accuracy. Subsequently, the methodology for interface development was covered, focusing on the methods for transforming the selected SER system into a functional, real-time application capable of detecting and displaying emotions instantly. It detailed the design and development of the user interface for the SER system, ensuring seamless integration with the backend. Lastly, the methodology for the usability study was described, which outlined the methods used to evaluate the usability and user experience of the developed SER system in supporting late-deafened educators during online teaching.

In the next chapter, Chapter 4, the results from the preliminary study will be discussed and analysed first, followed by the results from the three selected hybrid architectures: Model A (hybrid averaging), Model B (hybrid parallel), and Model C (hybrid sequential). The experiments demonstrated the superiority of Model A in achieving the highest accuracy. Consequently, Model A was incorporated into a real-time application with a suitable graphical user interface. Finally, the results from user testing and evaluation of the developed SER system will be presented, analysed, and discussed.

Chapter 4

4. Results

4.1 Introduction

This chapter presents the results and analysis of the study, offering insights into each phase of the research. It begins with the findings from the preliminary study, which aimed to understand the needs and perspectives of late-deafened educators. This is followed by the experimental results from the three hybrid models, highlighting their performance and identifying the best-performing model. Subsequently, the chapter discusses the outcomes of integrating the selected (best-performing) model into a graphical user interface (GUI) for real-time emotion detection. Finally, the results of evaluating the effectiveness and user experience of the developed SER system in an online teaching environment are presented.

The results of the study are presented and analysed in the following sequence:

1. **Preliminary Study Results**: This section discusses the findings from the preliminary study, focusing on the needs and perspectives of late-deafened educators.
2. **Experimental Results and Performance Comparison of Three Selected Hybrid Models**: This section presents the experiments conducted using three hybrid models, highlighting and comparing their performance and identifying the model with the highest speech emotion recognition accuracy.
3. **GUI-Integrated SER System**: This section covers the results of integrating the selected SER model with the graphical user interface, emphasizing its functionality as a real-time emotion detection application.
4. **Evaluation of the SER system**: This section discusses the outcomes of evaluating the developed SER system for its effectiveness, efficiency and perceived impact on late-deafened educators for effective engagement in online education.

In the next sections, the results and analysis of the preliminary study are presented first, followed by the experimental results from the three hybrid models. Subsequently, the outcomes of integrating the best-performing model into a GUI for real-time emotion detection are discussed. Finally, the results of the usability study are presented.

4.2 Results and Analysis of the Preliminary Study

In Section 3.3, the preliminary study's methodology was described in detail. The study explored the potential benefits of using an SER system for late-deafened educators in online classes. A survey targeted educators with and without hearing impairments to achieve this goal. The final step of the preliminary user requirements study, as shown in Figure 3.3 (Section 3.3), involved the analysis of survey results. Hence, this section presents the results and analysis to determine whether the aims of the preliminary study were achieved. The aims of the preliminary study, as outlined in the methodology chapter, are as follows:

1. Investigate the perspectives of educators (with and without hearing impairments) on the necessity of detecting student emotions to enhance teaching outcomes.
2. Understand the challenges faced by educators (with and without hearing impairments) in discerning student emotions during online teaching.
3. Identify the benefits of implementing an emotion-detection system (SER) for detecting student emotions.

The next section first presents the demographic information.

4.2.1 Demographic Information

The survey conducted for the preliminary study received responses from 33 respondents who met the targeted criteria under purposive sampling. Of the 33, 10 were late-deafened educators, which was the targeted number, while the remaining 23 participants were educators without hearing impairment.

As seen in Table 4.1, among the 33 respondents, 42% were male and 58% were female. Regarding age distribution, 21% were between 30 and 40 years old, 36% between 41 and 50, and another 36% between 51 and 60. Around 7% were above 60. The snowballing sampling method was instrumental in recruiting educators with hearing impairment. The late-deafened educators were predominantly in the age group of 40-60 years. This made them highly relevant to this study's focus. This research focuses on the challenges late-deafened educators face in discerning student emotions in online classes. Additionally, 33% of the respondents had more than 20 years of teaching experience, demonstrating high expertise. Using snowballing and purposive sampling, the study successfully targeted experienced educators, particularly in online teaching. Further, 31% of the sample had 11-15 years of experience, 18% had 16-20 years, and another 18% had 5-10 years of teaching experience.

Regarding discipline, 64% of educators were from STEM (Science, Technology, Engineering, and Mathematics) fields, 24% from social sciences and business, and 11% from other areas. 68% of the respondents taught at the university level, 12% at high school or pre-university levels, and 18% were involved in corporate training or primary education.

Regarding formal training in education, 91% had received formal training, while 9% had not. Since the hearing condition of the respondent was crucial for this study, a section of the survey focused on this aspect. 70% of respondents reported normal hearing, which means these were educators without hearing impairment. 30% (or 10 educators) had some form of hearing impairment. Notably, the target of recruiting 10 late-deafened educators was met. Only 15% of the educators reported difficulty hearing students' vocal feedback in face-to-face classrooms, while 85% could hear adequately. Educators with mild to moderate hearing impairments could hear student feedback in the class as they were using hearing aids.

In terms of online teaching experience, 100% of respondents had taught online. Most of the educators used platforms such as MS Teams or Zoom. This provided essential insights for this preliminary study, which focused on educators with online teaching experience.

Table 4.1: Respondent's profile, demographic, and hearing condition

CRITERIA	PERCENTAGE (%)
Respondent's profile, demographics, and hearing condition	
Gender Details	
Male	42%
Female	58%
Age Group	
30 to 40 years	21%
41 to 50 years	36%
51 to 60 years	36%
Above 60 years	7%
Teaching experience	
More than 20 years	33%
16 to 20 years	18%
11 to 15 years	31%
5 to 10 years	18%
Area of teaching	
STEM (Science, Technology, Engineering, and Math)	64%
Social sciences and business	24%
Miscellaneous fields	12%

Students' level taught by the respondent	
University level	70%
High school or pre-university education	15%
Corporate training, primary schools, or other similar settings	15%
Formal training in education received by the respondent	
Formal training in education	91%
Not undergone any formal training	9%
Respondent's hearing condition	
Current hearing condition	
Normal hearing (without hearing imparment)	70%
Some form of hearing difficulty or impairment	30%
Respondent's hearing condition when taking feedback in class	
Could hear clearly	85%
Could not hear clearly or hear at all	15%
Respondent's frequency in conducting online classes in the last two years	
Have conducted classes online	100%

Quantitative and qualitative methods [198] were employed in this preliminary study to gather insights from educators, both with and without hearing impairments, on how they receive student feedback and discern emotions in that feedback during online sessions. In the next section, the analysis of the quantitative data is presented first.

4.2.2 Student Feedback and their Emotion Recognition in Online Systems - Analysis of the Quantitative Data

Statistics is a set of procedures defined to systematically gather, measure, classify, compute, describe, synthesize, analyse, and interpret quantitatively collected data [199]. It can be classified into two types – descriptive and inferential statistics. Numerical and graphical ways to summarize datasets can be parked under descriptive statistics. On the other hand, inferential statistics is drawing inferences or conclusions about a larger population based on data collected from a sample [199, 200]. For the survey data collected for the preliminary study, descriptive statistics using frequency count were used to analyse quantitative data. Frequency refers to the count or the number of times a particular value or a category occurs in a dataset. It provides information about data distribution by indicating how frequently each value or category appears [201]. This distribution can be represented in tabular or any graphical form.

From Figures 4.1 to 4.7 illustrate the respondents' perspectives on the importance of student feedback and recognizing students' emotions, which are crucial for adjusting teaching strategies in physical (face-to-face) and online classes. The questionnaire included some general questions about face-to-face classes, followed by more detailed questions focused on online classes, as this study primarily centres on online sessions. Responses were collected from educators with and without hearing impairments. The quantitative data analysis, using descriptive frequency counts, is presented below.

The survey received responses from 33 educators, both with and without hearing impairment. The questionnaire included closed-ended questions to explore the benefits of gauging emotions in student feedback during an ongoing class. Additionally, it aimed to understand how educators can leverage these detected emotions to improve their teaching approaches. Responding to the first survey question, all educators unanimously agreed to intermittently seeking student feedback about the ongoing session during face-to-face interactions. They also concurred that student responses and the emotions conveyed in those responses help assess a student's engagement in the session. Regardless of the educators' hearing conditions, this unanimous agreement is clearly depicted in Figure 4.1.

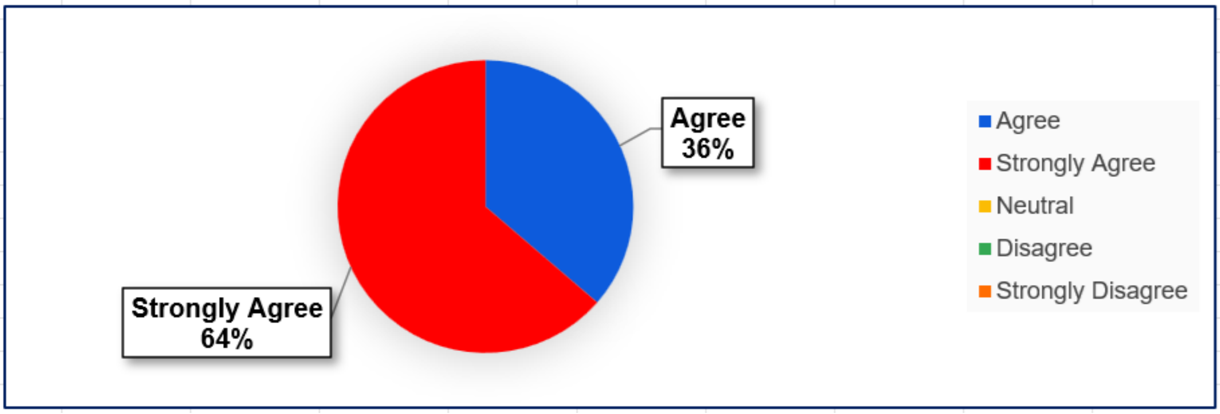


Figure 4.1: Respondents' views about student emotion and their engagement

The previous survey question (Figure 4.1) explored the link between educators' views on emotion in student feedback and engagement. Building on that, the next question aimed to find out from educators how student emotions conveyed in verbal feedback affect their teaching strategies. Figure 4.2 indicates the importance of affective engagement and its benefits in conducting classes. Most educators (94%), regardless of their hearing condition, agreed that student vocal feedback and the emotions conveyed help adjust teaching strategies. This perspective is crucial for understanding the need to recognize emotions in student feedback and its link to adapting teaching approaches accordingly. Only 6% of educators were neutral in their responses.

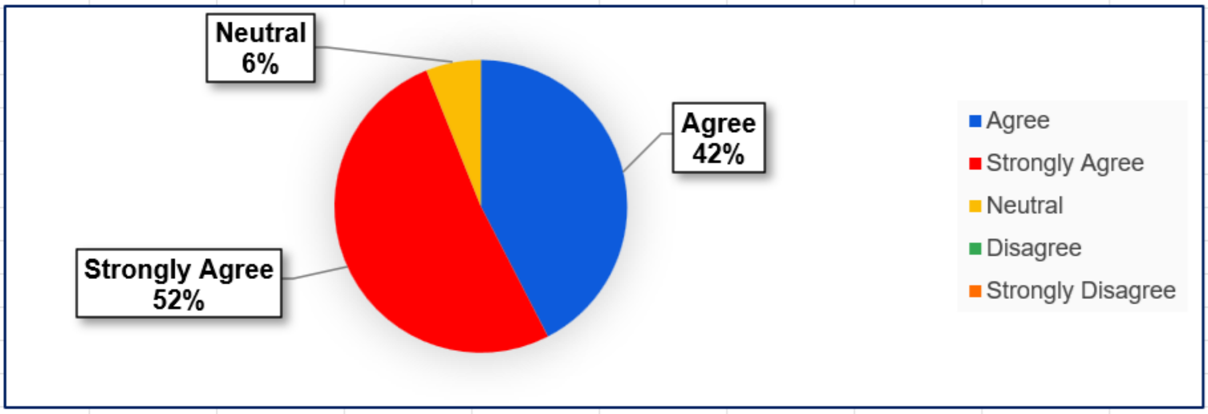


Figure 4.2: Respondents' view on the link between teaching strategy and students' emotions

The next questions focused exclusively on educators' opinions about online classes. Hence, the next question was to find out if the real-time student feedback for understanding student engagement was even more critical in an online class, as there is no physical interaction. As shown in Figure 4.3, 94% of the educators agreed with this, while 6% remained neutral. This feedback was crucial for the proposed SER system, which aims to facilitate emotion recognition in online environments.

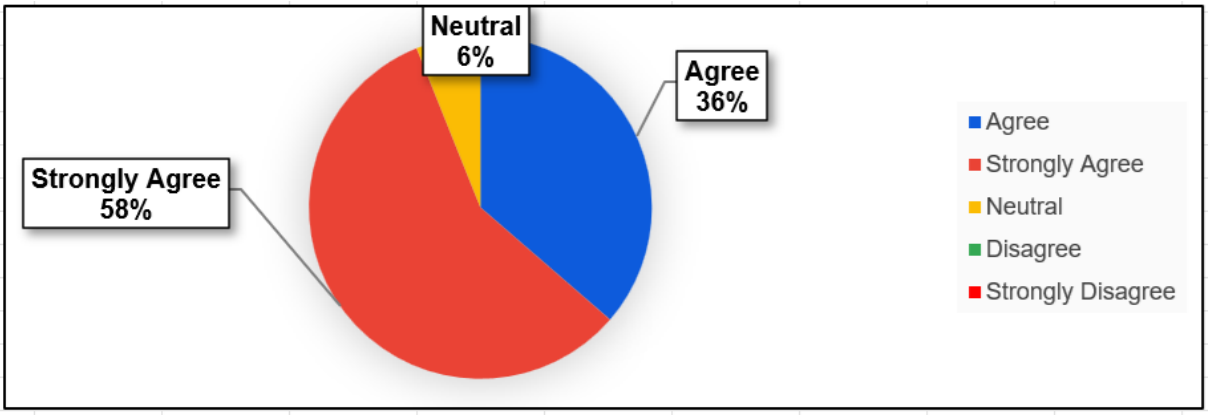


Figure 4.3: Respondents' views on student feedback are valued more in online classes

Figure 4.4 shows how educators collect feedback from students during online classes. Most respondents selected the chat box or verbal feedback in the multiple-choice question. This provides a clear understanding of educators' preferences based on available resources.

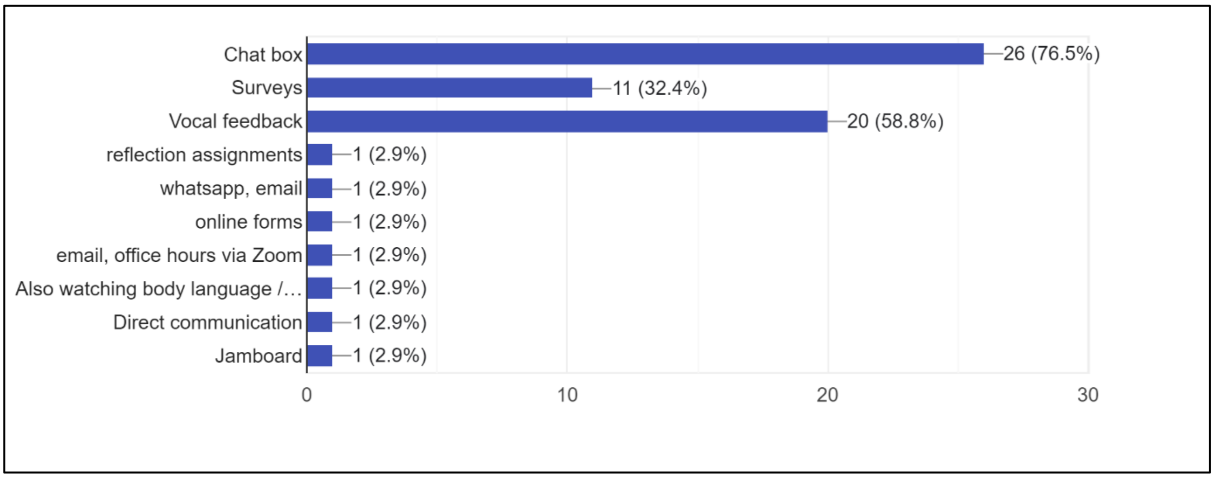


Figure 4.4: Types of methods used by respondents to gather feedback in online classes

As illustrated in Figure 4.5, 85% of the educators could hear the students' verbal feedback in online classes. This included educators without hearing impairment and educators with mild to moderate hearing impairment. The remaining 15% who could not hear were educators with profound hearing loss. Out of these, 6% could not hear at all, whereas the other 9% could hear using hearing aids.

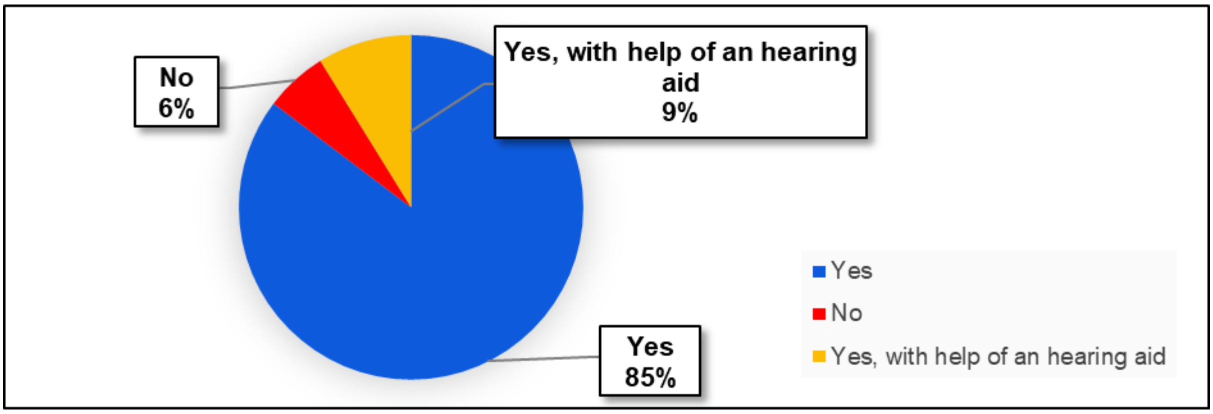


Figure 4.5: Respondents' views on hearing student vocal feedback in online classes

The next question sought respondents' views on the benefits of automatically detecting and displaying students' underlying emotions (such as happy, sad, neutral, or bored) as images based on student verbal feedback. Figure 4.6 clearly illustrates that among the 33 educators who participated in the study, 76% agreed, while 9% disagreed and 15% remained neutral. This feedback is crucial for this study, indicating strong support for the automatic display of student emotion as an image and highlighting the need for an SER system.

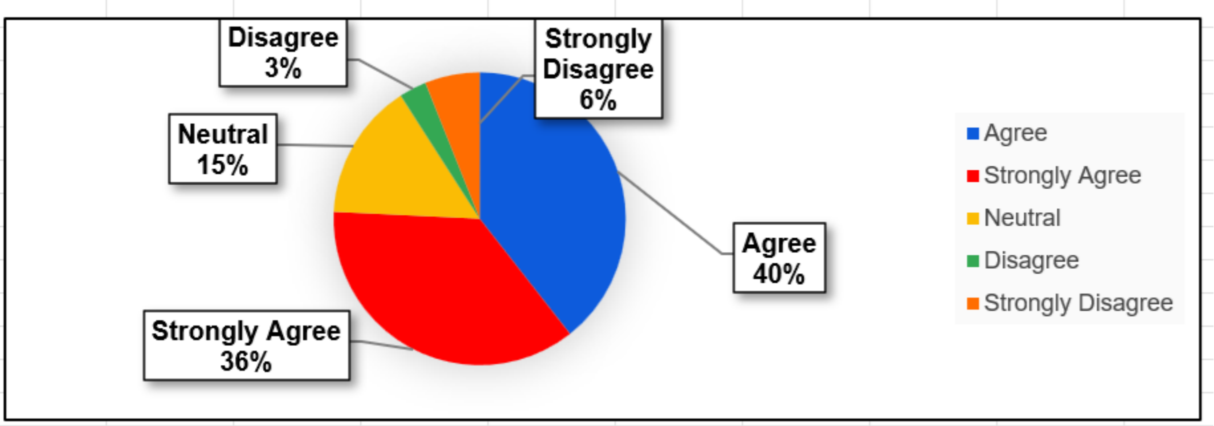


Figure 4.6: Respondents' view on automatically detecting and displaying student vocal feedback as an image

Finally, according to Figure 4.7, 94% of the educators agreed that the teaching approach can be adjusted based on student's emotions during online classes. Only 3% disagreed, and another 3% remained neutral. This overwhelming support was critical as it highlighted the significance of implementing an SER system for online teaching. It indicated the potential benefits of such a system in enhancing the teaching process by effectively considering and responding to students' emotions.

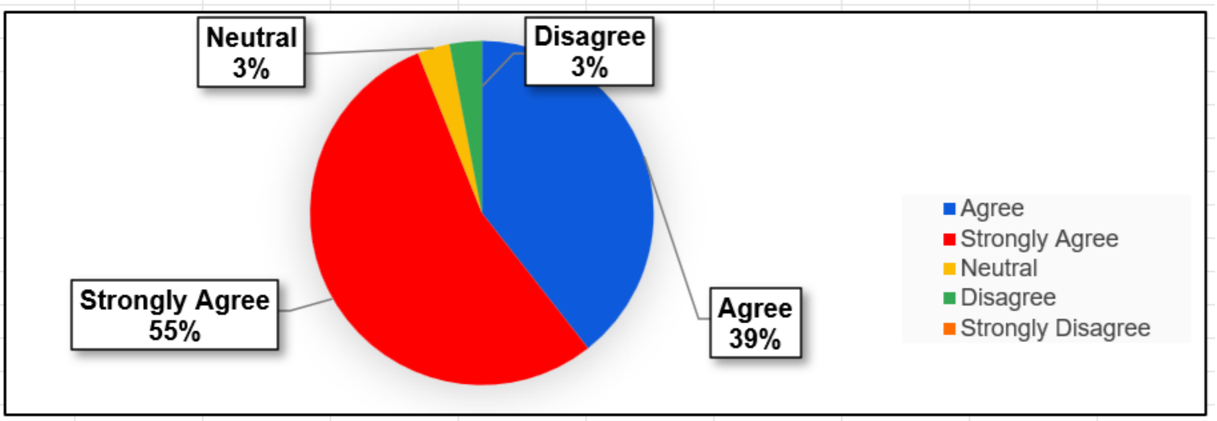


Figure 4.7: Respondents view on how student emotion helps in adjusting online teaching

4.2.3 Student Feedback and their Emotion Recognition in Online Systems - Analysis of the Qualitative Data

The survey also included an open-ended question asking respondents to describe the methods they use to assess student emotions during online sessions. The goal was to understand educators' diverse strategies to discern student emotions, considering it is very challenging. These responses were analysed using thematic analysis, a qualitative data analysis method introduced by Braun and Clarke [202]. Thematic analysis is a powerful method that systematically identifies, analyses, and interprets recurring patterns in data [203]. This approach was used to organize and interpret quantitative feedback from the educators, both with and without hearing impairment. Codes were assigned to recurring responses, and three key themes emerged: (i) Emotional Accessibility Challenges, (ii) Facial Expression, and (iii) Engagement and Interaction, as shown in Figure 4.8.

Emotional Accessibility Challenges:

Most respondents expressed that assessing students' emotions during online classes is challenging. A significant reason for this difficulty is that students often turn off their cameras during the session. Cameras are switched off primarily because leaving them on can be distracting and also to conserve bandwidth. However, in such cases, when a student's face is not visible, it becomes difficult to interpret their emotions. Many educators have openly admitted that they do not know how to assess students' emotions in an online class setting. Below are some of the responses from the participants: -

- *"It's challenging as many times students don't switch on cameras ... so a system to gauge emotions will be very useful".*
- *"It is not easy to detect emotions of the students in online class especially if the class is big and it is not possible to have videos on".*
- *"They are online, but do not want to answer the questions".*
- *"No idea".*
- *"Not aware".*

Facial Expression

Some respondents mentioned that they could recognize students' emotions by observing facial cues or expressions. However, to do so, they would initially request the students to turn on their cameras when a question was directed at them individually. Then, emotions were gauged by watching their facial expressions as they responded. Below are some selected responses from the participants: -

- *"Give quiz and see the responses of students frequency, use technology to detect the facial movement and analyse the feedback acoustics to know the emotions".*
- *"Via video conversation & from facial expression".*
- *"Ask the participants to on the webcam and give responses independently thru the webcam".*
- *"Their gestures and face expressions".*

- *"Facial gestures".*

Engagement and Interaction

Some educators stated they could determine students' emotions by observing their reactions during class participation and engagement. The following are selected responses from the participants: -

- *"Interaction with the students to gauge their enthusiasm, motivation, and attitude after a brainstorming session with the students"*
- *"I give instructions with specific words (that keep changing) to put into chat to gauge their attention and to keep them engaged. If I can get them to engage in the chat, I can detect their emotions partly from their comments and banters with each other on the chat".*
- *"By asking questions".*
- *"None other than verbal feedback, i.e get the student to ask or answer questions".*
- *"Voice".*

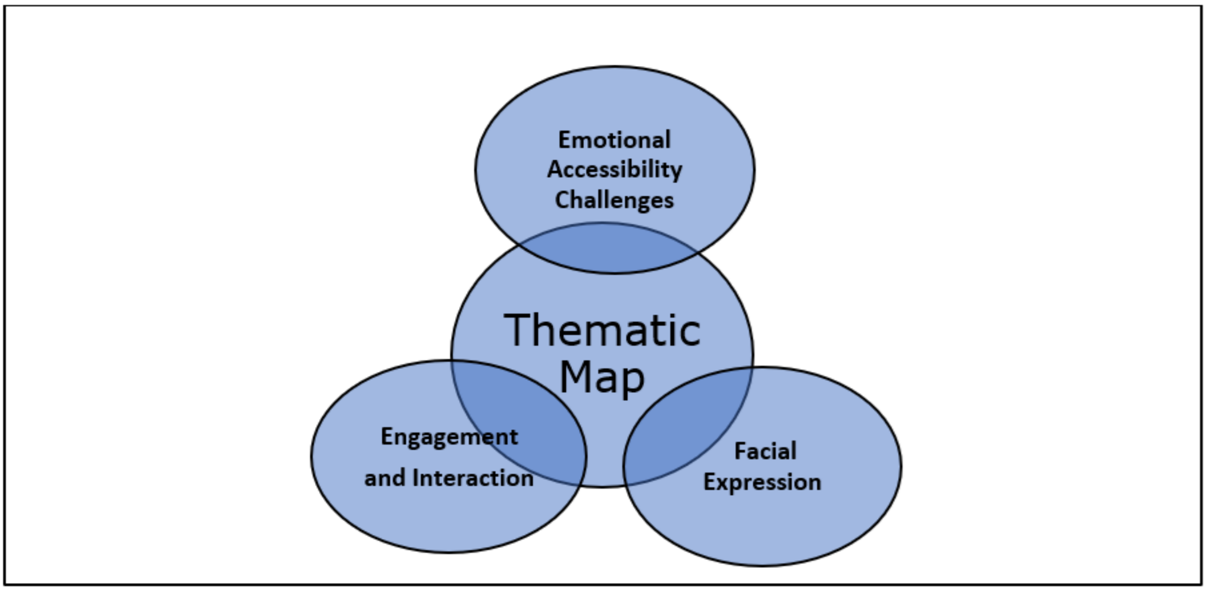

Emotional Accessibility Challenges
Thematic Map
Engagement and Interaction
Facial Expression

Figure 4.8: Thematic map: gauging student emotions during online sessions.

4.2.4 Discussion

The quantitative and qualitative findings from the survey, as outlined in Sections 4.2.2 and 4.2.3, which were based on both open-ended and close-ended questions, are triangulated and discussed in this section. The results clearly indicate that gauging student emotions through feedback taken intermittently during class is crucial for enhancing the quality of teaching. This is evident in the

‘Results and Analysis’ section and applies to traditional face-to-face and online classes. Additionally, the findings revealed a shared uncertainty among educators, regardless of their hearing ability, in assessing student emotions with certainty during online sessions. This uncertainty could lead to a negative impact on the teaching or learning process. The responses also strongly highlight the benefits of incorporating an automatic speech emotion detection system into online classes. Furthermore, the analysis confirms a strong visual bias in emotion recognition, further supporting the need for a clear visual representation of the SER output.

The 10 respondents with hearing impairments reported varying degrees of hearing loss, ranging from mild to profound. As a result, most of them struggled to hear students' verbal feedback clearly in online. Because these verbal responses go unheard, there is a high chance the students’ emotions may often remain undetected during online classes. However, all respondents emphasized the importance of emotions conveyed through vocal feedback for understanding student engagement, especially in online classes with limited physical interaction. Additionally, all ten respondents unanimously agreed that an automatic system for detecting and displaying students’ underlying emotions (such as happy, sad or neutral) through visual cues would be highly beneficial. This feature would greatly assist in adjusting teaching strategies to better meet students’ needs. Table 4.2 summarises the survey results specific to respondents with hearing impairments.

Table 4.2: Results of the respondents suffering from some form of hearing impairment

Sno.	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
	Hearing Condition of the respondent?	Were you able to hear student vocal feedback in a face-to-face with your current hearing condition?	Do you agree emotions carried in the student vocal feedback (in a face-to-face class) are helpful in understanding their engagement ?	Educators can adjust teaching strategies based on the student vocal response and emotions carried in them?	In an online class, real-time student feedback for understanding student engagement is more critical, as there is no physical interaction?	Were you able to hear the student's voice in vocal feedback, in an online class?	In an online class, would it be beneficial if the underlying emotion of the student is automatically detected from and displayed using an image?	Understanding the student emotion during online teaching is helpful in adjusting the teaching approach for the educator.
1	Moderate	Yes	Agree	Agree	Agree	Yes	Agree	Agree
2	Mild	Yes	Strongly Agree	Strongly Agree	Agree	Yes	Strongly agree	Strongly agree
3	Severe	No	Strongly Agree	Strongly Agree	Strongly Agree	No	Agree	Agree
4	Severe	Yes, with help of an hearing aid	Strongly Agree	Agree	Strongly Agree	Yes, with help of an hearing aid	Agree	Strongly agree
5	Profound	Yes, with help of a hearing aid	Strongly Agree	Strongly Agree	Strongly Agree	Yes, with help of an hearing aid	Agree	Strongly agree

6	Moderate	Yes, with help of a hearing aid	Strongly Agree	Strongly Agree	Strongly Agree	Yes, with help of an hearing aid	Neutral	Strongly agree
7	Mild	Yes	Agree	Agree	Agree	No	Agree	Agree
8	Mild	Yes	Strongly Agree	Strongly Agree	Strongly Agree	Yes	Strongly agree	Strongly agree
9	Moderate	No	Agree	Agree	Agree	Yes	Neutral	Agree
10	Mild	Yes	Agree	Agree	Neutral	Yes	Strongly agree	Strongly agree

The above discussion shows that all the aims of the preliminary study have been achieved. The first task of the preliminary study was setting up the aims, as seen in Section 3.3.

The aims were:

1) Investigate the perspective of educators (with and without hearing impairment) on the necessity of detecting student emotions to enhance teaching outcomes for effective teaching.
2) Understand the challenges faced by educators (with and without hearing impairment) in discerning student emotions during online teaching.
3) Identify the benefits of having an emotion-detection system (SER) for detecting student emotions.

A preliminary study survey was conducted to gather information on the challenges educators face in discerning student emotions during online teaching sessions. The participants included educators with and without hearing impairments, allowing them to gain insights into their specific needs. An analysis of the survey results indicated the potential benefits of implementing an emotion-detection system via speech, which was the intention of this research work.

This signifies the completion of stage 1 of the HCI research process for SER, as depicted in Figure 3.1. The corresponding section in Figure 4.9 has been shaded to illustrate this.

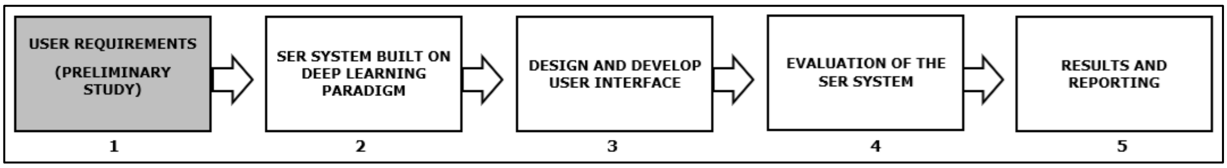

USER REQUIREMENTS
(PRELIMINARY STUDY)
SER SYSTEM BUILT ON DEEP LEARNING PARADIGM
DESIGN AND DEVELOP USER INTERFACE
EVALUATION OF THE SER SYSTEM
RESULTS AND REPORTING
1
2
3
4
5

Figure 4.9. Updated process of the HCI research used for SER

In the next section, results from the experiments conducted using three hybrid models are presented, their performance is highlighted and compared, and the model with the highest speech emotion recognition accuracy is identified.

4.3 Experimental Results and Performance Comparison of Three Selected Hybrid Models Integrating 1D, 2D, and 3D CNNs

In Section 3.4, the network configuration of the three hybrid architectures, Model A (hybrid average), Model B (hybrid merge), and Model C (hybrid sequential) were presented. These models were designed to leverage the fusion of acoustic features, each combining three types of CNNs (1D, 2D, and 3D) using different fusion techniques. The first two architectures, Models A (hybrid average) and Model B (hybrid merge), are multi-stream CNNs. Model A (hybrid average) fuses streams by averaging decisions, while Model B (hybrid merge) concatenates the outputs from the multi-streams for further processing. The third architecture, Model C (hybrid sequential), combines 3D, 2D, and 1D CNNs sequentially. The following section compares and presents results from all three hybrid architectures. The experiments were conducted based on the methodology described in Section 3.4, for the three hybrid classifiers. As detailed below, the experiments demonstrated that Model A (hybrid average) achieved the highest accuracy among the tested models.

To evaluate and compare the hybrid multi-dimensional CNN models - Model A (hybrid average), Model B (hybrid merge), and Model C (hybrid sequential), five datasets were utilized: two semi-natural and three acted. These datasets, which are IEMOCAP, DEMoS, RAVDESS, EMODB and TESS, are described in detail in Section 3.4.1. The models were implemented and executed on the Google Colab platform using Python, with GPU acceleration provided by Colab to expedite training. After importing the necessary libraries, the datasets were loaded for analysis. The programming environment used has been described in detail in Chapter 3, Section 3.7. The following sections will outline the experimental process, beginning with a detailed description of data processing and exploration.

4.3.1 Data Processing and Exploration

Under the data processing and exploration step, the audio files are first processed in each dataset sequentially. All the files were in .wav format. Each file represented a unique emotion, with its emotion label embedded in the file name. The initial step involved extracting these emotion labels from the file names and storing them in a data frame. The file name and its corresponding emotion label were placed in separate data frame columns. For the next step, data exploration was performed to analyse the dataset's characteristics and gain better insights into its nature. This was done mainly to examine whether any biases in the data needed to be addressed and to provide insights for subsequent analysis and modelling. This included determining whether the dataset was imbalanced by checking if certain emotions had more samples than others. Figures 4.10 to 4.14 display the emotions present in the five datasets and the sample count for each emotion.

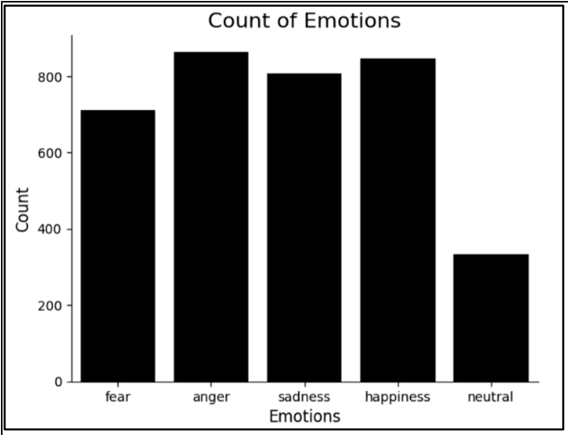


Figure 4.10: Count of emotions in DEMoS

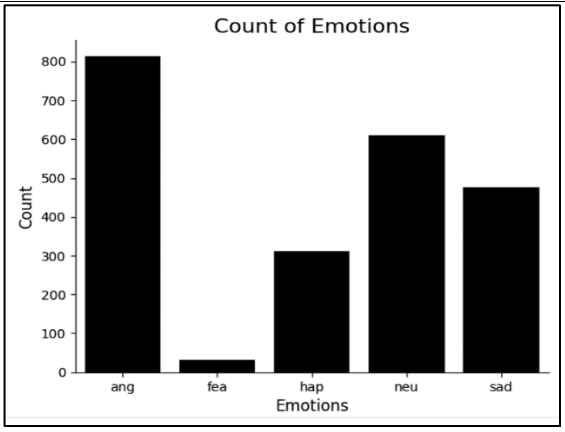


Figure 4.11: Count of emotions in IEMOCAP

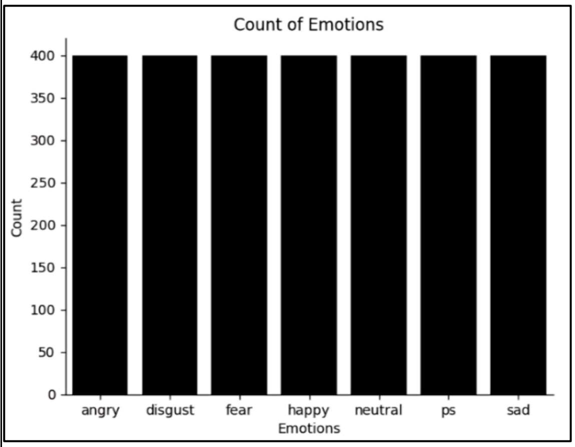


Figure 4.12: Count of emotions in TESS

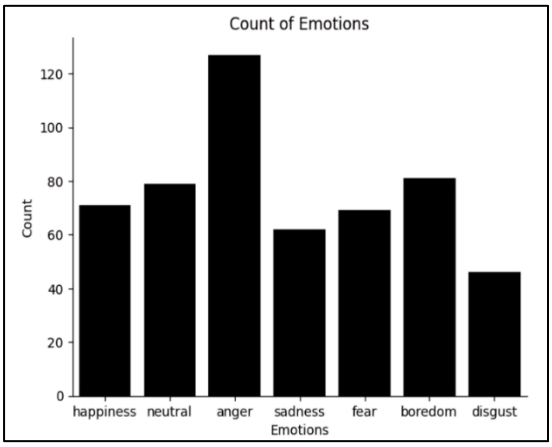


Figure 4.13: Count of emotions in EMO-DB

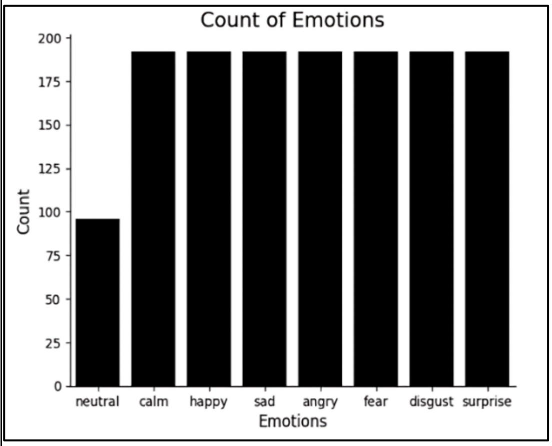


Figure 4.14: Count of emotions in RAVDESS

4.3.2 Data Balancing

After data exploration, the next step was balancing the datasets using under-sampling or over-sampling techniques. As can be seen from Figures 4.10 to 4.14, some of the datasets are imbalanced. Data imbalance in machine learning occurs when the distribution of classes or labels is uneven [204], which can negatively impact training and classification performance. This issue can be addressed in different ways, including over-sampling, which adds more records to the minority class, or under-sampling, which removes records from the majority class [204]. For this study, random over-sampling was applied to balance the datasets by duplicating instances from minority classes to match the size of the majority class. This method was chosen for its simplicity and effectiveness in preserving the original emotional characteristics of speech without introducing synthetic noise or altering the data distribution [204]. Unlike methods such as SMOTE, SIEOS and others that generate synthetic samples through interpolation and risk distorting subtle emotional cues, random over-sampling retains the authenticity of the original data [244]. It is especially suitable for deep learning models that require clean, consistent inputs and avoids the added complexity and potential instability of synthetic augmentation. Using the imbalanced dataset would have resulted in biased predictions favouring majority emotions and poor recognition of minority classes [27]. The number of audio files after balancing is shown in Tables 4.3, 4.4, and 4.5 for the selected datasets. Further data exploration was conducted, including visualizations of raw waveforms, spectrograms, MFCC features, and audio renditions. Figure 4.15 illustrates these visualizations for each unique emotion using representative files from the DEMoS dataset.

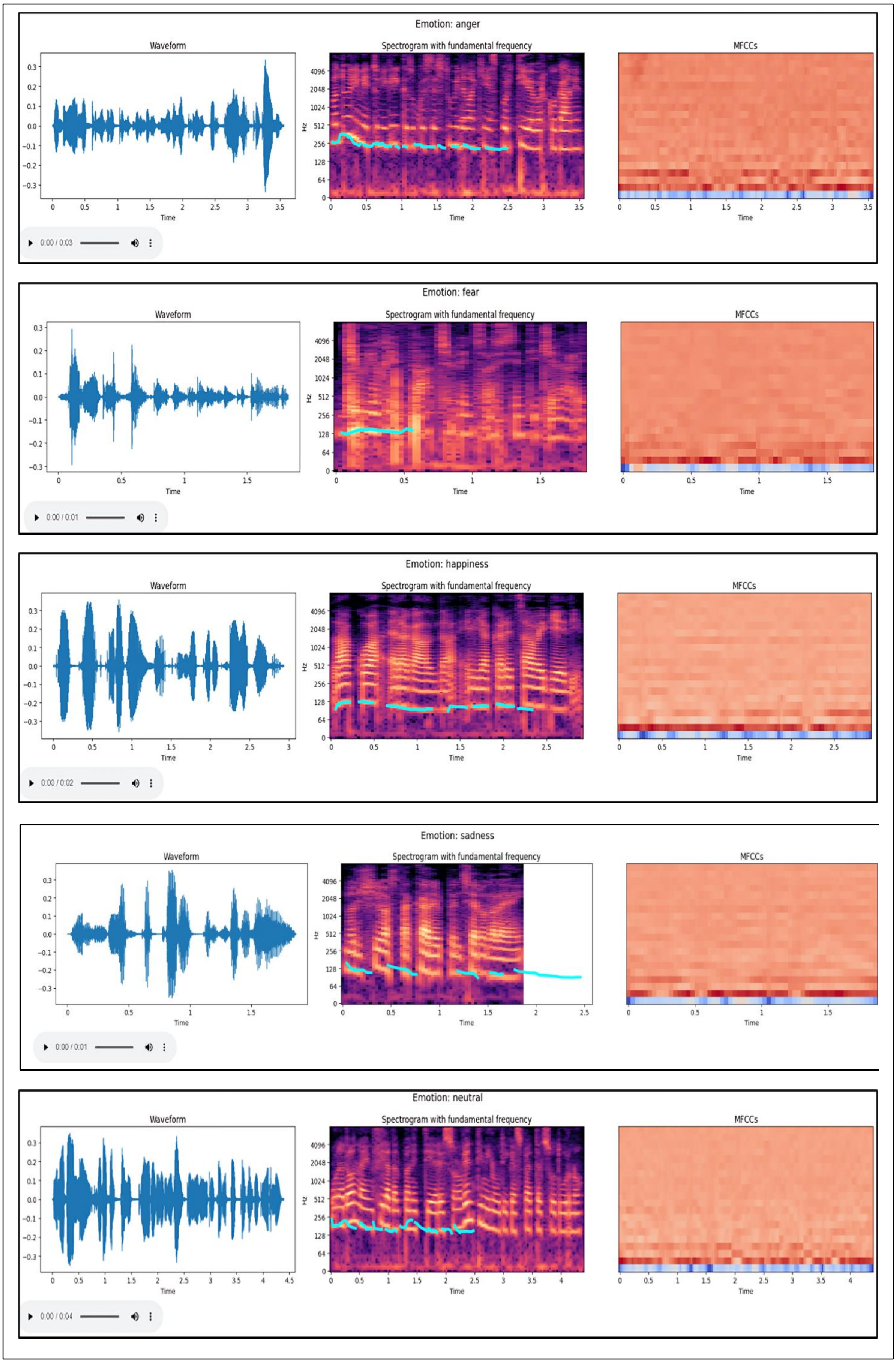

Emotion: anger
Waveform
Spectrogram with fundamental frequency
MFCCs
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
Hz
4096
2048
1024
512
256
128
64
0
0.5
1
1.5
2
2.5
3
3.5
Time
0:00 / 0:03
Emotion: fear
Waveform
Spectrogram with fundamental frequency
MFCCs
Time
0:00 / 0:01
Emotion: happiness
Waveform
Spectrogram with fundamental frequency
MFCCs
Time
0:00 / 0:02
Emotion: sadness
Waveform
Spectrogram with fundamental frequency
MFCCs
Time
0:00 / 0:01
Emotion: neutral
Waveform
Spectrogram with fundamental frequency
MFCCs
4
4.5
Time
0:00 / 0:04

Figure 4.15. Displaying waveform, spectrogram, MFCCs, and audio for every unique emotion (for selected files of the DEMoS dataset)

After data balancing, the next steps—pre-processing, feature extraction, and classification—were carried out, as described in the methodology (Section 3.4.2). The results from these experiments are discussed in the following sections.

4.3.3 Results from the Experiments Conducted

An experimental comparison of three hybrid architectures, each combining 1D, 2D, and 3D CNNs with different fusion techniques—hybrid average (Model A), hybrid merge (Model B), and hybrid sequential (Model C)—was conducted. The models were trained on five datasets: DEMoS, IEMOCAP, TESS, EMODB, and RAVDESS. Detailed descriptions of the datasets and model configurations are provided in Section 3.4.1 of the methodology chapter. Accuracy, Confusion Matrix, and Classification Report were employed as evaluation metrics to compare the performance of the three hybrid architectures. Below, the Accuracy Score, Confusion Matrix, and Classification Report are first defined to highlight their significance as evaluation metrics. Subsequently, the experimental results for the three hybrid architectures across the five datasets are presented.

a. Accuracy Score

The accuracy score is the metric used in this work, which gives the classification rate of the models based on the test data. It is used for the evaluation of the classification algorithm in the context of a multi-class classification problem. Classification accuracy is an important parameter for evaluating a developed model [94]. Achieving high accuracy ensures that the developed model can be applied for real-world applications like a speech emotion recognition system [118]. The accuracy score gives the correctly classified samples out of the total samples provided by the test dataset [94]. It is the ratio of correct predictions to the total number of predictions [69]. It's a simple metric that indicates how often the model's predictions are accurate and is typically expressed as a percentage:

$$\text{Accuracy} = \frac{correct\ predictions}{total\ predictions} * 100 \tag{4.1}$$

b. Confusion Matrix (CM)

Besides accuracy, the Confusion Matrix is a valuable evaluation metric for classification models. It provides a tabular representation of the counts of actual versus predicted classifications, highlighting four key values:

- **True Positives (TP):** Correctly predicted positive cases.
- **True Negatives (TN):** Correctly predicted negative cases.
- **False Positives (FP):** Incorrectly predicted as positive when actually negative.
- **False Negatives (FN):** Incorrectly predicted as negative when actually positive.

The confusion matrix illustrates the model's ability to correctly and incorrectly classify audio data. The horizontal axis represents the predicted labels, while the vertical axis represents the actual labels. Each intersection indicates the confusion ratio between different emotion classifications. The diagonal values correspond to the recall for each emotion, representing the proportion of correctly classified samples within each category [71, 98].

c. Classification Report: Precision, Recall and F1 Score

The classification report lists the network model's precision, recall, and F1 score for recognizing each emotion classification. It also includes the macro average, weighted average, and accuracy, providing an overall evaluation of the model's performance across the dataset. This report offers a comprehensive understanding of the recognition strength of each emotion classification and the model's overall performance. As per [73, 98], precision, recall, and F1 score are defined below:

- **Precision**: The proportion of correctly predicted positive observations out of all predicted positives.

$$\text{Precision} = \frac{TP}{TP+FP} \tag{4.2}$$

- **Recall**: The proportion of correctly predicted positive observations out of all actual positives.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad (4.3)$$

- **F1 Score**: The harmonic mean of precision and recall, balancing the two. It is especially useful when both false positives and false negatives need to be considered.

Below are the accuracy score, confusion matrix, and classification report results from experiments conducted on the three selected hybrid architectures using different fusion techniques: averaging (Model A), parallel merging (Model B), and sequential integration (Model C) have been displayed for the 5 datasets.

1. Model A (hybrid average)

The proposed hybrid average model produced results for the five selected datasets, as displayed in Table 4.3 below.

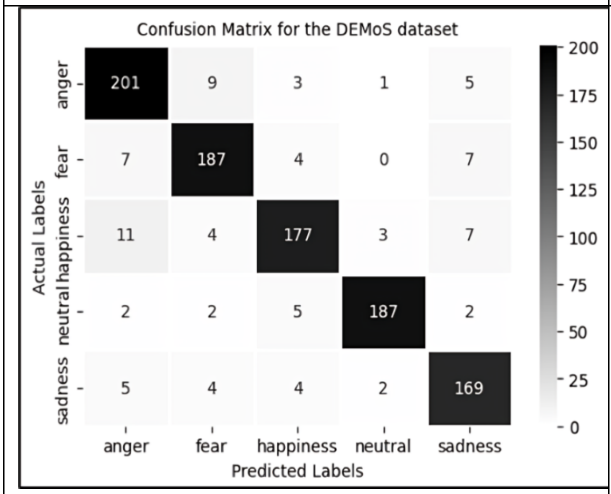
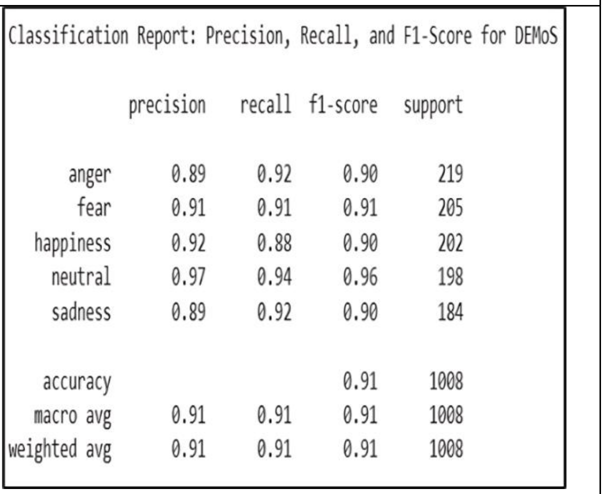
Table 4.3. Accuracy score of Model A (hybrid average)

DATASET	DATASET TYPE	EMOTIONS	NUMBER OF AUDIO FILES	CNN MODEL	INDIVIDUAL ACCURACY SCORE	ACCURACY SCORE SCORE FUSION (HYBRID AVERAGE)
DEMoS	Semi Natural	Fear, Anger, sadness, happiness, neutral	4030	1D CNN	43.15%	91%
				2D CNN	90.87%	
				3D CNN	79.46%	
IEMOCAP	Semi Natural	Fear, sadness, anger, happiness, neutral	2380	1D CNN	68.40%	82%
				2D CNN	84%	
				3D CNN	81.8%	
TESS	Acted	Angry, disgust, fear, happy, neutral, pleasant surprise, sad	2800	1D CNN	82.14%	100%
				2D CNN	99.86%	
				3D CNN	98.29%	
EMODB	Acted	Anger, boredom, anxiety, happiness, sadness, disgust. neutral	889	1D CNN	42.15%	91%
				2D CNN	84.30%	
				3D CNN	88.79%	
RAVDESS	Acted	Neutral, calm, happy, sad, angry, fear, disgust, surprise	1536	1D CNN	47.66%	91%
				2D CNN	90.89%	
				3D CNN	72.40%	

If the datasets are imbalanced, it can lead to lower emotion recognition accuracy. For example, when using the imbalanced EMODB dataset without any over-sampling, the accuracy scores were significantly lower. The overall emotion recognition accuracy (%) across the three models with and without data balancing were as follows:

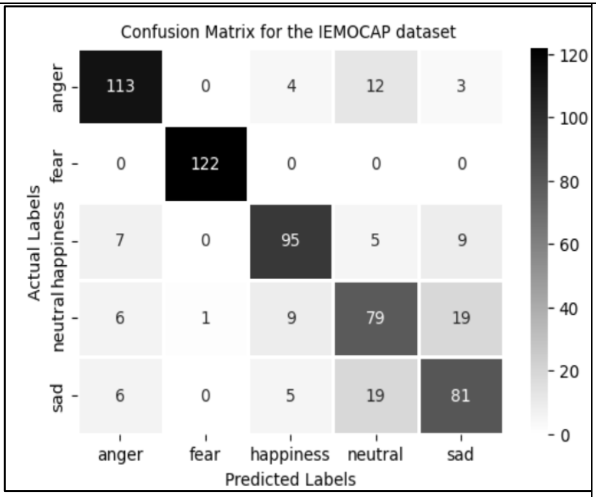
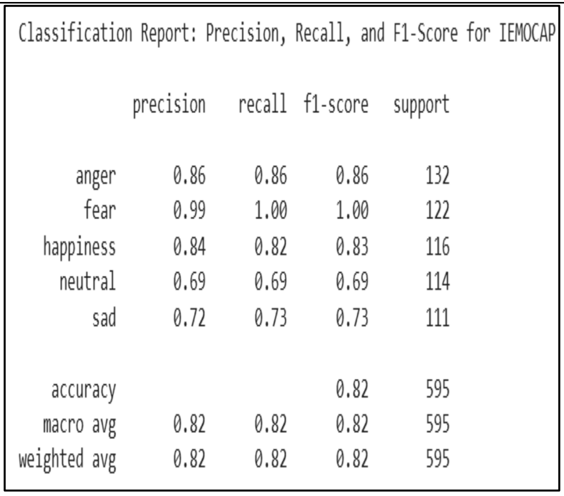
- **With data balancing**: Model 1 – 42.15%, Model 2 – 84.30%, Model 3 – 88.79%
- **Without data balancing**: Model 1 – 37.31%, Model 2 – 74.63%, Model 3 – 76.12%

These results clearly demonstrate the negative impact of class imbalance and emphasize the importance of applying random over-sampling to enhance both model performance and fairness.

CM	PRECISION, RECALL AND F1 SCORE
	

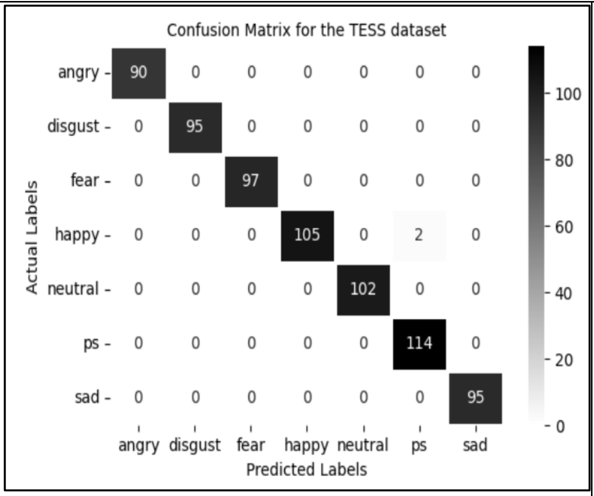
```
Classification Report: Precision, Recall, and F1-Score for DEMoS

              precision    recall  f1-score   support

       anger       0.89      0.92      0.90       219
        fear       0.91      0.91      0.91       205
   happiness       0.92      0.88      0.90       202
     neutral       0.97      0.94      0.96       198
     sadness       0.89      0.92      0.90       184

    accuracy                           0.91      1008
   macro avg       0.91      0.91      0.91      1008
weighted avg       0.91      0.91      0.91      1008
```

Figure 4.16: Confusion Matrix and Classification Report for the DEMoS Dataset Using Model A (Hybrid Average)

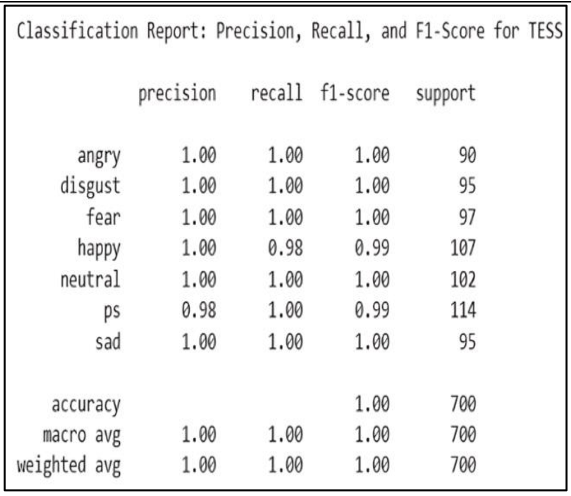
	

```
Classification Report: Precision, Recall, and F1-Score for IEMOCAP

              precision    recall  f1-score   support

       anger       0.86      0.86      0.86       132
        fear       0.99      1.00      1.00       122
   happiness       0.84      0.82      0.83       116
     neutral       0.69      0.69      0.69       114
         sad       0.72      0.73      0.73       111

    accuracy                           0.82       595
   macro avg       0.82      0.82      0.82       595
weighted avg       0.82      0.82      0.82       595
```

Figure 4.17: Confusion Matrix and Classification Report for the IEMOCAP Dataset Using Model A (Hybrid Average)

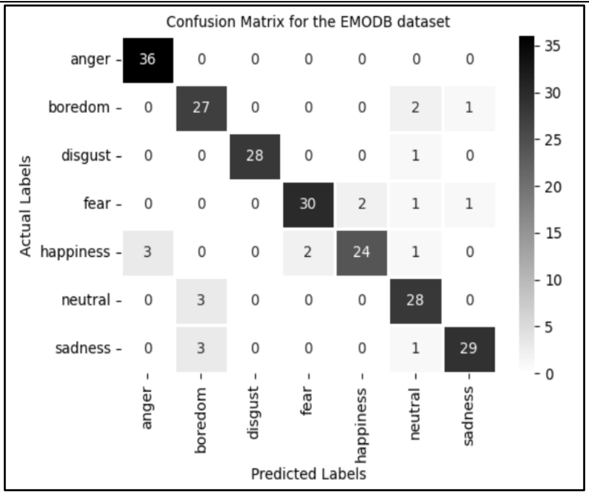


	angry	disgust	fear	happy	neutral	ps	sad
angry	90	0	0	0	0	0	0
disgust	0	95	0	0	0	0	0
fear	0	0	97	0	0	0	0
happy	0	0	0	105	0	2	0
neutral	0	0	0	0	102	0	0
ps	0	0	0	0	0	114	0
sad	0	0	0	0	0	0	95

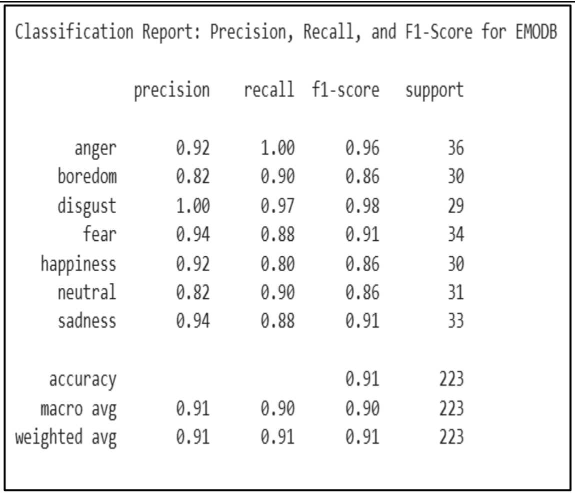


```
Classification Report: Precision, Recall, and F1-Score for TESS

              precision    recall  f1-score   support

       angry       1.00      1.00      1.00        90
     disgust       1.00      1.00      1.00        95
        fear       1.00      1.00      1.00        97
       happy       1.00      0.98      0.99       107
     neutral       1.00      1.00      1.00       102
          ps       0.98      1.00      0.99       114
         sad       1.00      1.00      1.00        95

    accuracy                           1.00       700
   macro avg       1.00      1.00      1.00       700
weighted avg       1.00      1.00      1.00       700
```

Figure 4.18: Confusion Matrix and Classification Report for the TESS Dataset Using Model A (Hybrid Average)

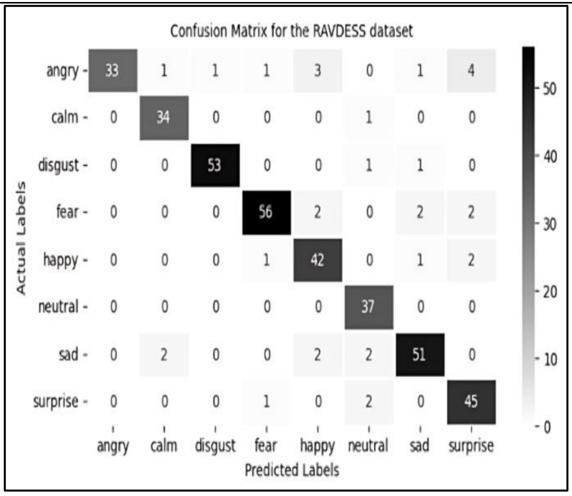


	anger	boredom	disgust	fear	happiness	neutral	sadness
anger	36	0	0	0	0	0	0
boredom	0	27	0	0	0	2	1
disgust	0	0	28	0	0	1	0
fear	0	0	0	30	2	1	1
happiness	3	0	0	2	24	1	0
neutral	0	3	0	0	0	28	0
sadness	0	3	0	0	0	1	29

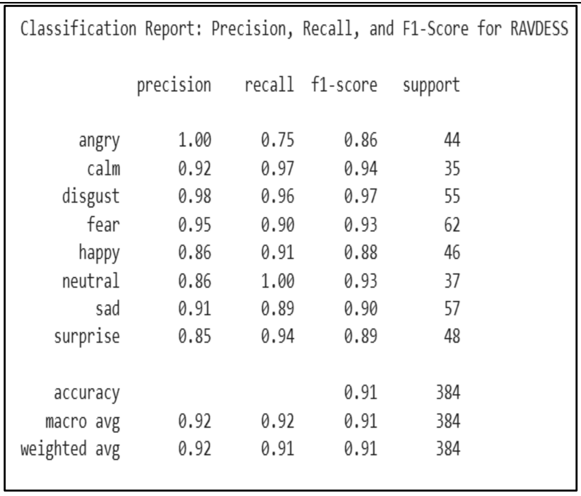


```
Classification Report: Precision, Recall, and F1-Score for EMODB

              precision    recall  f1-score   support

       anger       0.92      1.00      0.96        36
     boredom       0.82      0.90      0.86        30
     disgust       1.00      0.97      0.98        29
        fear       0.94      0.88      0.91        34
   happiness       0.92      0.80      0.86        30
     neutral       0.82      0.90      0.86        31
     sadness       0.94      0.88      0.91        33

    accuracy                           0.91       223
   macro avg       0.91      0.90      0.90       223
weighted avg       0.91      0.91      0.91       223
```

Figure 4.19: Confusion Matrix and Classification Report for the EMODB Dataset Using Model A (Hybrid Average)



	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	33	1	1	1	3	0	1	4
calm	0	34	0	0	0	1	0	0
disgust	0	0	53	0	0	1	1	0
fear	0	0	0	56	2	0	2	2
happy	0	0	0	1	42	0	1	2
neutral	0	0	0	0	0	37	0	0
sad	0	2	0	0	2	2	51	0
surprise	0	0	0	1	0	2	0	45



```
Classification Report: Precision, Recall, and F1-Score for RAVDESS

              precision    recall  f1-score   support

       angry       1.00      0.75      0.86        44
        calm       0.92      0.97      0.94        35
     disgust       0.98      0.96      0.97        55
        fear       0.95      0.90      0.93        62
       happy       0.86      0.91      0.88        46
     neutral       0.86      1.00      0.93        37
         sad       0.91      0.89      0.90        57
    surprise       0.85      0.94      0.89        48

    accuracy                           0.91       384
   macro avg       0.92      0.92      0.91       384
weighted avg       0.92      0.91      0.91       384
```

Figure 4.20: Confusion Matrix and Classification Report for the RAVDESS Dataset Using Model A (Hybrid Average)

2. Model B (hybrid merge)

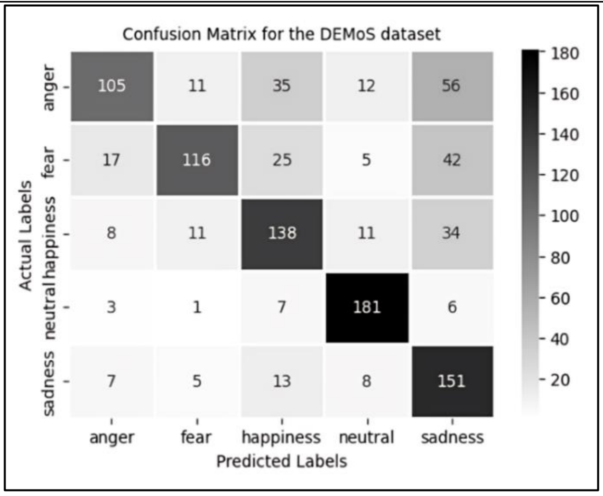
The proposed hybrid merge model generated results for the five selected datasets as displayed in Table 4.4.

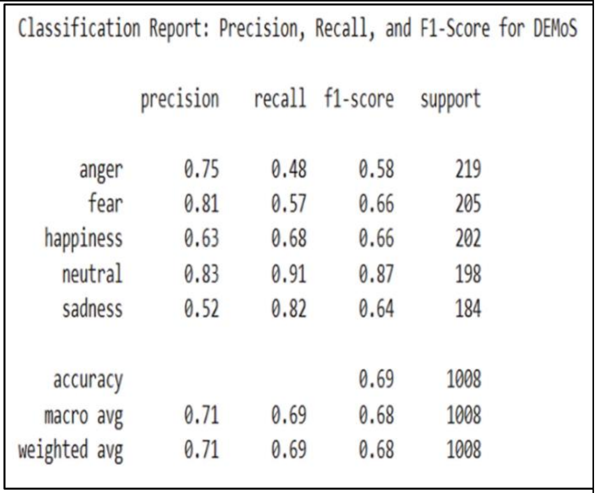
Table 4.4: Accuracy score of Model B (hybrid merge)

DATASET	DATASET TYPE	EMOTIONS	NUMBER OF AUDIO FILES	ACCURACY SCORE (HYBRID MERGE)
DEMoS	Semi Natural	Fear, Anger, sadness, happiness, neutral	4030	69%
IEMOCAP	Semi Natural	Fear, sadness, anger, happiness, neutral	2380	70%
TESS	Acted	Angry, disgust, fear, happy, neutral, pleasant surprise, sad	2800	100%
EMODB	Acted	Anger, boredom, anxiety, happiness, sadness, disgust. neutral	889	73%
RAVDESS	Acted	Neutral, calm, happy, sad, angry, fear, disgust, surprise	1536	45%

CM

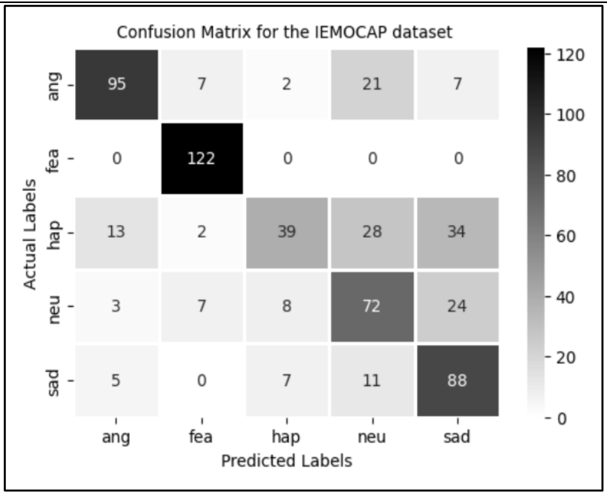
PRECISION, RECALL AND F1 SCORE

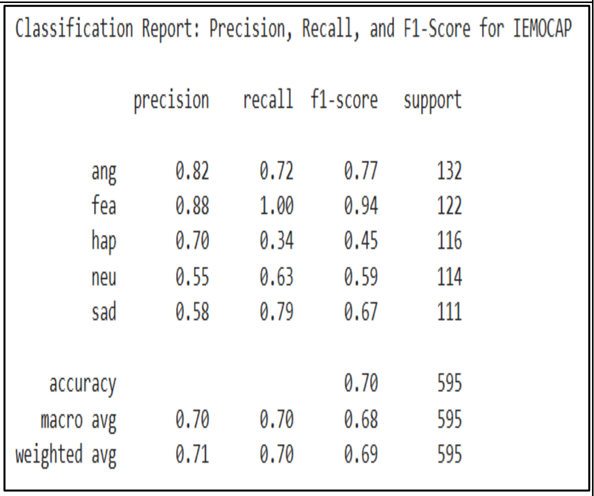




```
Classification Report: Precision, Recall, and F1-Score for DEMoS

              precision    recall  f1-score   support

       anger       0.75      0.48      0.58       219
        fear       0.81      0.57      0.66       205
   happiness       0.63      0.68      0.66       202
     neutral       0.83      0.91      0.87       198
     sadness       0.52      0.82      0.64       184

    accuracy                           0.69      1008
   macro avg       0.71      0.69      0.68      1008
weighted avg       0.71      0.69      0.68      1008
```

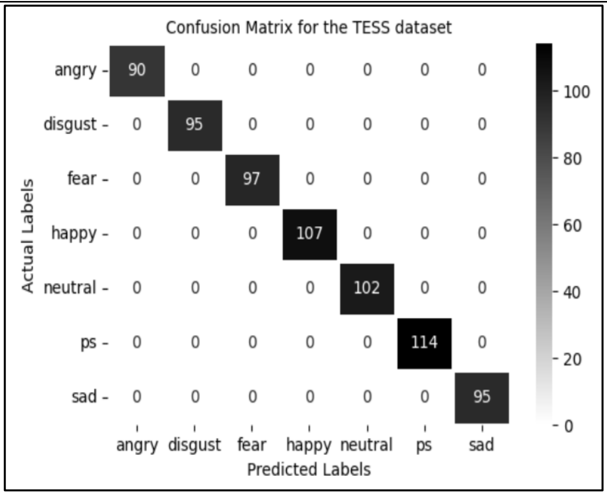
Figure 4.21: Confusion Matrix and Classification Report for the DEMoS Dataset Using Model B (Hybrid Merge)

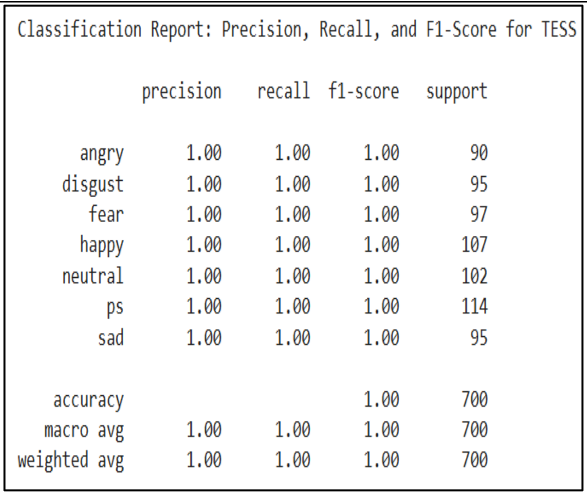




```
Classification Report: Precision, Recall, and F1-Score for IEMOCAP

              precision    recall  f1-score   support

         ang       0.82      0.72      0.77       132
         fea       0.88      1.00      0.94       122
         hap       0.70      0.34      0.45       116
         neu       0.55      0.63      0.59       114
         sad       0.58      0.79      0.67       111

    accuracy                           0.70       595
   macro avg       0.70      0.70      0.68       595
weighted avg       0.71      0.70      0.69       595
```

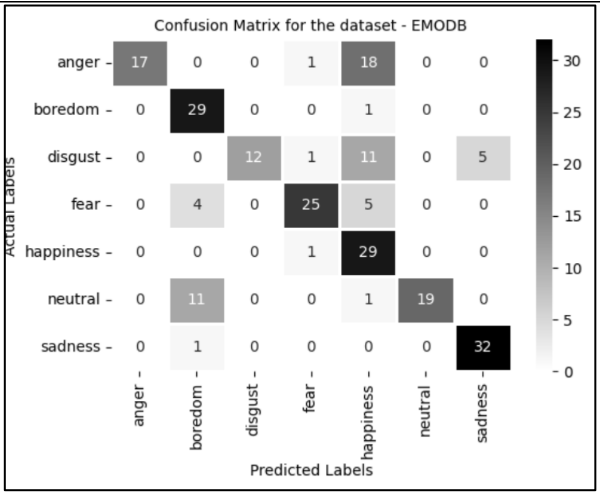
Figure 4.22 Confusion Matrix and Classification Report for the IEMOCAP Dataset Using Model B (Hybrid Merge)

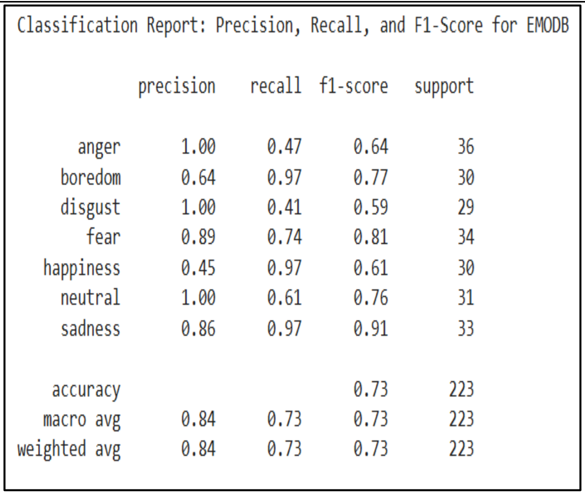




```
Classification Report: Precision, Recall, and F1-Score for TESS

              precision    recall  f1-score   support

       angry       1.00      1.00      1.00        90
     disgust       1.00      1.00      1.00        95
        fear       1.00      1.00      1.00        97
       happy       1.00      1.00      1.00       107
     neutral       1.00      1.00      1.00       102
          ps       1.00      1.00      1.00       114
         sad       1.00      1.00      1.00        95

    accuracy                           1.00       700
   macro avg       1.00      1.00      1.00       700
weighted avg       1.00      1.00      1.00       700
```

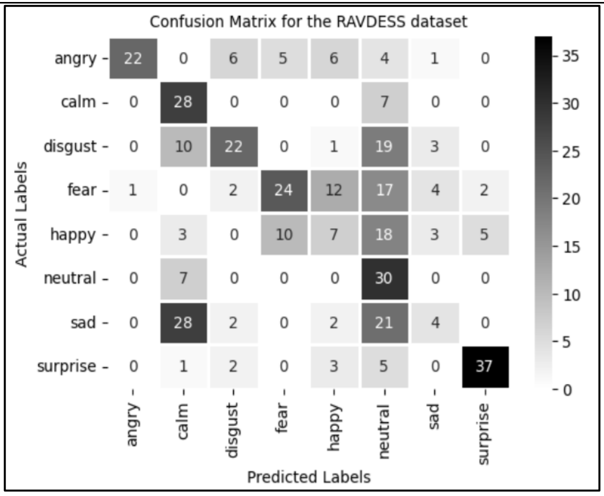
Figure 4.23 Confusion Matrix and Classification Report for the TESS Dataset Using Model B (Hybrid Merge)

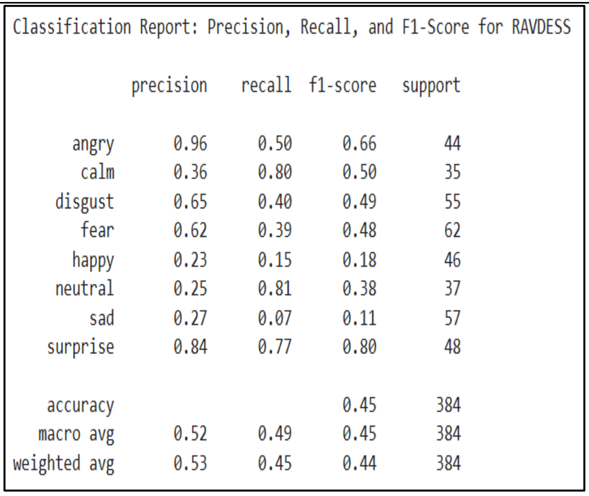




```
Classification Report: Precision, Recall, and F1-Score for EMODB

              precision    recall  f1-score   support

       anger       1.00      0.47      0.64        36
     boredom       0.64      0.97      0.77        30
     disgust       1.00      0.41      0.59        29
        fear       0.89      0.74      0.81        34
   happiness       0.45      0.97      0.61        30
     neutral       1.00      0.61      0.76        31
     sadness       0.86      0.97      0.91        33

    accuracy                           0.73       223
   macro avg       0.84      0.73      0.73       223
weighted avg       0.84      0.73      0.73       223
```

Figure 4.24 Confusion Matrix and Classification Report for the EMODB Dataset Using Model B (Hybrid Merge)





```
Classification Report: Precision, Recall, and F1-Score for RAVDESS

              precision    recall  f1-score   support

       angry       0.96      0.50      0.66        44
        calm       0.36      0.80      0.50        35
     disgust       0.65      0.40      0.49        55
        fear       0.62      0.39      0.48        62
       happy       0.23      0.15      0.18        46
     neutral       0.25      0.81      0.38        37
         sad       0.27      0.07      0.11        57
    surprise       0.84      0.77      0.80        48

    accuracy                           0.45       384
   macro avg       0.52      0.49      0.45       384
weighted avg       0.53      0.45      0.44       384
```

Figure 4.25 Confusion Matrix and Classification Report for the RAVDESS Dataset Using Model B (Hybrid Merge)

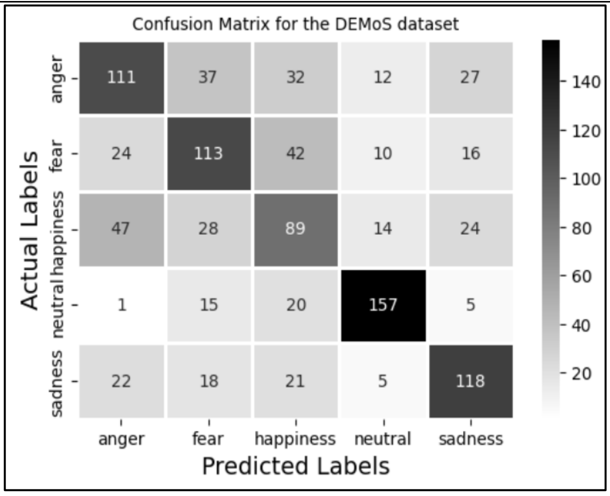
3. Model C

The proposed hybrid sequential model produced for the five selected datasets the results as displayed in Table 4.5 below.

Table 4.5:Accuracy score of Model C (hybrid sequential)

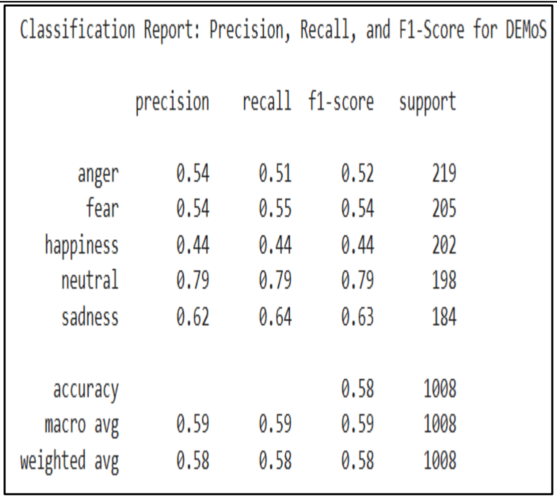
Dataset	Dataset Type	Emotions	Audio files	Accuracy Score (Hybrid sequential)
DEMoS	Semi Natural	Fear, Anger, sadness, happiness, neutral	4030	58%
IEMOCAP	Semi Natural	Fear, sadness, anger, happiness, neutral	2380	63%
TESS	Acted	Angry, disgust, fear, happy, neutral, pleasant surprise, sad	2800	98%
EMODB	Acted	Anger, boredom, anxiety, happiness, sadness, disgust. neutral	889	80%
RAVDESS	Acted	Neutral, calm, happy, sad, angry, fear, disgust, surprise	1536	53%

CM | **Precision, Recall and F1 Score**



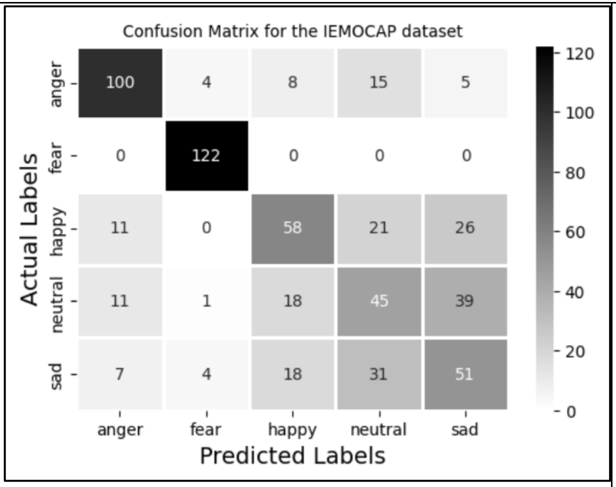
Confusion Matrix for the DEMoS dataset

Actual Labels \ Predicted Labels	anger	fear	happiness	neutral	sadness
anger	111	37	32	12	27
fear	24	113	42	10	16
happiness	47	28	89	14	24
neutral	1	15	20	157	5
sadness	22	18	21	5	118



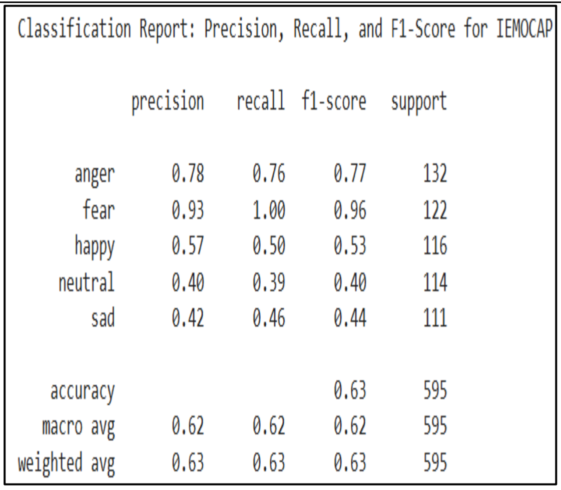
```
Classification Report: Precision, Recall, and F1-Score for DEMoS

              precision    recall  f1-score   support

       anger       0.54      0.51      0.52       219
        fear       0.54      0.55      0.54       205
   happiness       0.44      0.44      0.44       202
     neutral       0.79      0.79      0.79       198
     sadness       0.62      0.64      0.63       184

    accuracy                           0.58      1008
   macro avg       0.59      0.59      0.59      1008
weighted avg       0.58      0.58      0.58      1008
```

Figure 4.26: Confusion Matrix and Classification Report for the DEMoS Dataset Using Model C (Hybrid Sequential)



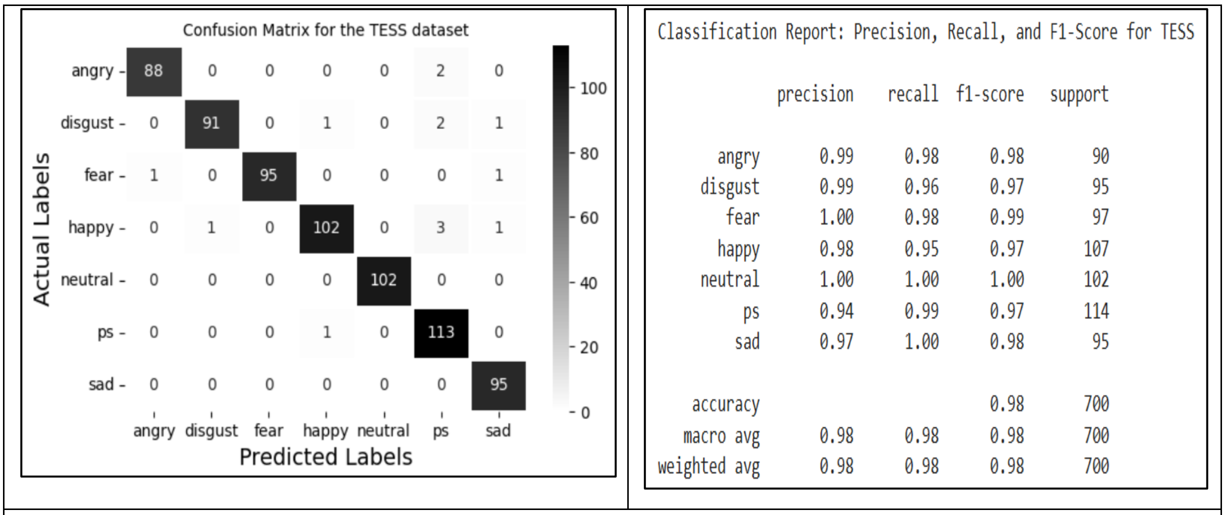
Confusion Matrix for the IEMOCAP dataset

Actual Labels \ Predicted Labels	anger	fear	happy	neutral	sad
anger	100	4	8	15	5
fear	0	122	0	0	0
happy	11	0	58	21	26
neutral	11	1	18	45	39
sad	7	4	18	31	51



```
Classification Report: Precision, Recall, and F1-Score for IEMOCAP

              precision    recall  f1-score   support

       anger       0.78      0.76      0.77       132
        fear       0.93      1.00      0.96       122
       happy       0.57      0.50      0.53       116
     neutral       0.40      0.39      0.40       114
         sad       0.42      0.46      0.44       111

    accuracy                           0.63       595
   macro avg       0.62      0.62      0.62       595
weighted avg       0.63      0.63      0.63       595
```

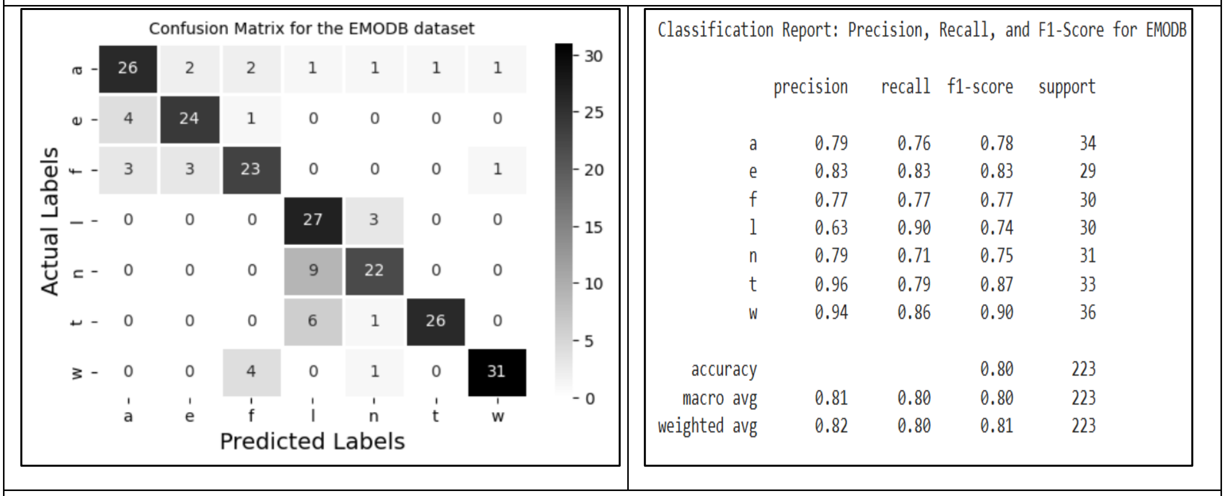
Figure 4.27: Confusion Matrix and Classification Report for the IEMOCAP Dataset Using Model C (Hybrid Sequential)



	angry	disgust	fear	happy	neutral	ps	sad
angry	88	0	0	0	0	2	0
disgust	0	91	0	1	0	2	1
fear	1	0	95	0	0	0	1
happy	0	1	0	102	0	3	1
neutral	0	0	0	0	102	0	0
ps	0	0	0	1	0	113	0
sad	0	0	0	0	0	0	95

```
Classification Report: Precision, Recall, and F1-Score for TESS

              precision    recall  f1-score   support

       angry       0.99      0.98      0.98        90
     disgust       0.99      0.96      0.97        95
        fear       1.00      0.98      0.99        97
       happy       0.98      0.95      0.97       107
     neutral       1.00      1.00      1.00       102
          ps       0.94      0.99      0.97       114
         sad       0.97      1.00      0.98        95

    accuracy                           0.98       700
   macro avg       0.98      0.98      0.98       700
weighted avg       0.98      0.98      0.98       700
```

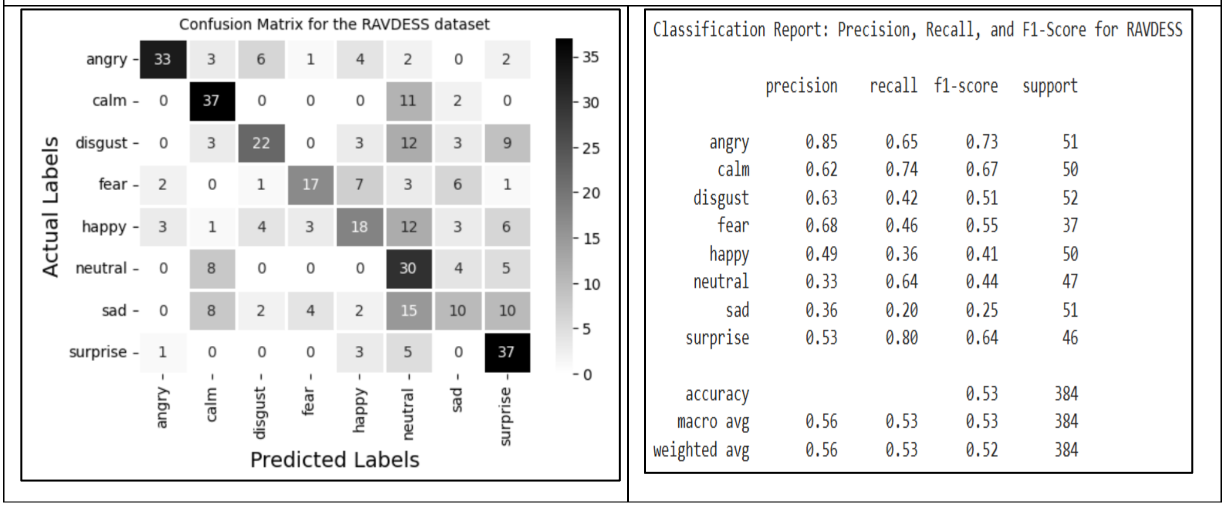
Figure 4.28: Confusion Matrix and Classification Report for the TESS Dataset Using Model C (Hybrid Sequential)



	a	e	f	l	n	t	w
a	26	2	2	1	1	1	1
e	4	24	1	0	0	0	0
f	3	3	23	0	0	0	1
l	0	0	0	27	3	0	0
n	0	0	0	9	22	0	0
t	0	0	0	6	1	26	0
w	0	0	4	0	1	0	31

```
Classification Report: Precision, Recall, and F1-Score for EMODB

              precision    recall  f1-score   support

           a       0.79      0.76      0.78        34
           e       0.83      0.83      0.83        29
           f       0.77      0.77      0.77        30
           l       0.63      0.90      0.74        30
           n       0.79      0.71      0.75        31
           t       0.96      0.79      0.87        33
           w       0.94      0.86      0.90        36

    accuracy                           0.80       223
   macro avg       0.81      0.80      0.80       223
weighted avg       0.82      0.80      0.81       223
```

Figure 4.29: Confusion Matrix and Classification Report for the EMODB Dataset Using Model C (Hybrid Sequential)



	angry	calm	disgust	fear	happy	neutral	sad	surprise
angry	33	3	6	1	4	2	0	2
calm	0	37	0	0	0	11	2	0
disgust	0	3	22	0	3	12	3	9
fear	2	0	1	17	7	3	6	1
happy	3	1	4	3	18	12	3	6
neutral	0	8	0	0	0	30	4	5
sad	0	8	2	4	2	15	10	10
surprise	1	0	0	0	3	5	0	37

```
Classification Report: Precision, Recall, and F1-Score for RAVDESS

              precision    recall  f1-score   support

       angry       0.85      0.65      0.73        51
        calm       0.62      0.74      0.67        50
     disgust       0.63      0.42      0.51        52
        fear       0.68      0.46      0.55        37
       happy       0.49      0.36      0.41        50
     neutral       0.33      0.64      0.44        47
         sad       0.36      0.20      0.25        51
    surprise       0.53      0.80      0.64        46

    accuracy                           0.53       384
   macro avg       0.56      0.53      0.53       384
weighted avg       0.56      0.53      0.52       384
```

Figure 4.30: Confusion Matrix and Classification Report for the RAVDESS Dataset Using Model C (Hybrid Sequential)

4.3.4 Experimental Comparison of the Three Proposed Hybrid Models

An experimental comparison of the three proposed hybrid models was made to identify the one with the highest speech emotion recognition accuracy. The selected model was then applied to detect emotions in real-world scenarios by making it a real-time application. The experimental results from the three selected hybrid models have been depicted in the tables (4.3-4.5). Table 4.6 compares accuracy % across the three hybrid models, highlighting the best performances in bold. The results clearly demonstrate that the hybrid average model (A) consistently achieves the highest accuracy across all five selected datasets.

Table 4.6: Comparison of the accuracy % of the three selected hybrid models

DATASET	DATASET TYPE	HYBRID AVERAGE	HYBRID MERGE	HYBRID SEQUENTIAL
DEMoS	Semi Natural	**91%**	69%	58%
IEMOCAP	Semi Natural	**82%**	70%	63%
TESS	Acted	**100%**	100%	98%
EMODB	Acted	**91%**	73%	80%
RAVDESS	Acted	**91%**	45%	53%

Table 4.6 compares the experimental results for all three hybrid models. Model A (hybrid average) was the one that achieved the highest accuracy for speech emotion recognition. It was selected to be subsequently implemented as a real-time application to detect emotions in practical scenarios.

Section 4.3 presented the results of experiments conducted using three hybrid models, comparing their performance and identifying the model with the highest speech emotion recognition accuracy. Model A (hybrid average), which achieved the highest accuracy, was selected from the three to be developed into a real-world SER application. This signifies the completion of stage 2 of the HCI research process for SER, as depicted in Figure 3.2.

The corresponding section in Figure 4.31 has been shaded to illustrate this.

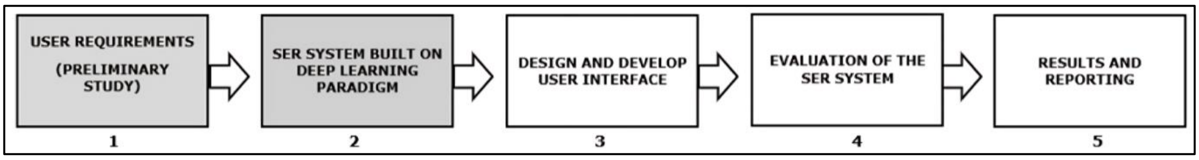

USER REQUIREMENTS (PRELIMINARY STUDY)
SER SYSTEM BUILT ON DEEP LEARNING PARADIGM
DESIGN AND DEVELOP USER INTERFACE
EVALUATION OF THE SER SYSTEM
RESULTS AND REPORTING
1
2
3
4
5

Figure 4.31: Updated process of the HCI research used for SER

Furthermore, the results in Section 4.3.5 indicate that the second research objective has been successfully addressed, as stated below:–

RO2 - To compare and assess the effectiveness of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers through three selected fusion techniques, identifying the approach with the highest accuracy and efficiency and adapting it to build a real-world SER application.

The next section presents the results of integrating the selected SER model with the graphical user interface, highlighting its functionality as a real-time emotion detection application. The subsequent sections focus on the research efforts to achieve the next two objectives (RO3 and RO4), which are as follows:

RO3 - To design and develop a graphical user interface (GUI) that integrates SER-derived emotional feedback and displays emotions to late-deafened educators accurately and in real-time.

RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency and perceived impact on late deafened educators for an effective engagement in online education.

4.4 Real-Time Application of the SER system

Chapter 3 Section 3.5 presented the methodology for transforming the selected SER system into a functional, real-time application capable of detecting and displaying emotions. It covered details of the design and development of the user interface for the SER system. This section presents further enhancements to the selected SER system (Model A – hybrid average) to convert it into a functional, real-time application capable of instantly detecting and displaying emotions.

4.4.1 Model A (Hybrid Average) - Retraining with a Combination of Datasets

All the selected hybrid models discussed in Section 4.3 were trained to detect emotions using pre-processed datasets containing recordings labelled with specific emotions. The accuracy results for the five selected datasets (DEMoS, IEMOCAP, EMODB, RAVDESS, and TESS) across the three models are presented in Table 4.6. However, this research also emphasizes accurately detecting emotions from real-time utterances in real-world scenarios. To achieve this, the first step was to retrain the selected Model A (hybrid average) using a combination of datasets to increase the training sample size. Three of the five datasets with the highest accuracy for Model A (Hybrid Average) were selected: one semi-natural dataset (DEMoS) and two acted datasets (RAVDESS and TESS), as described in Table 4.6. Figure 4.32 illustrates the combination of the three selected datasets and displays the sample count for each emotion.

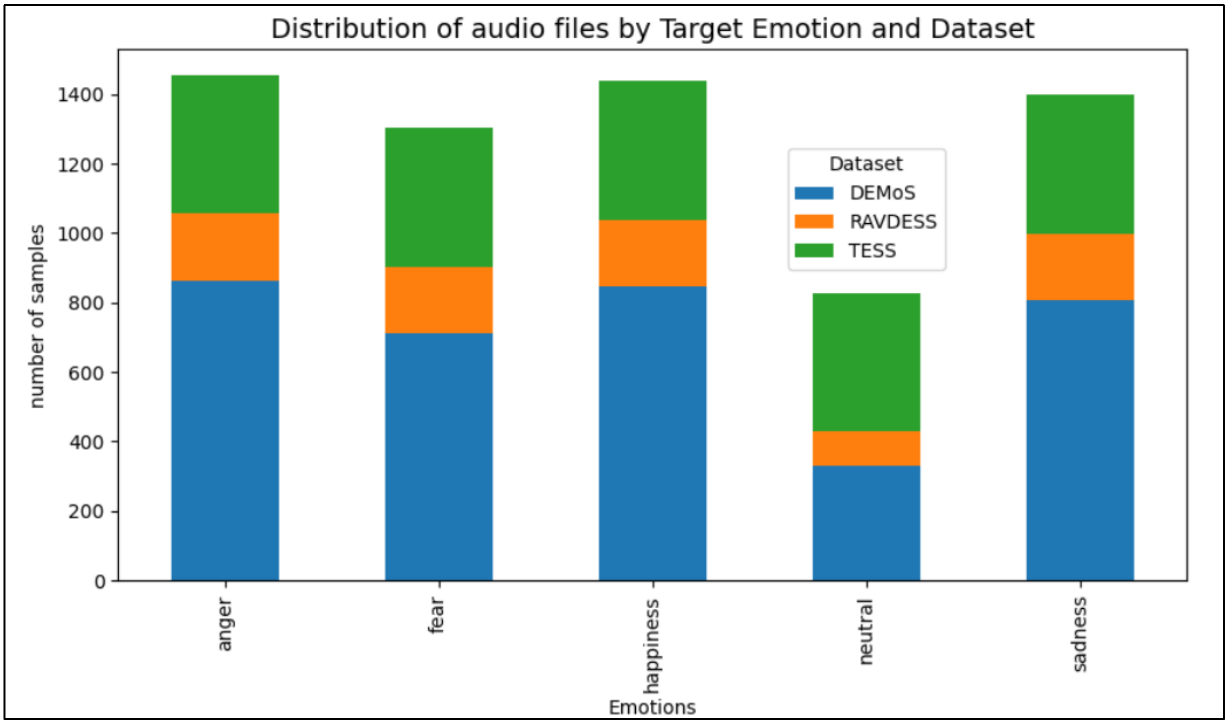


Figure 4.32: Combined distribution of audio files by target emotion across DEMoS, RAVDESS and TESS datasets

In this case, the data was balanced, as described in Section 4.3.3, bringing the total number of audio files to 7,280 for the 5 emotions: anger, fear, happiness, neutral, and sadness.

4.4.2 Selection of Emotions

As shown in Section 4.4.1, 5 emotions were chosen for the selected Model A (hybrid average), which needs to be built into a real-time application. Out of the selected 5 emotions – happiness, anger, sadness, and fear are part of Ekman's basic emotions, which are widely studied in psychology. They are universally recognized and considered fundamental across different cultures, making them ideal for this research work [19, 37, 43]. Also selected was the emotion neutral. Neutral, while not an emotion in Ekman's theory, serves as a baseline for emotional absence or no reaction, making it essential for practical classification tasks. All five are sufficiently distinct from one another in terms of acoustic features, reducing overlap and improving classification accuracy. In supporting late-deafened educators, these emotions align well with common emotional states observed in students [205, 206]. Categorizing them as positive (happy, neutral) or negative (anger, sad, fear) simplifies the feedback and makes it more actionable for late-deafened educators. By limiting the number of emotions, the SER system simplifies the computational process, ensuring more accurate and meaningful real-time feedback without overwhelming the user.

4.4.3 Data Augmentation

A data augmentation technique was applied to the (combined) dataset to enhance its size and variability, ensuring better generalization during model training. Data augmentation is a technique commonly practised in machine learning and deep learning that aims to increase the number of training samples via an automated approach [207]. Dynamic data augmentation improves the model's robustness, generalization, and overall performance, making it more suitable for real-world applications [158]. One approach involves creating new samples by applying various transformations and modifications to the original data, thereby artificially increasing the dataset's diversity [73]. For this work, random noise was introduced using functions from different Python libraries. The Librosa library was used to load the audio files and determine the sample rate, while NumPy was used to calculate the noise amplitude. Our process involved (1) loading all original audio files into memory, (2) applying Gaussian noise to each file when creating a batch and extracting features. Gaussian noise, a random signal often used to simulate real-world background noise, was added to enhance the model's learning and generalization capabilities [73]. This makes the model more suitable for real-world scenarios where background noise is common. Gaussian noise was specifically chosen over other augmentation techniques such as pitch shifting or time-stretching because it realistically simulates background interference typical of online educational environments, without altering the core emotional cues in the speech signal. This approach preserves the emotional integrity of the data while enhancing the model's generalization ability and performance in practical settings. The effectiveness of this augmentation is further discussed in the Results section, where improved generalization and validation performance were observed compared to training without noise augmentation.

4.4.4 Results Analysis

After data augmentation, the dataset, a combination of three datasets (DEMoS, TESS and RAVDESS), was divided into Training, Validation and Test sets. The training set is used to train the model, whereas the validation set is used to evaluate the model's performance during training. Lastly, the test set evaluates the model's performance after training is complete (this set is not involved during training) [208]. 25 % of the data is used for model testing, and the remaining 75 % is used to train the model. Out of the 75% separated for training, 25% of this training data was allocated to validation. The splitting of the dataset was used to plot the training and validation accuracy and loss graphs as shown in Figures 4.33, 4.34 and 4.35 below for the models 1D CNN, 2D CNN and 3D CNN, respectively, for the combined dataset (DEMoS, TESS and RAVDESS). These graphs help visualize the model's performance over time. It shows whether it is learning effectively (decreasing loss and increasing accuracy) and whether overfitting or underfitting occurs [105, 208].

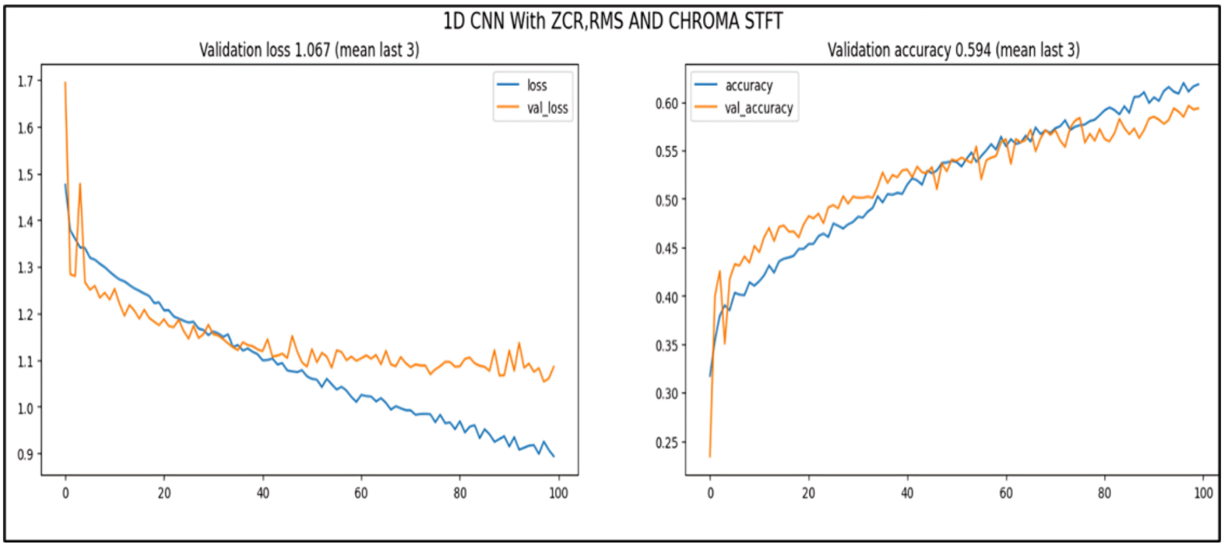


Figure 4.33: 1D CNN Performance Using ZCR, RMS, and Chroma-STFT Features: Validation Loss and Accuracy Graphs on Combined Datasets (DEMoS, TESS, and RAVDESS)

Figure 4.33 shows the two graphs showing the training and validation performance of a 1D CNN trained with the audio features - ZCR, RMS, and Chroma STFT over 100 epochs. The loss graph illustrates the loss curves for training and validation. Both losses sharply decrease during the initial epochs, with the training loss exhibiting a steady decline and the validation loss closely following the training curve. By the final epochs, the validation loss stabilizes, achieving a mean value of 1.067 (over the last three epochs), indicating reasonable generalization. In the accuracy graph, the accuracy curves rise rapidly in the early epochs for both training and validation data. The training accuracy consistently improves, while the validation accuracy follows a similar trend with slight fluctuations. By the final epochs, the accuracy stabilizes, achieving a mean validation accuracy of 0.594 (over the last three epochs). This suggests moderate model performance, with some potential for improvement in generalization or feature selection to further enhance validation accuracy.

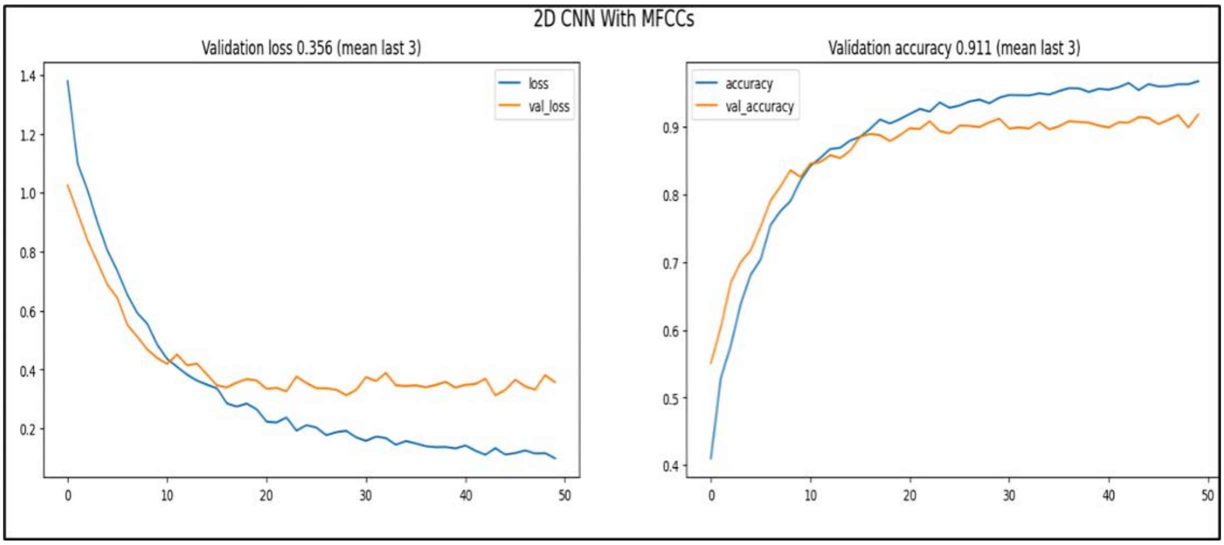


Figure 4.34: 2D CNN Performance Using MFCC Features: Validation Loss and Accuracy Graphs on Combined Datasets (DEMoS, TESS, and RAVDESS)

As seen in Figure 4.34, the two graphs display the training and validation performance of a 2D CNN trained with the audio features - MFCCs over 50 epochs. The loss graph illustrates the loss curves for training and validation. Both losses decrease significantly during the initial epochs, with the training loss steadily declining throughout the training process. The validation loss closely follows the training curve and stabilizes by the later epochs, achieving a mean value of 0.356 (over the last three epochs), indicating excellent generalization. In the accuracy graph, the accuracy curves for both training and validation data show a sharp increase during the early epochs. Training accuracy steadily improves; validation accuracy mirrors this trend with minimal fluctuations. By the final epochs, both curves stabilize, achieving high accuracy values. The validation accuracy reaches a mean value of 0.911 (over the last three epochs), demonstrating strong model performance with minimal overfitting and consistent alignment between training and validation accuracy.

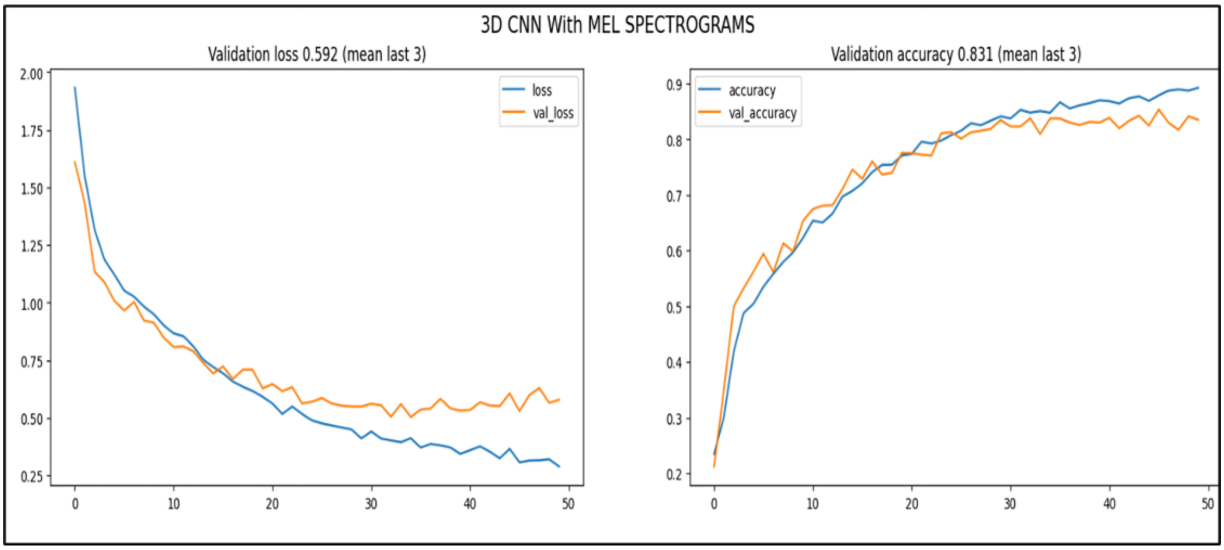
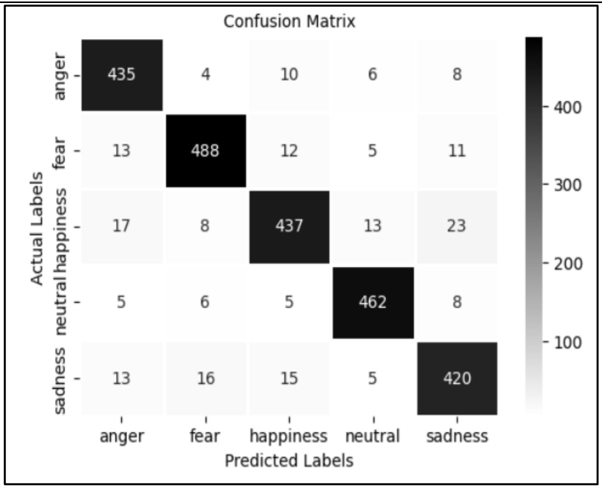


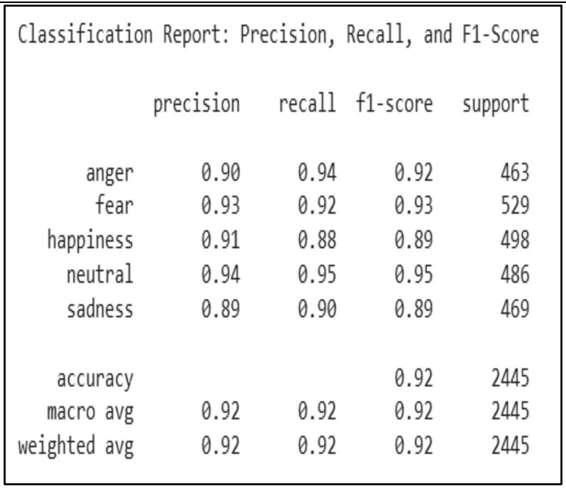
Figure 4.35. 3D CNN Performance Using Mel Spectrogram Features: Validation Loss and Accuracy Graphs on Combined Datasets (DEMoS, TESS, and RAVDESS)

As seen in Figure 4.35, the two graphs represent the training and validation performance of a 3D CNN trained with the audio features - Mel Spectrograms over 50 epochs. The loss graph shows the loss curves for both training and validation. The losses decrease sharply during the initial epochs, with the training loss continuing to decline steadily. The validation loss closely follows the training loss curve and stabilizes towards the later epochs, achieving a mean value of 0.592 (over the last three epochs), indicating good generalization. In the accuracy graph, the accuracy curves for training and validation increase rapidly during the early epochs. The training accuracy continues to improve consistently, while the validation accuracy follows a similar trend with slight fluctuations. By the final epochs, both curves stabilize, achieving a high validation accuracy of 0.831 (mean of the last three epochs). This demonstrates strong model performance with a good balance between training and validation accuracy, suggesting minimal overfitting and effective feature learning from the Mel Spectrograms. Figure 4.36 displays the Confusion Matrix and Classification Report for the combined Dataset (DEMoS, RAVDESS and TESS) using the Hybrid Average Model.

Further, to evaluate the impact of Gaussian noise augmentation, model performance was compared with and without the augmentation. With Gaussian noise, the accuracy scores were Model 1 – 59.30%, Model 2 – 91.37%, and Model 3 – 83.20%. In contrast, without Gaussian noise, the accuracies dropped to: Model 1 – 58.20%, Model 2 – 88.84%, and Model 3 – 68.68%. These results confirm that Gaussian noise augmentation contributed to improved performance across all models for the combined datasets - DEMoS, TESS, and RAVDESS. This justifies the choice of Gaussian noise as an effective augmentation strategy for preparing the model for real-world audio conditions.

Confusion Matrix and Classification Report





```
Classification Report: Precision, Recall, and F1-Score

              precision    recall  f1-score   support

       anger       0.90      0.94      0.92       463
        fear       0.93      0.92      0.93       529
   happiness       0.91      0.88      0.89       498
     neutral       0.94      0.95      0.95       486
     sadness       0.89      0.90      0.89       469

    accuracy                           0.92      2445
   macro avg       0.92      0.92      0.92      2445
weighted avg       0.92      0.92      0.92      2445
```

Figure 4.36 Confusion Matrix and Classification Report for the combined Dataset (DEMoS, TESS and RAVDESS) using Hybrid Average Model

The evaluation metrics using the Confusion Matrix and Classification Report for the combined Dataset (DEMoS, TESS and RAVDESS) using the Hybrid Average Model, as shown in Figure 4.36, indicate a high accuracy score of 92%.

4.4.5 Converting the SER System into a Real-time Application

The predictions from the three stream CNNs - 1D CNN, 2D CNN, and 3D CNN were saved as three pickle files named modelk1.pkl, modelk2.pkl and modelk3.pkl respectively. The trained deep learning models were stored as pickle files to facilitate testing the model (hybrid average) with real-time speech. This approach enabled quick application loading at runtime and faster responses to user requests. Additionally, saving the model as a pickle file allowed seamless dynamic updates without disrupting the application [191, 197]. Notably, pickle files are highly compatible with Flask, the Python web framework used to develop the web interface for our trained SER model. Pickle files can be easily saved and loaded using the pickle module functions dump() and load(), respectively. An independent application was created to integrate the trained pickle files with the graphical user interface. The idea is that, with the help of the interface, the educator can interact with the SER system, and the system can display the live results, which is the emotion recognised from the student's live voice/speech. This new application was created to run the live application meant to detect emotions via live utterance. First, all the required libraries were loaded. Next, the three pickle files were loaded which contain the predictions from the hybrid CNNs – 1D CNN, 2D CNN, and 3D CNN. Further feature preprocessing, feature extraction and classification were performed as discussed in section 3.4.2 earlier. The Score-level fusion combined the outputs from multiple models to get the final prediction during live testing.

The dynamic web application was built using Flask, a powerful framework provided by Python. Flask makes developing web applications very convenient [191]. Defining and linking URL routes to Python functions is much simpler with the Flask framework. It also makes handling various HTTP requests easier and enables the dynamic serving of content. Its built-in development server makes processing and debugging web applications locally easy, especially during development. Hence, our web application was built on Flask, which also handled the backend – which is the SER system. The library IPython was installed for real-time audio processing of the incoming speech stream. The real-time audio is received from the front end with the help of the written code. A Python code was written to handle and process the live audio file recorded via the front end. The recorded audio file gets saved in a specific directory. Then, the necessary processing is done to detect the emotion using the proposed emotion detection algorithm. The user interface element is automatically updated to show the emotion predicted in a user-friendly way.

Further, the web application running on a local server was hosted on the Internet with the help of Ngrok. Ngrok provides the network services to host one's website or a webpage [192]. Ngrok helps generate a temporary URL for the application on our local server. This intention is to host a small number of users at any point for small periods. After installing Flask and Ngrok, the Flask application was started locally first. Ngrok was executed in the same port number as where the Flask application was executing. Ngrok generated a URL that was shared publicly, making it convenient for everyone to access the Flash application running locally. The following section delves into the user interface aspect of the developed SER system.

4.4.6 Workflow for the SER Built with GUI

In this section, the process of using the SER interface is outlined. Educators could access the SER web interface via the provided Uniform Resource Locator (URL) on the webpage where SER is hosted. The SER system aims to support online teaching; it can be accessed with online teaching tools such as Zoom or MS Teams. By sharing the link to the online teaching tool, educators and students could participate in the session, which is essential for testing the SER system.

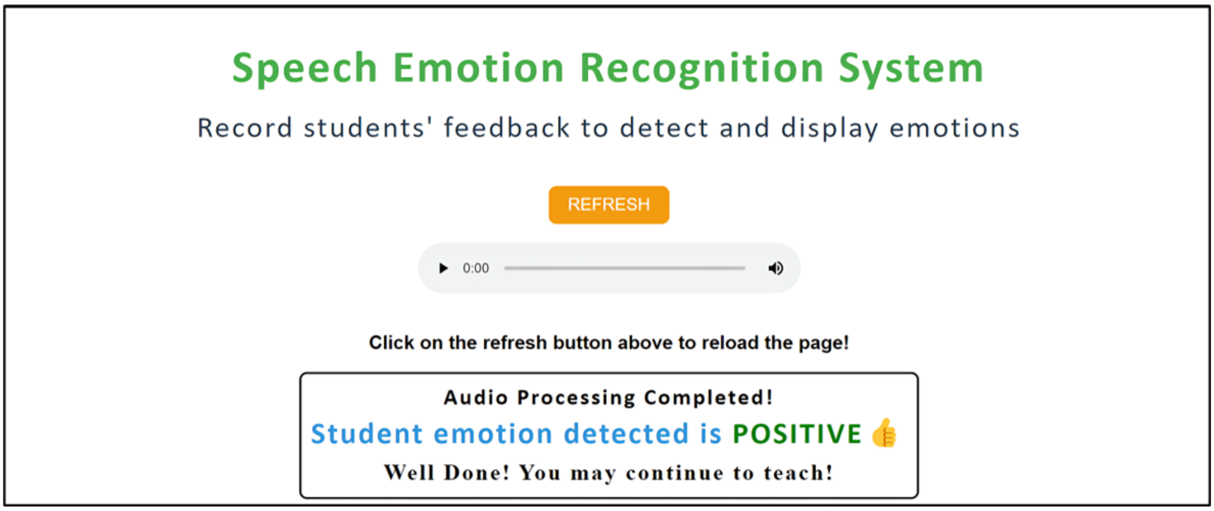
Detecting emotions as positive or negative:

Research shows that students' emotions are significantly influenced by their learning experiences, exhibiting positive or negative emotions based on their comprehension of the material being taught [205]. These emotions become apparent when students provide verbal feedback. The SER system is designed to detect the underlying emotions in students' speech during online classes and display the results as either positive or negative. Positive emotions, such as happiness, reflect favourable experiences and satisfaction, while negative emotions, such as sadness, anger, or fear, signal distress or dissatisfaction [206]. These emotional states are critical in shaping individual behaviour, decision-making, and social interactions, particularly in environments requiring effective communication.

The SER system developed in this study categorizes detected emotions as positive when it identifies happiness or neutrality, and as negative when it identifies sadness, anger, or fear. The resulting classification is then displayed to the educator as either positive or negative. The decision to categorize emotions as positive (happy, neutral) and negative (sad, anger, fear) was driven by both practical and functional considerations within the context of the study. While the developed SER system is capable of detecting five discrete emotions, converting these into a binary classification simplifies the interpretation of emotional feedback, particularly in real-time teaching scenarios. This approach not only enhances usability but also improves classification accuracy that occurs between closely related emotions grouped together. Binary categorization enables faster response times from educators, as they can quickly grasp whether a student's reaction is generally favourable or unfavourable without needing to interpret specific emotional labels. This is especially critical in online teaching environments where immediate instructional adjustments are essential. For instance, if a positive emotion is detected, the educator can continue teaching or proceed to get feedback from other students. However, if a negative emotion is detected, the educator may choose to re-explain the concept, provide additional examples, or offer more exercises to enhance the student's understanding and ensure clarity. This streamlined feedback mechanism is particularly beneficial for late-deafened educators, as it makes emotional cues more accessible, interpretable, and actionable without relying on the detail of the auditory input—by providing immediate insight into whether student feedback reflects overall understanding and engagement or if the student is struggling. This approach supports real-time instructional adjustments, making the emotional feedback actionable and accessible in an online teaching environment. This further aligns with the research scope of this study.

To implement this functionality, the SER system captures verbal feedback from students and displays the detected emotion to the educator. When an educator wishes to understand how a student feels about the ongoing class, they can prompt the student to provide verbal feedback. As described earlier, using the interface shown in Chapter 3, Figure 3.11, the educator clicks the 'RECORD' button to initiate the recording process. The student's voice is captured through the microphone. After providing feedback, the educator clicks the 'STOP' button to end the recording. The interface includes a playback feature, allowing the educator to replay the student's recorded feedback as needed. The educator clicks the 'SHOW EMOTION' button to activate emotion recognition from the student feedback. This feature processes the recording and provides an emotional assessment from the backend, where the neural network application is running to discern the emotion. When clicked, it displays a message which can be divided into two sections –

1. The first section message says – "Audio Processing Completed", which indicates that the processing of the captured live audio is completed.
2. The second section of the message gives information on whether the student's emotion detected is positive or negative. If a positive emotion is detected in the captured student vocal feedback, it gets displayed as depicted in Figure 4.37, which is "Student emotion detected is POSITIVE". Additionally, a message is provided to the educator: "Well Done! You may continue to teach!" This indicates that since the student feedback is positive, they may proceed with the class. Conversely, if a negative emotion is detected, it gets displayed as shown in Figure 4.38, which is "Student emotion detected is NEGATIVE". It also displays a message to the educator: "Kindly consider reexplaining or taking a different approach to what you are currently teaching or ask for specific feedback from the student!"



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

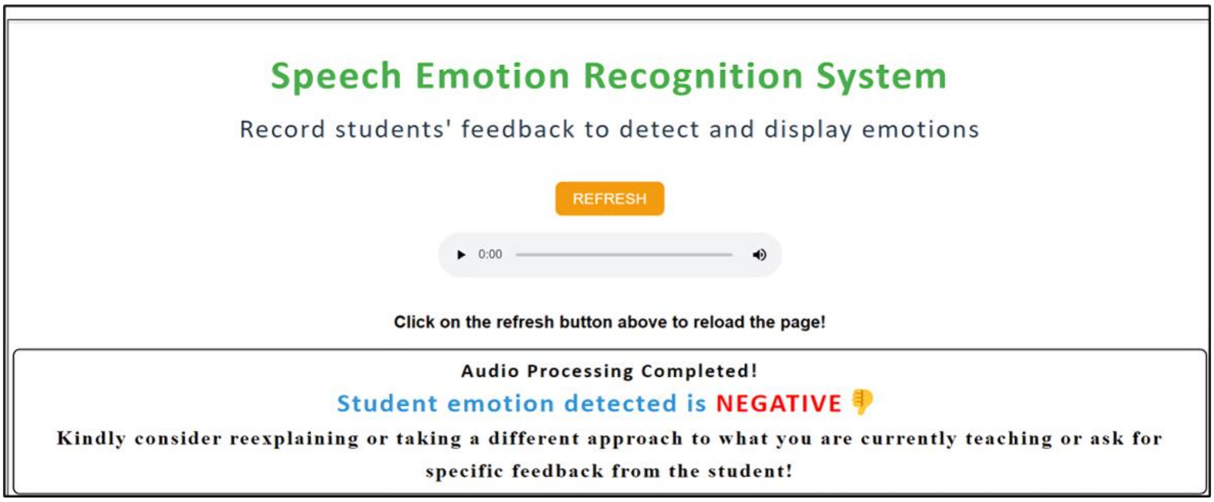
Click on the refresh button above to reload the page!

Audio Processing Completed!

Student emotion detected is POSITIVE 👍

Well Done! You may continue to teach!

Figure 4.37: Detection and display of positive emotion via the web interface



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

Click on the refresh button above to reload the page!

Audio Processing Completed!

Student emotion detected is NEGATIVE 👎

Kindly consider reexplaining or taking a different approach to what you are currently teaching or ask for specific feedback from the student!

Figure 4.38: Detection and display of negative emotion via the web interface

In this way, the SER application identifies emotions in the student's live verbal feedback and presents them through a user-friendly interface. The interface provides convenient features for capturing the live verbal feedback, playing and replaying the recorded feedback, and displaying the detected emotion. Additionally, it includes an option to refresh or reload the web page for seamless operation.

This concludes the ‘Design and Development of the User Interface’ section. Section 4.4 presented the results of integrating the selected SER model with the graphical user interface, highlighting its functionality as a real-time emotion detection application. This signifies the completion of stage 3 of the HCI research process for SER, as depicted in Figure 3.1. The corresponding section in Figure 4.39 has been shaded to illustrate this.

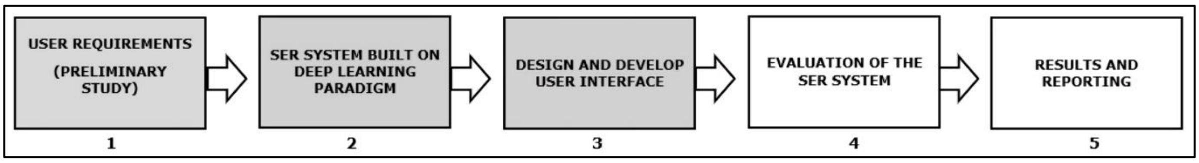

USER REQUIREMENTS (PRELIMINARY STUDY)
1
SER SYSTEM BUILT ON DEEP LEARNING PARADIGM
2
DESIGN AND DEVELOP USER INTERFACE
3
EVALUATION OF THE SER SYSTEM
4
RESULTS AND REPORTING
5

Figure 4.39: Updated process of the HCI research for the SER system

Furthermore, the results in Section 4.3 and 4.4 indicate that the third research objective has been successfully addressed, as stated below:–

RO3 - To design and develop a graphical user interface (GUI) that integrates SER-derived emotional feedback and displays emotions to late-deafened educators accurately and in real-time.

In the subsequent section, the evaluation of the SER system is discussed, focusing on usability testing and user experience. The aim is to achieve the 4th research objective, which is

RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency and perceived impact on late deafened educators for an effective engagement in online education.

4.5 Evaluation of the SER System

A preliminary study explored late deafened educators' challenges in discerning student emotions during online teaching. The findings, detailed in Section 4.2, provided valuable insights into these difficulties and emphasized the specific needs of these educators. The survey analysis highlighted the necessity for a speech emotion recognition (SER) system to support online teaching activities. This study focused on the real-world application of the developed SER system, with its evaluation described in this section. The methodology for evaluating the SER system is outlined in Section 3.6.

The SER system was tested in real-time and assessed for usability and user experience, targeting educators with and without hearing impairments. The evaluation examined the system’s effectiveness, efficiency, and perceived impact on late deafened educators. Usability testing was conducted with educators (both with and without hearing impairment) to determine whether the SER system addressed their challenges in discerning student emotions during online classes. This process involved selecting educators based on specific criteria, conducting system testing, and gathering feedback through questionnaires. The following analysis assessed the SER system's impact on educators’ ability to engage effectively with students and captured additional suggestions for improvement.

The aims of evaluating the developed SER system are:

1) To assess the effectiveness and efficiency of the developed SER system.
2) To obtain feedback on the perceived impact of the SER system on educators with and without hearing impairment. [3]

Achieving these aims would align with the accomplishment of our research objective, which is: RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency, and perceived impact on late-deafened educators for an effective engagement in online education.

To achieve the goals of measuring usability and evaluating the user experience of the SER system, Figure 4.40 illustrates the general flow of the evaluation process for the developed GUI-integrated SER system.

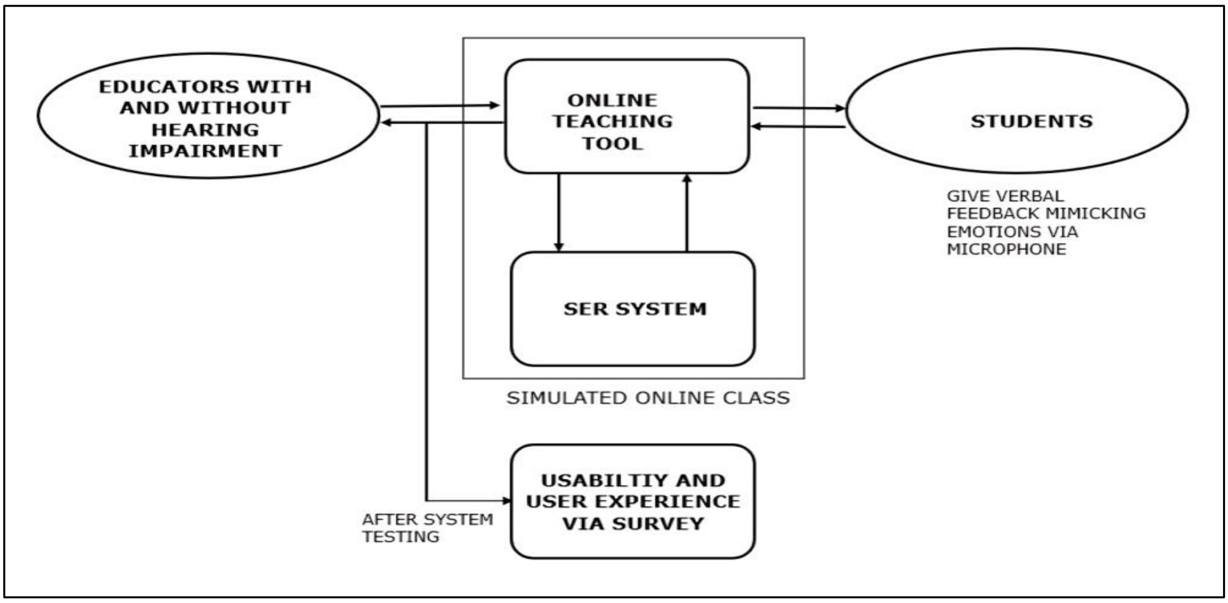

EDUCATORS WITH AND WITHOUT HEARING IMPAIRMENT
ONLINE TEACHING TOOL
STUDENTS
GIVE VERBAL FEEDBACK MIMICKING EMOTIONS VIA MICROPHONE
SER SYSTEM
SIMULATED ONLINE CLASS
AFTER SYSTEM TESTING
USABILTIY AND USER EXPERIENCE VIA SURVEY

Figure 4.40. Flow diagram of the evaluation process of the SER system

As shown in Figure 4.40, educators with and without hearing impairments tested the system remotely using the provided web interface (described in detail in Section 3.5). A simulated online teaching environment involving educators and students was created for this purpose. During the testing, students provided verbal feedback via their microphones, which the system analysed to detect and display emotions as either positive or negative on the interface. The evaluation focused on assessing usability and user experience from the educators' perspective.

The methodology for conducting the system evaluation is detailed in Section 3.6. The procedure for the system testing in the pilot phase by pilot users and the main phase by final

[3] *Wherever the term educators with hearing impairment is used, it refers to late-deafened educators, defined in Section 2.6.4.*

respondents has been outlined in Section 3.6.5. A survey with the respondents followed this to get feedback on their experience with the SER system. The survey was administered to the educators to collect structured quantitative data and qualitative data on various aspects such as demographic information (age, gender, hearing condition, and such), the effectiveness of the SER system, the efficiency of the SER system, and the perceived impact of the SER system on their teaching. The next sections provide an in-depth examination of the survey data, including the findings and the analysis. First, the pilot users' results, findings, and recommendations will be presented, followed by the results and analysis of the data collected from the final respondents. The following sections will address the final stage of the HCI research process: results and reporting.

4.5.1 Results, Findings and Recommendations for the SER system after the Pilot Phase

As outlined in Sections 3.6.4 and 3.6.5, six pilot users were chosen to assess the SER system. These users were selected using a purposive sampling technique based on the same criteria as the respondents for final testing. Five of the six pilot users were educators with normal hearing, which means they did not have any form of hearing impairment. One pilot user was an educator with a hearing impairment. After the completion of the usability testing of the SER system by pilot users, the link of the questionnaire was shared to gather quantitative data and qualitative data on various aspects such as demographic information (age, gender, hearing condition and such), effectiveness and efficiency of the SER system and the perceived impact of the SER system on their teaching. All the questions covered in the questionnaire are broadly described in Section 3.6.3.1.

Most survey outcomes have been represented in graphs, while some findings are presented descriptively.

4.5.1.1 Demographic Information

Table 4.7 presents the demographic information of the six pilot users. The Pilot User is represented as PU. As seen in Table 4.7, out of the six pilot users, four were females, and two were males. All had extensive teaching experience ranging from 15 to over 20 years. Most of them belonged to the age group 40 to 60 years. These educators specialized primarily in STEM (Science, Technology, Engineering, and Mathematics) and taught at higher education or university levels, with some also engaged in corporate training. The demographics of the pilot users below clearly demonstrate their suitability for participation in the study. The key criteria for purposive sampling for this study were selecting educators with at least four years of teaching experience and experience in teaching in higher education, university, or corporate training environments.

Table 4.7: Demographic information of the Pilot Users *(Pilot User = PU)*

Pilot User	Gender	Age Group	Years Of Teaching Experience	Area Of Teaching	Student's Level Taught by The Respondent
PU 1	Male	31 - 40	11-15 years	STEM	Others (Corporate training etc.)
PU 2	Female	41 - 50	16-20 years	STEM	Others (Corporate training etc.)
PU 3	Female	51 - 60	16-20 years	STEM	Higher Education or University
PU 4	Female	41 - 50	>20 years	Others	Higher Education or University
PU 5	Female	51 - 60	>20 years	Social Science / Business	Higher Education or University
PU 6	Male	41 - 50	16-20 years	STEM	Higher Education or University

4.5.1.2 Hearing Condition

The hearing condition of the pilot user (selected educators) is a crucial parameter for this research. It is another key criterion for the purposive sampling. 5 of the 6 pilot users were educators without hearing impairment, whereas 1 pilot user had moderate hearing impairment. Educators with moderate hearing impairment could hear without hearing aids but could not always understand the speech. This is shown in Figure 4.41 below. As per the responses in the questionnaire, all 6 pilot users could hear the students' verbal feedback during online teaching sessions.

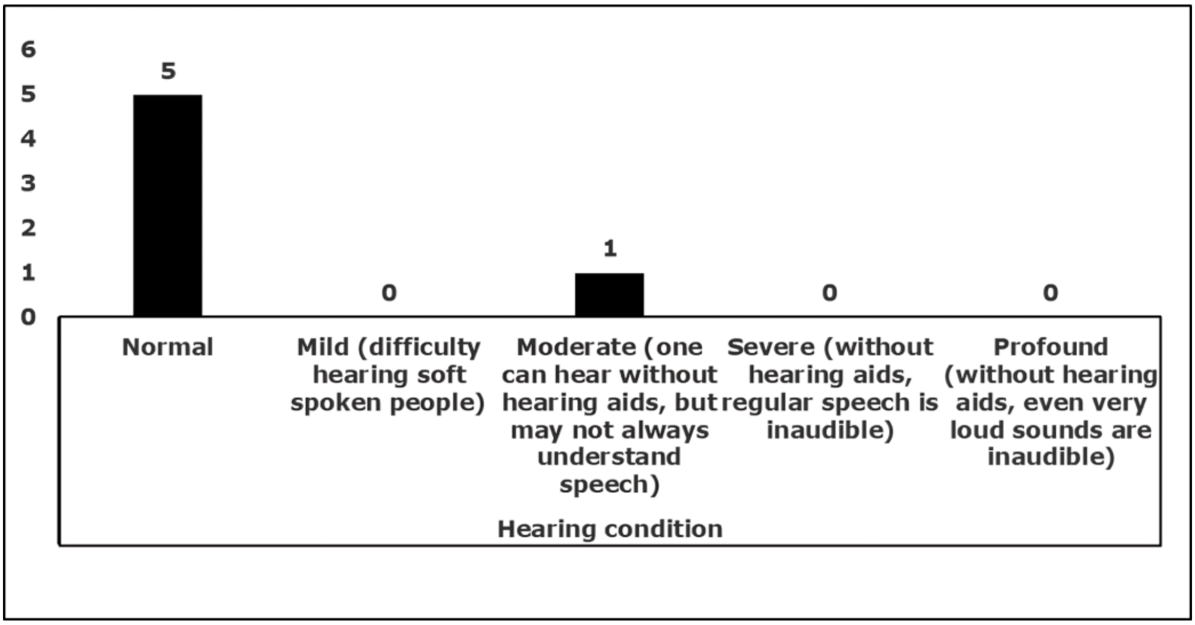


Figure 4.41. Hearing condition of the Pilot Users

(The degree of hearing loss is categorized based on available audiograms and classified according to the World Health Organization standards: Normal hearing (25 dB or better), mild impairment (26–40 dB), moderate impairment (41–60 dB), severe impairment (61–80 dB), and profound impairment (81 dB or higher) [209, 210])

4.5.1.3 Usage of Online Systems and Significance of Student Feedback during Online Classes

Data collected from the pilot users on the usage of online systems indicated that each one was familiar with online teaching and had conducted online classes frequently in the past three years. This was another key criterion for purposive sampling. It clearly shows the suitability of pilot users for our usability study, as all have experience teaching online. Each has taught using different online teaching tools like Zoom, MS Teams, and such. Each educator selected as a pilot user strongly agreed on the importance of real-time student feedback during online classes for enhancing student learning and engagement, as shown in Figure 4.42.

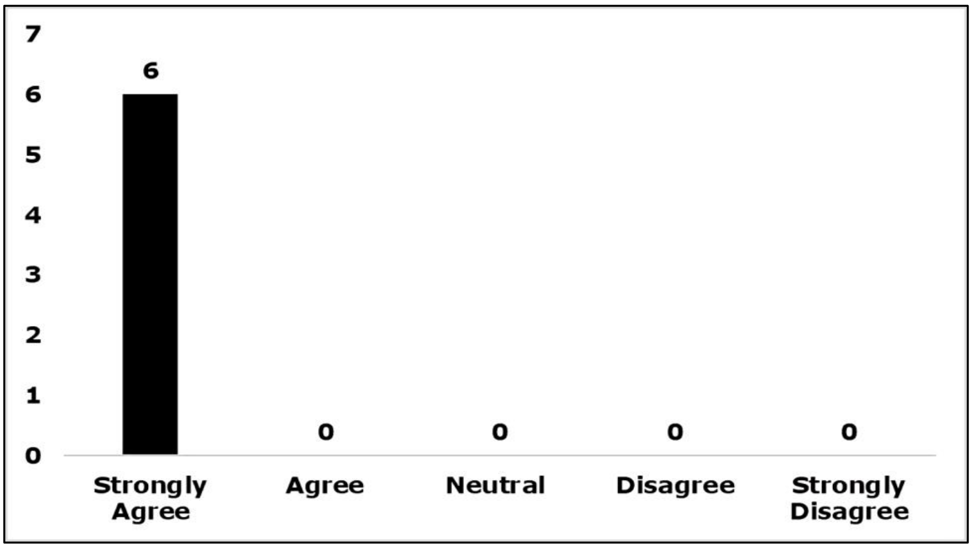


Figure 4.42: Importance of student feedback in online classes for their learning and engagement by pilot users

As seen in Table 4.8, each educator (Pilot User) seems to be using different methods to collect feedback from students during online classes. Some used a chat tool or a feedback form or asked the student to email their queries. Others would ask students to switch on their videos and give feedback where their facial expressions can be seen or verbal feedback heard via the microphone. The data in Table 4.8 further reveals that educators (Pilot Users) often struggled to confidently determine students' emotions, especially when the students' camera/video was off. This indicates that educators require significant effort to assess emotional responses to student's feedback. Each educator shared their challenges in discerning student emotions in online classes.

Table 4.8: Usage of online systems and significance of student feedback during online classes *(PU Pilot User)*

Pilot User	Tools used for collecting Student Feedback in Online Class	Ability to Discern Emotions (Positive/Negative) and Challenges Faced
PU 1	Through Chat, Call and Email.	Yes, I could tell if their emotion were **positive** or **negative**.
PU 2	Feedback form and **verbal** feedback.	**Negative** feedback is easier to tell.
PU 3	Using chat box and **verbal** communication.	**Sometimes,** only I could gauge.

PU 4	By posing questions and **asking** the students to keep the video and audio switched on when answering.	I often could sense the underlying emotion, positive or negative, in the student's voice. However, when the video is turned off, it becomes challenging to gauge their feedback and understand whether they understand the material or not.
PU 5	chat, Slido, Quizz, Kahoot.	I could **seldom** tell.
PU 6	Just **checking** with students during the online session.	**Yes, difficulties because** of poor network connection sometimes.

4.5.1.4 Purpose

As shown in Figure 4.43 below, all the pilot users understood the purpose of the developed SER system. This understanding was based on the provided demonstration and the guidelines shared before the usability testing.

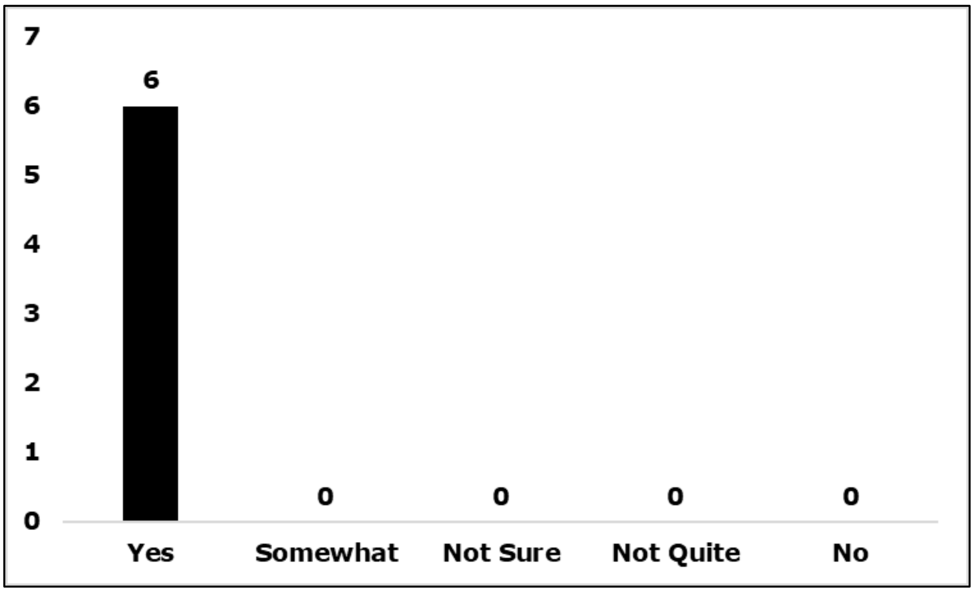


Figure 4.43: Purpose of the developed SER system

4.5.1.5 Usability - Effectiveness and Efficiency of the SER System

Effectiveness and Efficiency are defined in the methodology chapter's introduction section, Section 3.6.1. As evidenced in Figure 4.44 below, favourable responses from all six pilot users regarding the user-friendly user interface also highlight an aesthetically pleasing layout. Controls for recording and features like replay of the student recording are intuitive, making it much easier for the users to navigate. Most importantly, an affirmative response concerning the detected student emotion being displayed as positive or negative, along with an emoji, indicates the effectiveness of the SER system.

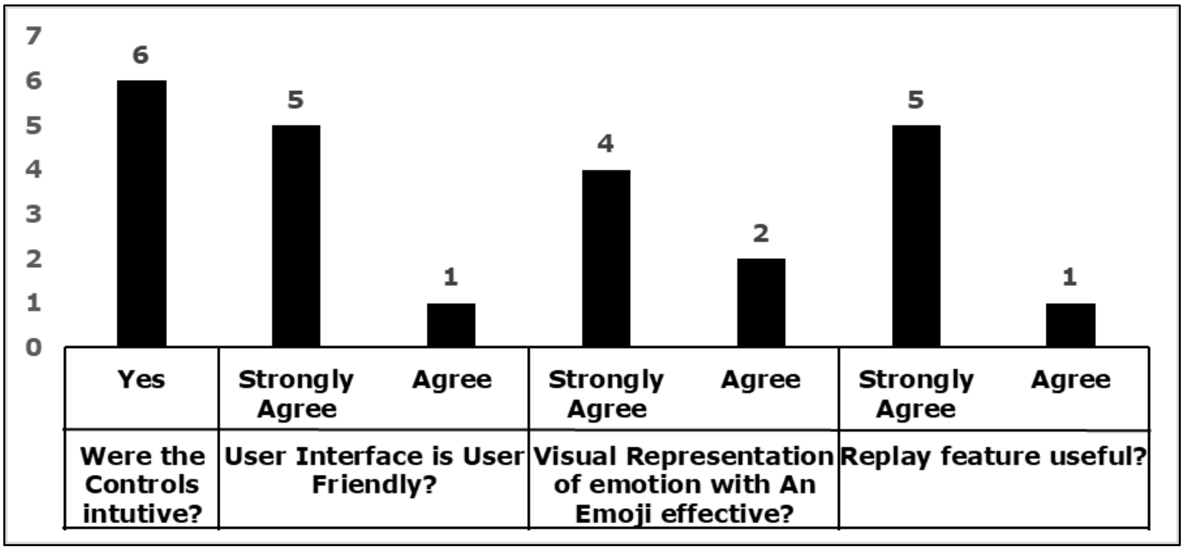


Figure 4.44: Feedback on the features of the SER system by pilot users

The SER system identifies the emotion in the student's verbal feedback as positive or negative. When the backend algorithm detects a happy or neutral emotion, it displays a message which states that a positive emotion has been detected with the thumbs-up emoji and when it detects a sad, angry or fearful emotion, it displays a message saying that a negative emotion has been detected with a thumb's down emoji. The effectiveness of the developed SER systems lies in accurately categorising and displaying the emotions as positive or negative. This is crucial for the efficacy of the SER system in aiding educators' understanding of student sentiments during online classes. As indicated in Table 4.9 below, the system demonstrates a high accuracy rate in correctly detecting emotions. It effectively identifies negative emotions from students' voices; however, there were occasional instances of incorrect detection in categorising happy emotions as negative. Some pilot users suggested adding more training samples to make the system more efficient in correctly detecting positive emotions. Some suggested that a very high pitch could be one of the reasons for emotions like happiness being detected as negative (as it is detected as an angry emotion).

Table 4.9. Feedback on the functionality of the SER system (*PU - Pilot User*)

Respondent	The system was able to detect and display the real-time emotion (positive/negative) of the student's vocal feedback correctly.	Any instances where the emotions were not correctly interpreted? if so, please provide some specific comments on your assessment. ?	Any suggestions for enhancing the efficiency and effectiveness of the speech emotion recognition system?
PU 1	Yes, it was able to detect **correctly**.	No. There were **no such instances**.	**Add more** training samples to deal with diversity of vocals
PU 2	Yes, it was able to **show** the real time emotion as positive/negative.	Negative is **correctly** interpreted; positive emotions were getting detected as **negative sometimes**.	Sometimes Happy emotions are showing as negative by the system as it has similar pitch as angry. This can be **improved**.
PU 3	Negative emotions were **correctly** detected.	Few Positive emotions were **not** being interpreted correctly.	**No** suggestions.
PU 4	The system's accuracy rate in detecting and presenting the real-time emotional tone (positive/negative) of the student's vocal responses was nearly **95%.**	The majority of the responses were accurately detected. However, I believe the **incorrect detections** may have been caused by the trembling or shakiness in the students' voices, particularly during positive comments.	Yes. Think about how people from different places might express emotions differently. May include like **facial expressions or body movements**, along with the voice.

PU 5	**Not precisely**. Some of the positive emotions were being recognised as negative.	Sometimes the positive emotions were recognised **wrongly** as negative.	Positive emotions must be **recognised** correctly.
PU 6	Yes, it displayed **correctly**.	No. There were **no wrong** detections.	Not now. System is working **fine.**

In conclusion, the pilot user seems to have found the SER system's web interface visually appealing, intuitive, accessible, and responsive. By prioritizing these aspects, the system enhances user engagement and satisfaction, contributing to the success of online teaching endeavours. Moreover, this aspect was effectively achieved based on the experience the pilot users recorded.

4.5.1.6 User Experience - Perceived Impact of SER on the Pilot User

Efficiency and effectiveness are important factors in the usability of a system and can contribute to a good user experience. A positive user experience occurs when the user finds the system enjoyable and satisfying. The system must be appealing and valued by the user. User experience evaluation involves assessing users' perceptions and emotions when interacting with the system [170, 173, 174, 211]. Some pilot users' feelings are described after interacting with the SER system, particularly regarding its impact on their teaching experience.

When each pilot user strongly agreed that a negative emotion detected in a student's feedback through the SER system helps adjust teaching strategies in an online class, it suggests an encouraging perceived impact on personalized instruction. This is evident from the data gathered from pilot users and presented in Figure 4.45. It also displays a unanimous strong agreement that a positive emotion detected in student feedback reinforces confidence in the educator to continue with the ongoing class. It indicates a perceived positive impact on the educator's morale to continue with the class. As seen in Figure 4.45, every pilot user has strongly agreed to recommend using the SER system for online classes. This indicates a clear benefit in enhancing teaching effectiveness. This suggests that the system gives an overall perception of student involvement in the class. It also indicated that student feedback is valuable for understanding their comprehension level in the learning process.

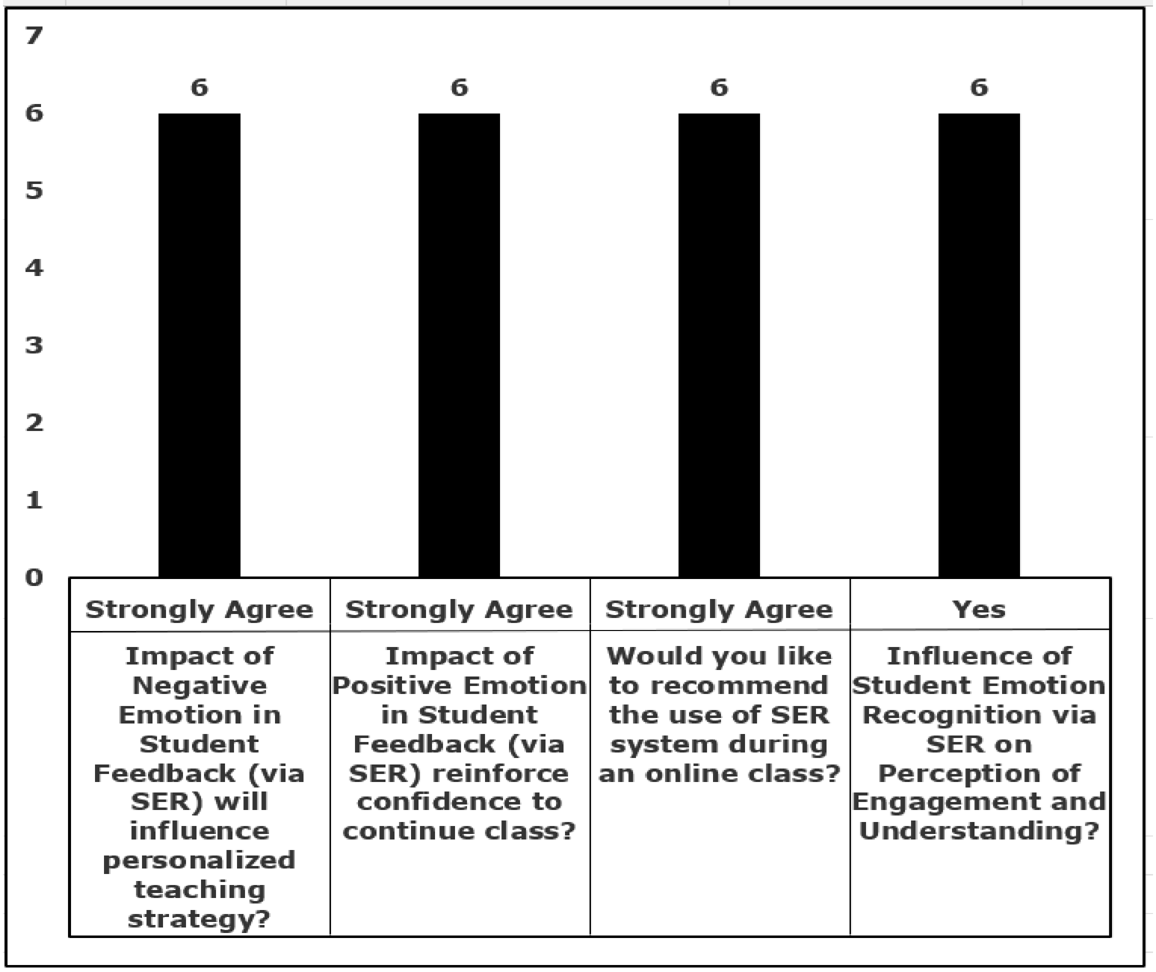


Figure 4.45. Perceived impact of SER on pilot users

4.5.1.7 Further Comments or Improvements Needed in the SER System

As seen in Table 4.10 below, all the pilot testers have unanimously declined any further enhancements to the system. They felt the system was sound and only needed to be tested by more students.

Table 4.10: Comments or improvements in the SER system (*PU - Pilot User*)

Respondent	Any other feedback or comments?
PU 1	The system is **good**.
PU 2	**Wonderful** app. Helps the online trainers specially if you have hearing issues.
PU 3	No feedback as the system is **good**.
PU 4	The system helps us grasp students' emotions through their voices, which is **useful.** This can **enhance** teaching by encouraging student participation and improving their performance.
PU 5	It would be better to test the system in an online class with many students.

PU 6	No feedback or comments. System is **good** as it is.

4.5.1.8 Overall Summary, Insights and recommendations from the Pilot Users

Testing by the Pilot users, followed by their feedback, was essential for identifying areas of improvement for refining the system's performance. These suggestions were implemented to enhance the developed SER system's efficiency and effectiveness in correctly displaying emotions. This would ultimately improve its utility for educators in online teaching environments.

Based on the results and their analysis from the usability testing of the developed SER system by the pilot users, the findings revealed several key insights. Below are the collective recommendations from the pilot users:

1) The user interface was found to be visually appealing and easily navigable.
2) The SER system was observed to integrate seamlessly with any online teaching tool.
3) The system was noted to promote user engagement and understanding.
4) It was determined that the system adds value by supplementing teaching approaches.
5) It was recommended that additional training data be incorporated to overcome occasional misclassifications of positive emotions as negative.
6) It was suggested that the negative emotion "Anger" should be removed.

The systems' interface was designed to be intuitive and user-friendly, enabling educators to navigate and utilize the tool easily during online classes. The pilot users agreed with the interface as reflected in their recorded user experiences, which were documented in the provided questionnaire. The real-time emotion detection feature was particularly effective, mostly identifying student emotions and allowing educators to adjust their teaching strategies. Integration with popular online teaching platforms would be seamless, as they could access the backend SER system via a user-friendly web interface. The system's emotion visualization was clear and straightforward, making it easy for educators to interpret and respond to student's emotional states. Feedback from pilot users has been overwhelmingly positive, highlighting the system's ability to enhance student engagement and understanding. The SER system would significantly improve teaching practices and student learning outcomes by providing valuable insights into student emotions during online classes**.** The pilot user data captured via the Google form was downloaded as an Excel sheet (Refer to Appendix 7 for sample pilot user data).

However, due to occasional misclassifications of positive emotions as negative and some pilot users suggesting additional training data to be incorporated, the dataset was expanded by two additional datasets: CREMA-D and EMODB. As recommended, the emotion ‘anger’ was removed. The positive emotions considered were ‘happy’ and ‘neutral,’ while the negative emotions were ‘sad’ and ‘fear.’ Five datasets were used to retrain the model (hybrid average). Figure 4.46 below illustrates the combination of the five selected datasets and displays the sample count for each emotion.

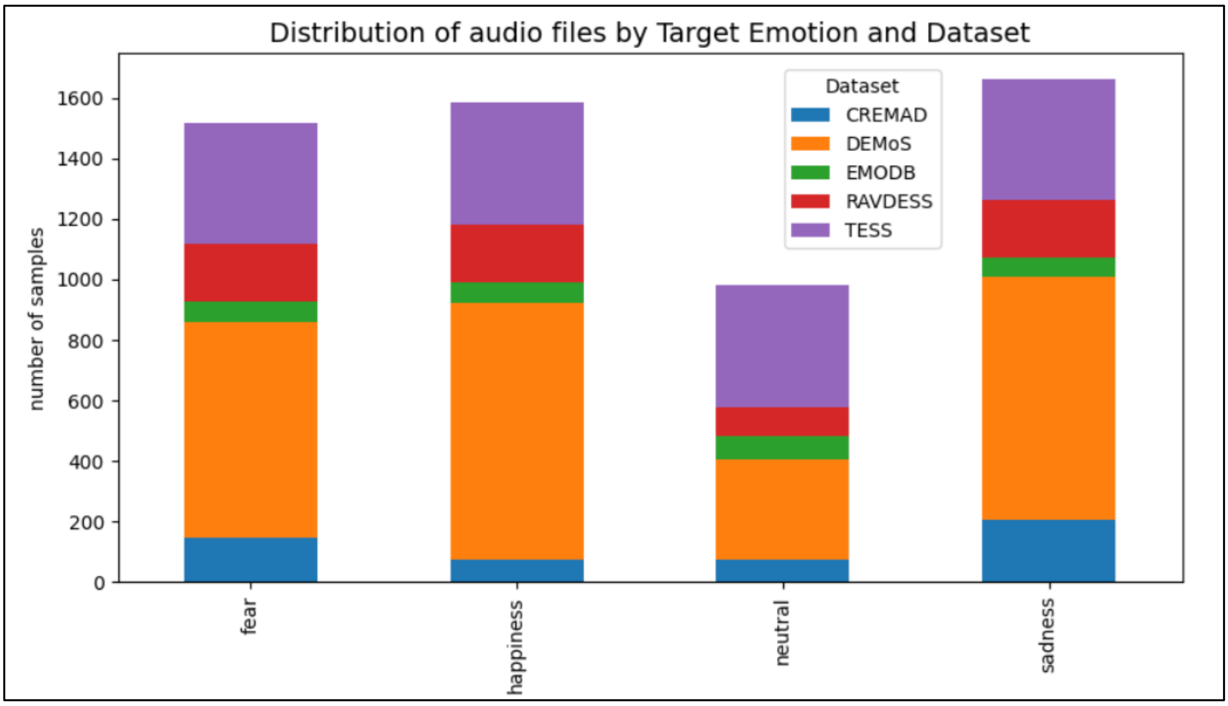
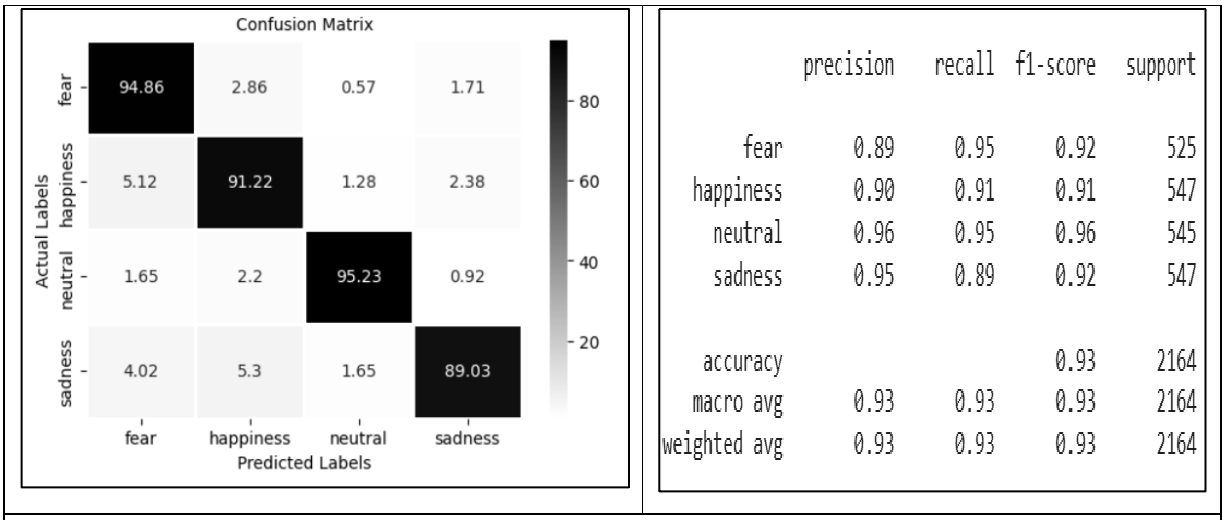


Figure 4.46 Combined distribution of audio files by target emotion across DEMoS, RAVDESS, CREMA-D, EMODB and TESS datasets

According to the researcher’s feedback documented in an Excel sheet (for the template, Refer to Appendix 8) and in the questionnaire, pilot users strongly suggested that students rarely exhibit ‘angry’ emotions during online classes. As seen in section 3.6.6 above, to display an emotion of anger, the student needs to raise the pitch, speak with a louder and more forceful tone, increase the pace slightly, and use assertive and direct language. As per one of the pilot users, to maintain the decorum of the class, no student would display such behaviour in an online class towards the educator. As a result, they requested the removal of the 'angry' emotion category from testing, leaving only four emotion categories: ‘Positive’ (comprising happy and neutral) and ‘Negative’ (comprising sad and fearful), as seen in Figure 4.46. In this case, the data was balanced, as described in Section 4.3, bringing the total number of audio files to 6656 for the 4 emotions: fear, happiness, neutral, and sadness. Figure 4.47. Displays the Confusion Matrix and Classification Report for the combined Dataset (DEMoS, RAVDESS, CREMA-D, EMODB and TESS) using the Hybrid Average Model.



```
              precision    recall  f1-score   support

        fear       0.89      0.95      0.92       525
   happiness       0.90      0.91      0.91       547
     neutral       0.96      0.95      0.96       545
     sadness       0.95      0.89      0.92       547

    accuracy                           0.93      2164
   macro avg       0.93      0.93      0.93      2164
weighted avg       0.93      0.93      0.93      2164
```

Figure 4.47. Confusion Matrix and Classification Report for the combined Dataset (DEMoS, RAVDESS, CREMA-D, EMODB and TESS) using the Hybrid Average Model

4.5.2 Results and Analysis of Respondent Data (Educators with and without Hearing Impairment)

The above section discusses the pilot users' findings and recommendations for the SER system. In this section, the results and analysis based on the survey data collected from the complete set of respondents who are educators with and without hearing impairments are presented. As outlined in the methodology chapter, Section 3.6.4, using purposive sampling, 20 respondents were recruited; 10 were educators with hearing impairment, and 10 were educators without hearing impairment. After the completion of the usability testing of the SER system, the link to the questionnaire was shared with the respondents to gather quantitative data and qualitative data on various aspects such as demographic information (age, gender, teaching experience and hearing condition), effectiveness and efficiency of the SER system and the perceived impact of the SER system on their teaching. The questionnaire was structured to offer valuable insights into all these components. As mentioned in methodology Section 3.6.5, the quantitative data was analysed using descriptive statistics, specifically through frequency counts. Qualitative data were analysed using content analysis.

4.5.2.1 Respondents Demographic/Profile Information

As seen from the respondent’s demographic information presented in Table 4.11, out of the 20 educators, 10 were educators with hearing impairment, whereas the remaining 10 were educators without any form of hearing impairment (within a normal range of hearing capability).

1) Educators with hearing impairment

With the help of purposive sampling and snowballing methods, ten respondents with hearing impairment were recruited, ranging from mild to profound. Of the ten, 60% were females and 40% were males. Most of these educators, 70%, had extensive teaching experience, ranging from 15 to over 20 years and more. And a handful of them, 30% had teaching experience between 1 to 10 years. Most (80%) belonged to the age group of 41 to 60 years, and a small percentage of 20% were below 40. Additionally, 50% specialized in STEM (Science, Technology, Engineering, and Mathematics) and 50% in social sciences or corporate training. 60% taught in higher education or at the university level, and the rest of the 40% gave corporate training. The demographics of the respondents above clearly demonstrate their suitability for the usability testing of our SER system, as discussed in Section 3.6.4.

2) Educators without any hearing impairment

For educators without hearing impairment (normal hearing), 10 respondents were recruited with the help of purposive sampling. Of ten respondents, 50% were females and 50% were males. 50% of the educators had teaching experience ranging from 15 to 20 years and more. And 50% had teaching experience between 1 and 10 years. 40% belonged to the age group of 41 to 60 years, another 50% belonged to the age group between 30 to 40 years of age, and just 10% were above 60 years. Additionally, 60% specialized in STEM (Science, Technology, Engineering, and Mathematics) and 40% in social sciences or corporate training. 90% taught in higher education or at the university level and the rest of the 10% gave technical and vocational training. The demographics of the respondents above clearly demonstrate their suitability for the usability testing of our SER system, as discussed in section 3.6.4.

Table 4.11: Respondents' demographic/profile information

Criteria	Educators with Hearing Impairment	Educators without Hearing Impairment
Gender Details		
Male	4 (40%)	5 (50%)
Female	6 (60%)	5 (50%)
Age Group		
Less than 30 years	1 (10%)	0 (0%)
30 to 40 years	1 (10%)	5 (50%)
41 to 50 years	5 (50%)	3 (30%)
51 to 60 years	3 (30%)	1 (10%)
Above 60 years	0 (0%)	1 (10%)
Teaching experience		
More than 20 years	3 (30%)	1 (10%)
16 to 20 years	3 (30%)	3 (30%)
11 to 15 years	1 (10%)	1 (10%)
1 to 10 years	3 (30%)	5 (50%)
Field of teaching		
STEM (Science, Technology, Engineering, and Math)	5 (10%)	6 (60%)
Social sciences and business	2 (10%)	3 (30%)
Others	3 (10%)	1 (10%)
Educational level taught		
Higher Education or University	6 (60%)	9 (90%)
TVET (Technical and Vocational Education and Training)	0 (0%)	1 (10%)
Others (Corporate training etc.)	4 (40%)	0 (0%)

4.5.2.2 Respondents Hearing Condition

1) Educators with hearing impairment

The hearing condition of educators is a crucial parameter for this research. It is one of the key criteria for our purposive sampling. All respondents in this category exhibited some form or percentage of hearing loss. As stated in Figure 4.48, among the ten educators shortlisted for our study, 40% had mild hearing impairments, 30% had moderate hearing impairments, 10% had profound hearing impairments, and 20% were severely deaf. This is displayed in Figure 4.49 below. According to the data shown in Figure 4.49, 50% of respondents with mild and moderate hearing impairments could hear student vocal feedback during online classes. In comparison, the other 50% either could not hear clearly or required a hearing aid.

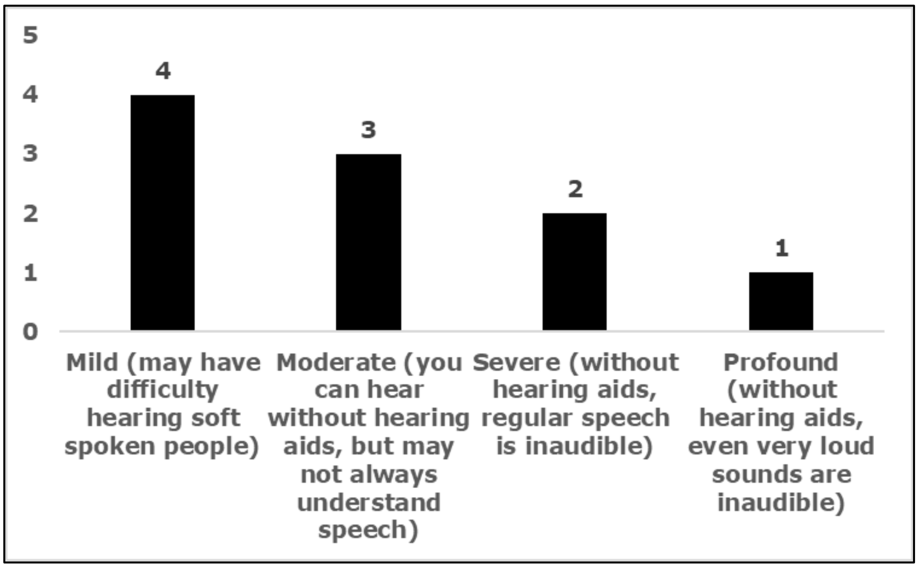


Figure 4.48: Respondents with different ranges of hearing impairments

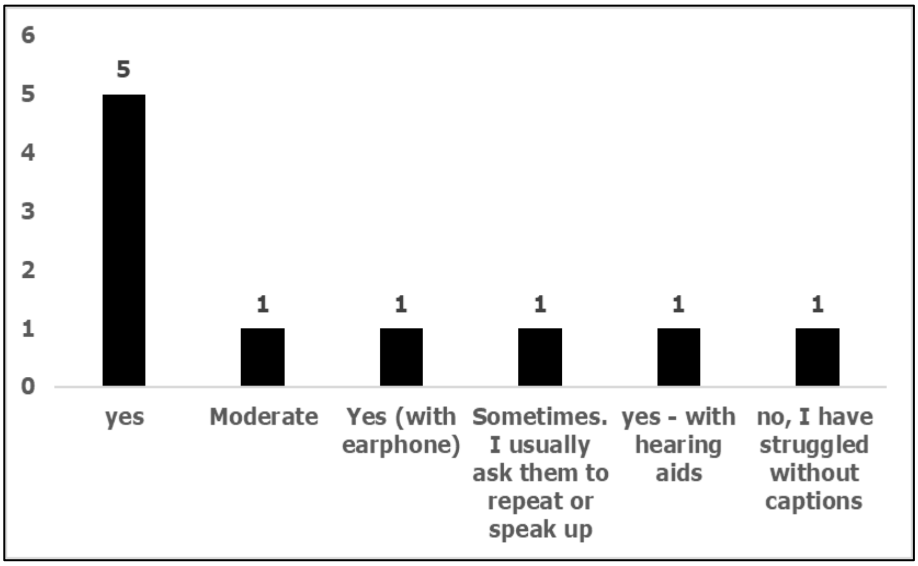


Figure 4.49: The ability of the respondents to hear student feedback based on their current hearing condition

2) Educators without hearing impairment

All the 10 respondents were educators without any hearing impairment. They could hear the student's verbal feedback during online teaching sessions.

4.5.2.3 Usage of online systems and significance of student feedback during online classes

1) Educators with hearing impairment

The data collected from the respondents on the usage of online systems clearly shows their suitability for our usability study, as all have experience teaching online. They have frequently conducted online classes in the past three years using different online teaching tools like Zoom, MS Teams, and such.

All respondents, despite their hearing impairment ranging from mild to profound, agreed (some strongly) on the importance of real-time student feedback during online classes for enhancing student learning and engagement, as seen in Figure 4.50 below.

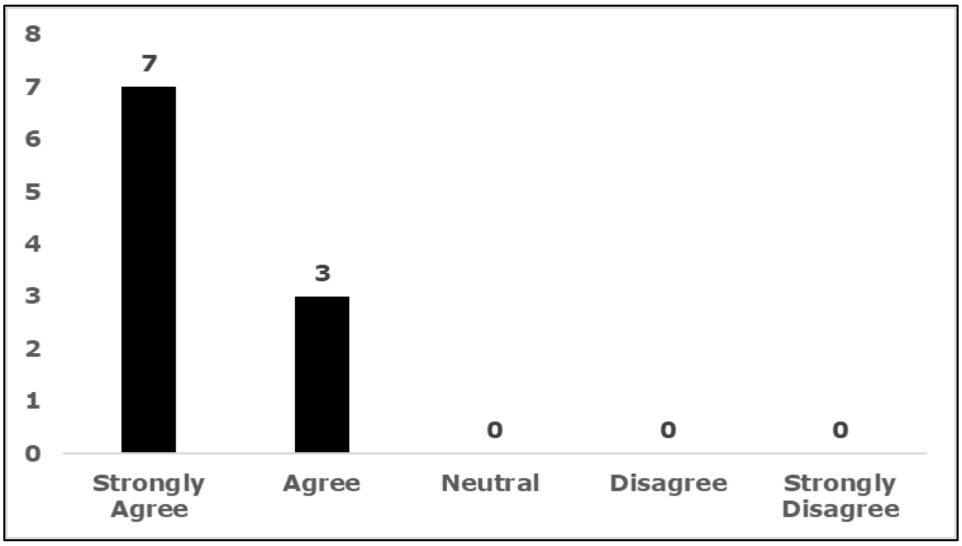


Figure 4.50: Importance of student feedback in online classes for their learning and engagement by educators with hearing impairment

As shown in Table 4.12, educators employed various methods to gather student feedback. Each respondent is represented with the letter 'R'. The data also reveals that educators often struggle to determine students' emotions, which becomes even more pronounced when students' cameras are switched off. Each educator reported difficulties in reliably discerning student emotions.

Table 4.12: Use of tools and the significance of student feedback during online classes by educators with hearing impairment *(R=Respondent)*

Respondent	Tools used for collecting Student Feedback in Online Class	Ability to Discern Emotion (Positive/Negative) and Challenges Faced
R1	**Verbal** feedback via microphone.	I could tell the **Positive** ones only.
R2	By **asking** question or giving a quiz.	Yes, I could tell. But **not always**.
R3	I usually have a **Q&A** slot at the end of the class.	**Unlikely**. Emotions cannot be determined from sounds only.
R4	Taking feedback **verbally**.	Yes, can discern information but **challenging** if the students do not put their camera on (so lack visual clues).
R5	**Online** Feedback Forms.	**Negative** Feedback would more like sad.
R6	May ask them to unmute and **respond** or post in Chat.	**Most often** yes. But at times would ask for reclarification.
R7	Through feedback **forms**.	Yes, i was able to tell when they are **happy**.
R8	Using Chat or **Offline mode**.	No, it was a **struggle**.
R9	Question and answer **verbally** and writing comment.	If the camera is switched off, **difficult** to tell which emotion.
R10	**Verbally**, but also looking at facial expressions if the camera is switched on.	Mostly I **would not** be able to tell.

2) Educators without any hearing impairment

The data collected from the respondents on the usage of online systems clearly shows their suitability for our usability study, as all have experience teaching online. Most have conducted online classes at least ten times in the past three years using online teaching tools like Zoom, MS Teams, and such.

Most of them have either strongly agreed or agreed on the importance of real-time student feedback during online classes for enhancing student learning and engagement, as seen in Figure 4.51 below.

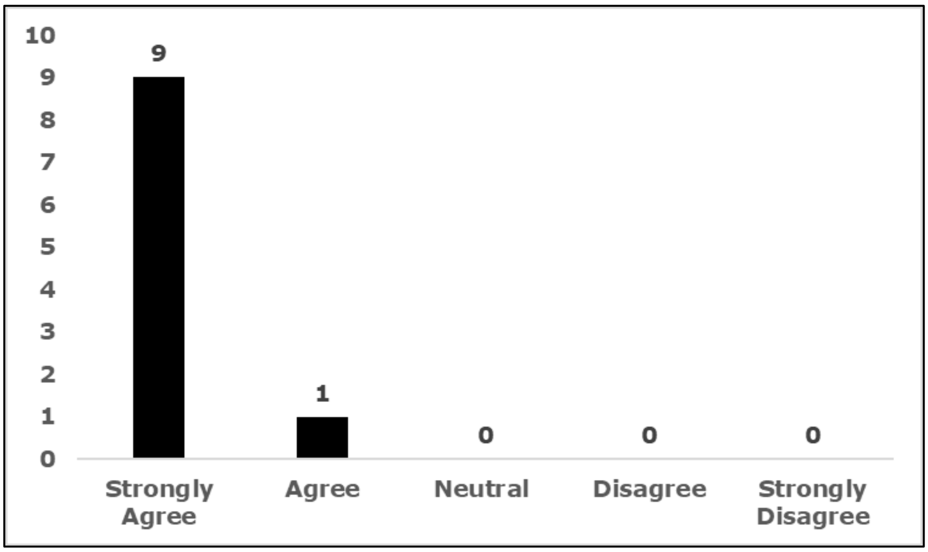


Figure 4.51: Importance of student feedback in online classes for their learning and engagement by educators with hearing impairment

As shown in Table 4.13, educators employ various methods to gather feedback from students, whether through verbal feedback or other means. The data reveals that educators often struggle to determine students' emotions, which becomes even more pronounced when students' cameras are switched off. Some educators reported difficulties in reliably discerning student emotions during online classes.

Table 4.13.Use of tools and the significance of student feedback during online classes by educators without hearing impairment *(R=Respondent)*

Respondent	Tools used for collecting Student Feedback in Online Class	Ability to Discern Emotions (Positive/Negative) and Challenges Faced
R1	I take **verbal** feedback. If I can't hear them well, I will ask them to write in chat.	**Yes**, I am able to know their emotion. They don't really show their emotions even if they don't understand.
R2	Virtually right after my class via **verbal** feedback.	**Yes**. I did not face any challenge.
R3	From the chat box and **verbal** feedback.	Some students **won't** express their emotions properly and as well verbally.
R4	using the zoom chat or **verbal** feedback	**Not** always, as the sounds may be unclear.
R5	By asking direct feedback from students and student's **survey**.	**Yes**, I can tell most of the times.
R6	Student **Evaluation** on Online Teaching at the end of the semester	**Yes**, I am able to understand to **certain** level

R7	**Directly** by asking them to speak up using microphone.	**Yes**.
R8	Poll, Chat, **Voice** feedback.	**Sometimes** only, it is highly dependent on the **network condition**.
R9	By asking their **verbal** feedback to know if they are clear.	**Yes**. At times, it is not audible due to **network issues**.
R10	Feedback via **chat**.	**Sometimes** based on student's sounds and voice.

4.5.2.4 Purpose of the SER System

All the educators, both with and without hearing impairment, irrespective of their hearing condition, understood the purpose of the SER system. This was also based on the provided demonstration and the guidelines shared before the usability testing. This is an important parameter as our usability testing is based on the understanding, usage and usefulness of the SER system.

4.5.3.5 Usability - Effectiveness and Efficiency of the SER System

The objective of usability is discussed in detail in Section 3.6.1 of the methodology chapter. A visually appealing and intuitive interface enhances the overall user experience, making it easier for educators to interact with the system during testing.

1) Educators with hearing impairment

As evidenced in Figure 4.52 below, responses of the hearing-impaired respondents indicate that they found the user interface to be user-friendly, highlighting its clear and concise layout. 90% of educators with hearing impairments agreed that the controls for recording and features such as replaying student recordings were intuitive, making navigation easier for users. Most importantly, the affirmative response regarding the detected student emotion being displayed as either positive or negative, along with an emoji, indicates the effectiveness of the SER system.

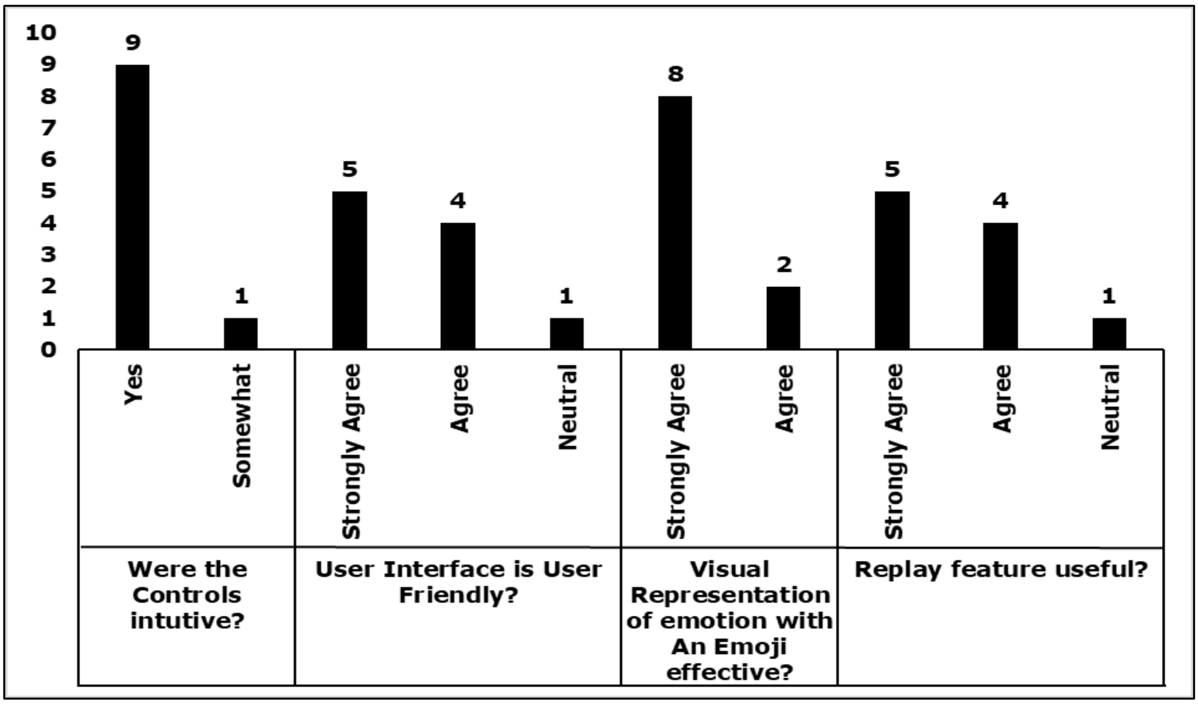


Figure 4.52. Feedback on the features of the SER system by educators with hearing impairment

The survey included some open-ended questions to encourage users to provide written feedback on the functionality of the SER system. Content analysis was used to analyse the qualitative data gathered through open-ended questions. Content analysis is a widely recognized method for data analysis of textual data [178]. It involves systematically examining the responses to identify and interpret patterns or themes. It allows researchers to quantify the presence of certain words, themes, or concepts in qualitative data and provides a systematic approach to understanding open-ended responses [179, 180]. Discussed below are the open-ended questions and their analysis.

a) The first question aimed to gather respondents' opinions on whether the system accurately identified and displayed real-time emotions (positive/negative) from student vocal feedback. This feedback from educators was crucial as it directly reflected the system's efficacy in recognizing student emotions. Using a content analysis approach, respondents' answers were categorized as 'positive' or 'mixed feedback,' as shown in Table 4.14. The 'positive feedback' classification was based on responses indicating that the system correctly detected emotions, while the 'mixed feedback' classification was based on responses indicating some issues with detection. Most respondents (80%) reported that the system could accurately detect and display real-time emotions from students' vocal feedback. However, 20% of respondents provided mixed feedback, and none gave negative feedback. This suggests that most educators with hearing impairments felt the system correctly detected student emotions.

b) The next question aimed to identify specific instances where the SER system did not correctly interpret students' emotions. Using a content analysis approach, respondents' answers were categorized as either 'no issues' or 'minor issues,' as shown in Table 4.14. The 'no issues' classification indicated no instances of incorrect emotion interpretation, which was reported by 50% of the hearing-impaired educators. The 'minor issues' classification was based on responses indicating some challenges with emotion interpretation, as reported by 50% of respondents. The issues were attributed to unstable internet, students failing to emote correctly despite the guidelines, or when the students mumbled with low volume. However, no educator with hearing impairment reported any major issues

c) The final question was directed at respondents asking for suggestions to enhance the efficiency and effectiveness of the SER system based on their experiences during usability testing. Using a content analysis approach, responses were categorized as 'no suggestions' or 'suggestions,' as shown in Table 1.14. The 'no suggestions' classification was based on responses indicating that the system was satisfactory, while the 'suggestions' classification included responses providing insights for further improvements. 40% of the respondents, who are educators with hearing impairments, had no suggestions. In contrast, the remaining 60% of respondents provided suggestions for future enhancements. Some recommended adding features beyond voice to detect student emotions, such as pairing student voice with textual cues or combining student voice with facial expressions. Other suggestions included expanding the range of emotions detected to respond more precisely to students' emotions.

Table 4.14. Feedback on the functionality of the SER system by educators with hearing impairment

(R=Respondent)

Respondent	SER system Detected & Displayed Real-Time Student Emotions Correctly?	Category	Instances of Incorrect Emotion Interpretation? Please Provide Comments	Category	Suggestions to Enhance the SER System?	Category

R1	Negative ones were always getting detected **correctly**. Some positives ones were not.	Mixed Feedback	**Few** positive emotions not recognized correctly.	Minor issues	**No** Suggestions	No Suggestions
R2	**Yes**	Positive Feedback	**Sometimes** positive feedback also goes in negative.	Minor issues	**Add** text detection with verbal.	Suggestions
R3	**Yes**	Positive Feedback	**No,** there were no such instances. It was displaying correctly. There are further comments.	No issues	Can **combine** with facial recognition.	Suggestions
R4	**Yes**	Positive Feedback	**No** such instances.	No issues	Echo, or able to detect with **noisy** environment.	Suggestions
R5	Sometimes **not** correctly.	Mixed Feedback	Sometimes issues	Minor issues	Speech emotions may not be able to reflect students' emotions **entirely** but a good start.	Suggestions
R6	Mostly **yes**	Positive Feedback	While having **connectivity** issues or when students were neutral.	Minor issues	**nothing** at present.	No Suggestions
R7	**yes**	Positive Feedback	**No** - the system was correct each time.	No issues	**Not** at the moment.	No Suggestions

R8	**Yes**, its correct	Positive Feedback	**No** such cases.	No issues	**No** suggestions.	No Suggestions
R9	**Yes**	Positive Feedback	When students just mumble, it cannot recognize properly	Minor issues	Be aware of the potential for disability to mask level of response **accurately**.	Suggestions
R10	**Yes**	Positive Feedback	**None**.	No issues	Would love a spectrum	Suggestions

2) Educators without hearing impairment

As evidenced in Figure 4.53, 100% of the respondents without hearing impairment agreed or strongly agreed that the user-friendly user interface highlights a clear and concise layout. 90% agreed that controls for recording and replaying student utterances were intuitive, which implies that these features make it easier for the users to navigate easily. Most importantly, an affirmative response regarding the detected student emotion being displayed as positive or negative, along with an emoji, indicates the effectiveness of the SER system.

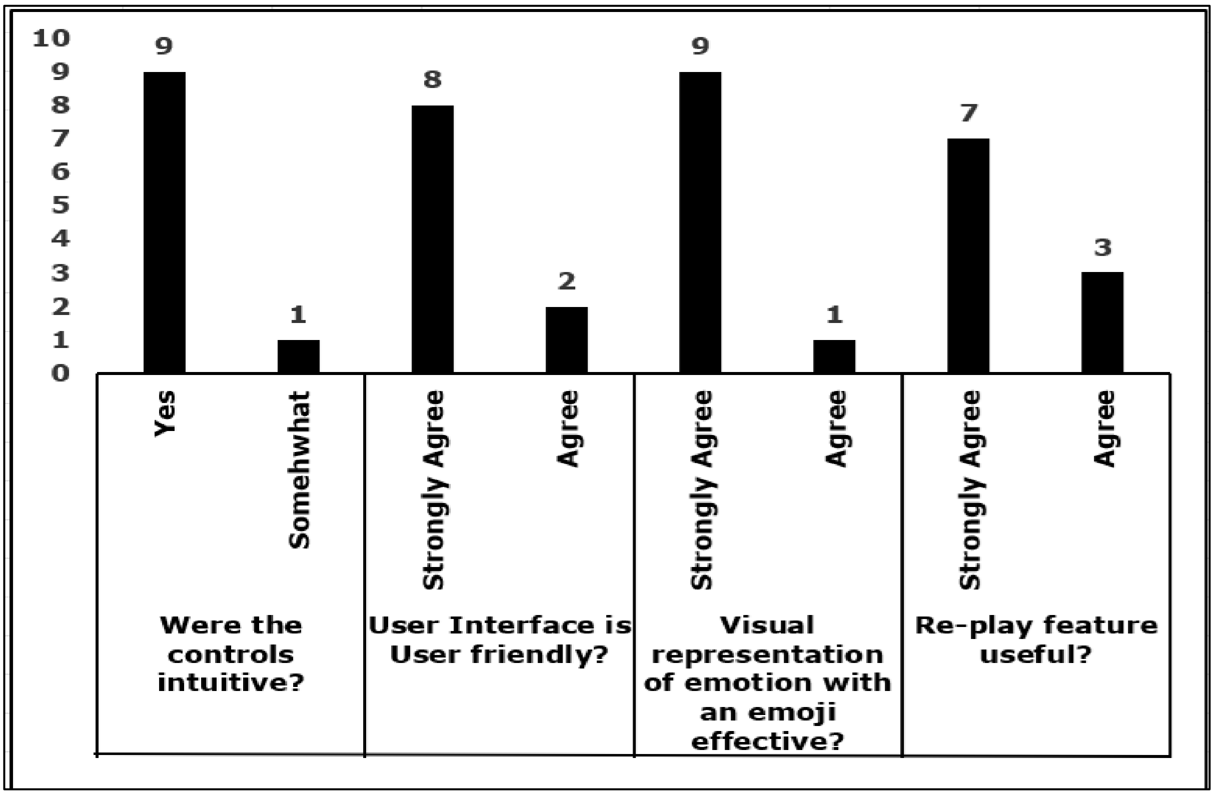


Figure 4.53: Feedback on the features of the SER system by educators without hearing impairment

The survey also featured some open-ended questions. The feedback on the functionality of the SER system from respondents without any hearing impairment issues was analysed as below:

a) The first question aimed to gather respondents’ opinions on whether the system accurately identified and displayed real-time emotions (positive/negative) from student vocal feedback. This feedback from educators was crucial as it directly reflected the system's efficacy in recognizing student emotions. Using the content analysis approach, respondents’ responses could be categorized exclusively as “positive feedback”, as seen in Table 4.15. All the respondents (100%) reported that the system accurately detected and displayed real-time emotions from students’ vocal feedback. As seen in Table 4.15, all respondents, educators without hearing impairment issues, felt that the system could correctly detect the student’s emotions.

b) The next question aimed to identify specific instances where the SER system did not correctly interpret students’ emotions. Using a content analysis approach, respondents’ answers were categorized as either ‘no issues’ or ‘minor issues,’ as shown in Table 4.15. The 'no issues' classification indicated no incorrect emotion interpretation, which 30% of educators reported without hearing impairment. The ‘minor issues’ classification was based on responses indicating some challenges with emotion interpretation, as reported by 70% of respondents. The issues were attributed to unstable internet, students failing to emote correctly despite the guidelines, or when the students’ volume was too low. No educator reported any major issues with the system detecting student emotions as positive or negative.

c) The final question asked of the respondents was suggestions to enhance the efficiency and effectiveness of the SER system based on their experiences during usability testing. Responses were categorised using a content analysis approach as ‘no suggestions’ or ‘suggestions,’ as shown in Table 4.15. The ‘no suggestions’ classification was based on responses indicating that the system was satisfactory, while the ‘suggestions’ classification included responses providing insights for further improvements. 40% of the respondents, educators without hearing impairments, had no suggestions. They expressed satisfaction with the system and made favourable comments about it. In contrast, the remaining 60% of respondents provided suggestions for future enhancements. Some recommended adding features beyond voice to detect student emotions, such as pairing student voice with textual cues or combining student voice with facial expressions. Other suggestions included expanding the range of emotions detected to respond more precisely to students' emotions. Some even suggested increasing the training samples during the models' training to detect the students' emotions more accurately. Some even advised to incorporate the SER system with any selected online teaching tool to make its usage more convenient.

Table 4.15. Feedback on the functionality of the SER system by educators without hearing impairment *(R=Respondent)*

Respondent	**SER system Detected & Displayed Real-Time Student Emotions Correctly?**	**Category**	**Instances of Incorrect Emotion Interpretation? Please Provide Comments**	**Category**	**Suggestions to Enhance SER System?**	**Category**
R1	**Yes**	Positive Feedback	Students **shaky** voice is detected as negative even if they are happy, because they were shy.	Minor issues	Camera (**face capture**) for emotion	Suggestions
R2	**Yes**.	Positive Feedback	I did not face **any issues**.	No issues	I am fully **satisfied** with the efficiency of this SER system	No Suggestions

R3	**Yes**, I believe.	Positive Feedback	Sometimes, when students are not able to express their emotions clearly.	Minor issues	Voice with **image** recognition may be enhance the efficiency of recognition	Suggestions
R4	**Yes**	Positive Feedback	No such **instances**.	No issues.	To make it omnipresent by pairing it with the online system like zoom	Suggestions
R5	Mostly **Yes**.	Positive Feedback	**Some challenges in** detecting positive emotions.	Minor issues	Need to be **sensitive** in voice recording in differentiating positive and negative emotions.	Suggestions
R6	To some extend **yes**.	Positive Feedback	Due to high pitch but the words represent for sad, it **unable** to detect correctly.	Minor issues	**No**	No Suggestions

R7	**Yes**	Positive Feedback	No. It detected **well**.	No issues	Increase **samples**	Suggestions
R8	**Yes.**	Positive Feedback	Yes. This could be due to the system's sensitivity towards **tone** and pauses.	Minor issues	Perhaps further usability testing to train the system to recognize the voices/tone for **more** accurate recognition.	Suggestions
R9	**Yes.**	Positive Feedback	The system could not capture the audio when the student's voice is **low**.	Minor issues	**No**. The existing system is good.	No Suggestion
R10	**Yes**	Positive Feedback	Sometimes **internet** bandwidths and other outside sounds.	Minor issues	**No**	No Suggestions

In conclusion, both respondents with and without hearing impairments found the SER system's web interface visually appealing and intuitive. This enhanced the overall user experience, making it easier for educators to interact with the system during testing. By prioritizing these aspects, the system enhances user engagement, satisfaction, and, ultimately, the success of online teaching endeavours. The SER system's efficiency and effectiveness in detecting students' emotions from verbal feedback during the usability testing has been fairly successful. This is based on the feedback from respondents (educators with and without hearing impairments) presented in the above section. These educators felt the system was generally effective in identifying the real-time emotions of student verbal feedback as positive or negative. This makes the system reliable to use. However, they noted instances where emotions were not correctly interpreted, often due to students' lack of expressiveness with certain emotions, like happiness. Their suggestions for enhancing the system primarily focused on incorporating multimodal options for emotion detection, such as using facial recognition or combining text analysis with voice input.

4.5.2.6 User Experience - Perceived Impact of SER system on Educator

1) Educators with hearing impairment

90% of the respondents with hearing impairment either agreed or strongly agreed that when negative emotion was detected in a student's feedback through the SER system, it helped adjust teaching strategies in an online class. It further suggests a perceived impact on the customization of instruction. This is evident from Figure 4.54. Similarly, a positive emotion detected in student feedback was seen to reinforce confidence in the educator to continue with the ongoing class. 90% of the educators unanimously agreed or strongly agreed regarding this point, as seen in Figure 4.54. This indicates a perceived positive impact on the educator's morale to continue with the class. Also, as seen in Figure 4.54, 90% of the educators, except for 10% who do not disagree but are just neutral, strongly recommend using the SER system during an online class. This implies an overall perceived benefit in enhancing teaching effectiveness using the proposed SER system.

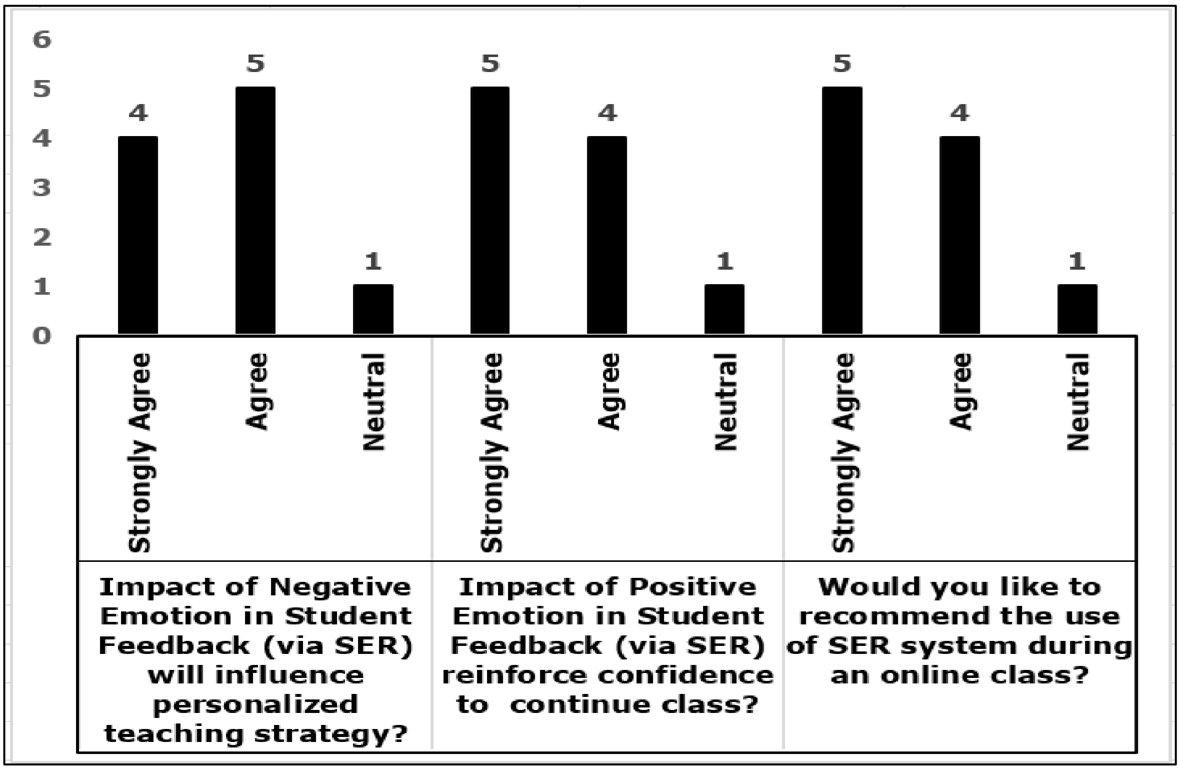


Figure 4.54: Perceived impact of SER on educators with hearing impairment

The next variable to be examined was whether student emotion detection through the SER system would influence educators' perceptions of student engagement and understanding during online classes. This question was posed to respondents as a text query (open-ended). The content analysis approach categorised responses as "positive feedback" or "negative feedback", as seen in Table 4.16. The "positive feedback" classification was based on responses indicating that the system detected emotions correctly. The "negative feedback" classification was based on responses indicating some issues with detection. Most respondents (90%) felt that student emotion recognition positively influences an educator's perception of student engagement and understanding during online classes. Only 10% felt that the student emotion recognized did not affect student engagement and teaching.

Table 4.16: The effect of the SER system on teaching by Educators with Hearing Impairment (*R=Respondent)*

RESPONDENT	Does emotion Recognition have an Impact on student Engagement?	Category
R1	yes	Positive feedback
R2	Yes	Positive feedback
R3	Yes	Positive feedback
R4	Yes	Positive feedback

R5	Unlikely.	Negative feedback
R6	Yes, very much	Positive feedback
R7	Yes	Positive feedback
R8	Yes	Positive feedback
R9	Yes	Positive feedback
R10	yes	Positive feedback

2) Educators without hearing impairment

All the respondents with normal hearing agreed or strongly agreed that detecting negative emotions in student feedback through the SER system helps adjust teaching strategies in online classes. This indicates a perceived impact on the customization of instruction, as shown in Figure 4.55 Similarly, as seen in the figure, a positive emotion detected in student feedback reinforces confidence in the educator to continue with the ongoing class. All the educators without hearing impairment have agreed or strongly agreed regarding this point. This indicates a perceived positive impact on an educator's morale to continue with the class. Finally, all the educators strongly recommend using the SER system during an online class. This implies a perceived benefit in enhancing teaching effectiveness using the SER system.

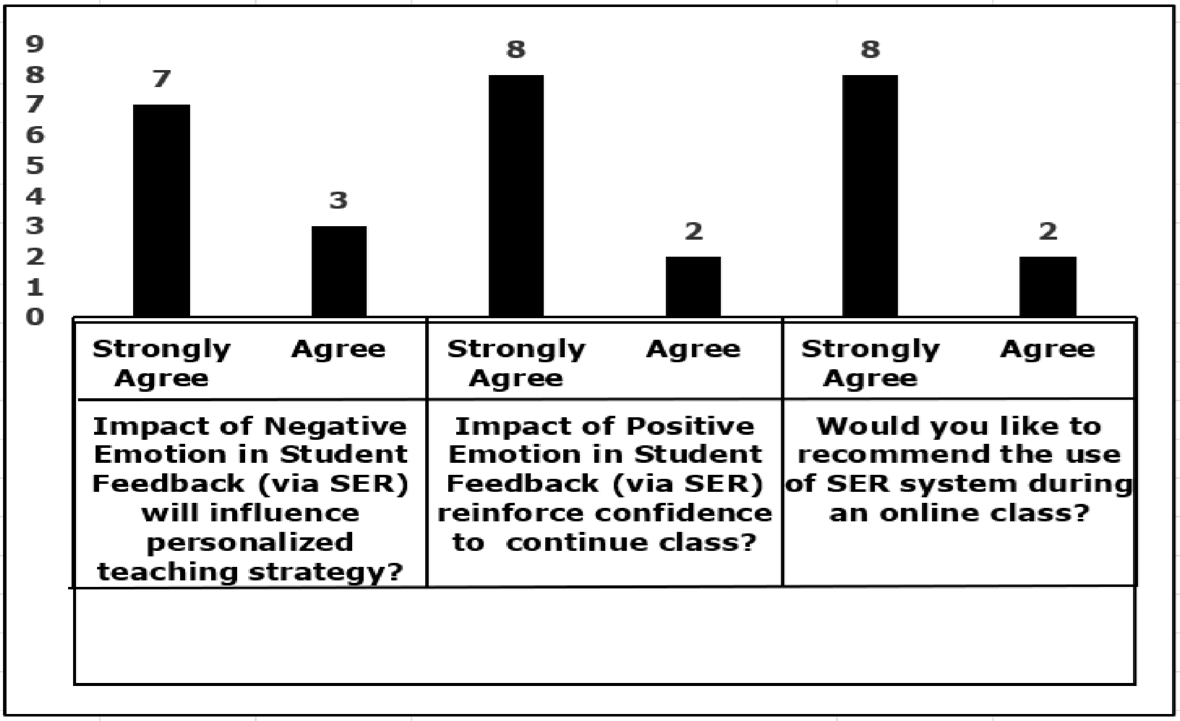


Figure 4.55: Perceived impact of SER system on educators without hearing impairment

The last variable in this category of questions was to examine whether student emotion recognition through the SER system would influence educators' perceptions of student engagement and understanding during online classes. This question was posed to respondents as an open-ended text query. The content analysis approach categorised responses as "positive feedback", as seen in Table 4.17. The "positive feedback" classification was based on responses indicating that the system detected emotions correctly. Respondents gave a 100% positive response. This shows that all the respondents felt that student emotion recognition positively influences educators' perceptions of student engagement and understanding during online classes.

Table 4.17. The effect of the SER system on teaching by Educators without Hearing Impairment

(R=Respondent)

RESPONDENT	Does emotion Recognition have an Impact on student Engagement?	Category
R1	Yes	Positive feedback
R2	Yes	Positive feedback
R3	yes	Positive feedback
R4	yes	Positive feedback
R5	Yes	Positive feedback
R6	Yes, of course.	Positive feedback
R7	Yes	Positive feedback
R8	yes	Positive feedback
R9	Yes.	Positive feedback
R10	yes	Positive feedback

4.5.3.7 Further Comments or Improvements Needed in the SER System

1) Educators with hearing impairment

The final open-ended question in the survey prompted users to give a final comment or any suggestions for improvements needed in the system. Using the content analysis approach, respondents' responses were categorized as "no comments," "suggestions", or "positive feedback", as seen in Table 4.18. The "no comments" classification was based on responses indicating no respondent feedback or comments. The "suggestions" classification was based on responses asking for a suggestion or recommendation from the respondents. The "positive feedback" classification was based on responses expressing positive feedback or appreciation from the respondents. As seen in Table 4.18 below, 90% of the respondents either gave positive feedback or provided no comment, indicating the system didn't require any further improvements. Some responses from respondents have been highlighted in bold in Table 4.18. Only 10% of the respondents gave suggestions to improve the system further by adding text as an emotion detection modality. Their comments overall reflect a high level of satisfaction with the system.

Table 4.18: Comments or improvements in the SER system *(R=Respondent)*

Respondent	Any other feedback or comments?	Category
R1	**No** comments.	No Comments
R2	Based on **words** it should give positive or negative.	Suggestion
R3	**No** comments.	No Comments
R4	**No** comments.	No Comments
R5	**None**.	No Comments
R6	It is a **good system**	Positive Feedback
R7	**Thank you for giving this opportunity participate in this research.**	Positive Feedback
R8	It is a **excellent app**	Positive Feedback
R9	**Thank you** for developing this work!	Positive Feedback
R10	**None**	No Comments

2) Educators without hearing impairment

The final open-ended question in the survey prompted users to give a final comment or any suggestions for improvements needed in the system. Using the content analysis approach, responses were categorized as "no comments," "suggestions", or "positive feedback", as seen in Table 4.19. The "no comments" classification was based on responses indicating no respondent feedback or comments. The "suggestions" classification was based on responses asking for a suggestion or recommendation from the respondents. The "positive feedback" classification was based on responses expressing positive feedback or appreciation from the respondents. 80% of the respondents either gave positive feedback or no comment response, indicating the system didn't require any further improvements. Only 10% of the respondents suggested improving the system further by including facial expressions in the emotion detection system and making it multi-modal. Their comments overall reflect a high level of satisfaction with the system.

Table 4.19. Comments or improvements in the SER system *(R=Respondent)*

Respondent	Any other feedback or comments?	Category
R1	It is a **good** technology to help lecturer understand their students	Positive feedback
R2	**No comments.**	No Comments
R3	Combination of image and vocal recognition may **enhance** the system more	Suggestion
R4	**Nice idea**	Positive feedback
R5	The system might need higher sensitivity to be more **effective**.	Suggestion
R6	**No feedback**	No Comments
R7	**System is good**	Positive feedback
R8	**No comments.**	No Comments
R9	**No comments.**	No Comments
R10	**No comments.**	No Comments

In conclusion, the feedback from educators, both with and without hearing impairments, highlights a significant positive impact of the SER system on teaching strategies and effectiveness in online classes. Most educators agreed that detecting negative emotions in student feedback through the SER system aids in adjusting teaching strategies, indicating a perceived benefit in customizing instruction. Similarly, detecting positive emotions reinforces the educator's confidence, as most suggested an improvement in morale and continuity of the class due to the positive emotions detected. Furthermore, educators agree that student emotion detection influences their perception of student engagement and understanding during online classes. This data collectively underscores the perceived advantages of the SER system in fostering a more adaptive and confident teaching environment for all educators (Refer to Appendix 9 for sample respondent data).

4.5.2.8 Summary of the Findings

The results and analysis of the respondent data detailed above offer a comprehensive understanding of the respondents' perceptions of the SER system. The feedback from educators, both with and without hearing impairments, has been analysed and presented separately. The findings reveal that both groups have responded very positively regarding the SER system's effectiveness, efficiency, and its impact on adjusting teaching strategies based on the student emotions detected. In this section, further analyses have been done on the collected data using the mean and standard deviation of the quantitative data extracted to gain deeper insights. A descriptive analysis for this purpose [201] has been used. The mean is meant to give an understanding of the central tendency of the data, while the standard deviation gives an idea of the spread of the data. In other words, the mean or average value indicates the typical value around which the data points cluster. The standard deviation measures the dispersion or spread of the data points around the mean. It quantifies the amount of variation or uncertainty in the dataset. A slight standard deviation indicates that the data points tend to be close to the mean, while a large standard deviation indicates that the data points are spread out over a broader range [201, 212, 213].

1) **Effectiveness and Efficiency of the Speech Emotion Recognition System – Usability analysis**

Usability refers to how well an experience enables users to meet their goals. Key aspects include efficiency, effectiveness, and user satisfaction [171]. The twenty respondents (educators with and without hearing impairment) rated on a 5-point Likert scale the features provided by the SER system. The quantitative data was analysed using frequency count, as discussed in the methodology Section 3.2. Features like the interface being user-friendly, controls being intuitive, visual representation of the emotion as negative or positive being effective, and features like a replay of the student feedback are analysed further in this section. The mean and standard deviation of the quantitative data are computed for deeper insights, as shown in Table 4.20, which is discussed in detail below:

a) The first category, which is the SER system having a user-friendly interface, shows a mean score of 4.6, 4.4 and 4.8 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. This indicates that, on average, respondents found the web interface of the SER system to be quite user-friendly, with the scores generally leaning towards the higher end of the scale. This suggests a positive reception overall. The standard deviation of approximately 0.59, 0.69 and 0.42 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively, indicates that there is a relatively low level of variability in the responses. Most users rated the system similarly, which implies consistent user experiences. The low standard deviation supports the conclusion that the perceived user-friendliness of the system is stable and widely accepted.

b) The second category, which is the SER system having intuitive recording controls, shows a mean score of 4.9 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10). This indicates that, on average, respondents found the controls of the SER system to be highly intuitive. The standard deviation of approximately 0.30, 0.32 and 0.32 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively, indicates that there is very low variability in the responses. The low standard deviation supports the conclusion that the perceived intuitiveness of the controls is consistent.

c) The third category, which is the effectiveness of the visual emotion feedback, shows a mean score of 4.8 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. This indicates that, on average, respondents found the visual representation of emotion very effective. The standard deviation of approximately 0.52, 0.42 and 0.63 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively, indicates a relatively low level of variability in the responses. The low standard deviation supports the conclusion that the visual representation of emotions is perceived uniformly well among users, with few deviations from the mean rating. This consistency suggests that nearly all users found the visual representation to meet their expectations effectively.

d) The last category of SER features, which is the usefulness of the replay feature, shows a mean score of 4.5, 4.4 and 4.4 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. This indicates that, on average, respondents found the replay feature of the student recording to be quite satisfactory. This suggests a generally positive reception, with users rating the feature on the higher end of the scale. The standard deviation of approximately 0.82, 0.70 and 0.96 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively, indicates a little level of variability in the responses. While most users rated the replay feature similarly, some variation indicates that not all users had the same experience. The moderately low standard deviation suggests that most users found the replay feature effective. Still, a few had different experiences, as reflected in the lower ratings. Most respondents perceive the replay feature positively, with a slight spread in opinions.

Table 4.20: Usability evaluation results of the quantitative data

Category	All respondents (n=20)		Respondents with hearing-impaired (n=10)		Respondents without hearing-impaired (n=10)	
SER Features	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
User-friendly Interface	4.6	0.59	4.4	0.69	4.8	0.42
Intuitive Recording Controls	4.9	0.30	4.9	0.32	4.9	0.32
Effectiveness of Visual Emotion Feedback	4.8	0.52	4.8	0.42	4.8	0.63
Usefulness of the Replay Feature	4.4	0.82	4.4	0.70	4.4	0.96

Sentiment analysis, also known as opinion mining, focuses on understanding opinions. It is a branch of NLP, part of computer science and artificial intelligence, dealing specifically with the interaction between human language and computers [181]. It involves collecting and examining users' views or opinions about a product, subject, or system [182, 183]. Sentiment analysis encompasses both subjectivity and polarity. It is a technique used to examine the polarity of the text to assess whether data expresses positive, negative, or neutral sentiments [184]. For this research, a lexicon-based analysis [185] was conducted. Lexicons are collections of tokens, where each token is assigned, a predefined score indicating its neutral, positive, or negative nature. These scores typically range from +1 for positive, 0 for neutral, and -1 for negative. This approach provides a numerical representation of sentiment useful for aggregating and analysing feedback [182, 184, 185]. Using this approach, the qualitative data from users was converted into numerical data. Table 4.21 displays the polarity of their opinion about real-time emotion detection by the SER system.

Table 4.21. Sentiment analysis of qualitative data

Educators with Hearing Impairment		Educators without Hearing Impairment	
Was Real-time Emotion Detection by SER system Accurate?	**Polarity**	**Was Real-time Emotion Detection by SER system Accurate?**	**Polarity**
Negative ones correctly. But positives ones sometimes were not getting detected correctly.	0	Yes	1
Yes	1	Yes.	1
Yes	1	Yes, I believe.	1
Yes	1	Yes	1
Sometimes not correctly.	0	Mostly Yes.	1
Mostly yes	1	To some extend yes.	0
Yes	1	Yes	1
Yes, its correct	1	Yes	1
Yes, within parameters	1	Yes	1
Yes	1	yes	1

The mean and standard deviation of the numerical qualitative data is computed for deeper insights, as shown in Table 4.22. A mean score of 0.85, 0.8 and 0.9 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) were obtained, respectively. The mean scores are close to 1, indicating that most users provided positive feedback regarding the system's effectiveness in identifying emotions. Most users rated the system as effective. A standard deviation of 0.36, 0.4 and 0.3 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) was obtained, respectively. A standard deviation 0.3 indicates that the feedback ratings are closely clustered around the mean. This suggests there is relatively low variability in the user feedback, meaning that most users have similar opinions about the system's effectiveness. The low standard deviation suggests this satisfaction is consistent across users, with few outliers or differing opinions.

Table 4.22: Usability evaluation results of the qualitative data

Category	**All respondents (N=20)**		**Respondents with hearing-impaired (N=10)**		**Respondents without hearing-impaired (N=10)**	
SER function	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Was Real-time Emotion Detection Accurate?	0.85	0.36	0.8	0.4	0.9	0.3

2) Impact of SER – User Experience analysis

a) The first SER impact category, which is adjusting teaching based on a negative emotion, shows a mean score of 4.5, 4.3 and 4.7 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. This mean score indicates that, on average, respondents found detecting negative student emotions in adjusting teaching strategies quite significant. This suggests that most educators see detecting negative emotions as an important factor in modifying their teaching approach. A standard deviation of approximately 0.60, 0.67 and 0.48 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) was obtained, respectively. This indicates a moderate level of variability in the responses. While most educators rated the impact similarly, some variation indicates different levels of perceived importance among the respondents. The moderate standard deviation suggests that, although most educators strongly agree on the significance of emotion detection, a few agree or do not feel very strongly about it. Overall, the system's ability to detect negative emotions is valuable for informing teaching strategies, with a generally consistent but slightly varied perception among respondents**.**

b) The second category of the impact of SER, which is the impact of positive feedback on class continuity, shows a mean score of 4.6, 4.4 and 4.8 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. These mean scores indicate that, on average, respondents found that detecting positive student emotions significantly impacted their confidence in continuing with the class. This suggests that most educators believe the system's ability to detect positive emotions is valuable for reinforcing their teaching strategy and boosting their confidence. A standard deviation of approximately 0.59, 0.70 and 0.42 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) was obtained, respectively. These values indicate a moderate level of variability in the responses. The moderate standard deviation implies that although there is general agreement on the positive influence of emotion detection, some educators may experience this impact more strongly than others. Overall, the system's ability to detect positive emotions is beneficial, with a generally high but somewhat varied level of perceived impact among respondents.

c) The third category of the impact of SER, which is whether the respondents would recommend SER, shows a mean score of 4.6, 4.4 and 4.8 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10), respectively. The mean scores indicate that, on average, respondents are highly likely to recommend the Speech Emotion Recognition (SER) system for online classes. This suggests a strong endorsement of the system's utility and effectiveness for enhancing online teaching experiences. A standard deviation of approximately 0.59, 0.70 and 0.42 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) was obtained, respectively. These values indicate a moderate level of variability in the responses. The moderate standard deviation implies that most respondents are aligned in their positive assessment of the system's suitability for online classes. Overall, the high mean and moderate standard deviation indicate that the SER system is generally perceived as valuable for online teaching, with some variation in the strength of recommendation among respondents.

Table 4.23. Impact of SER - Evaluation of the quantitative data

Category	All respondents (N=20)		Respondents with hearing-impaired (N=10)		Respondents without hearing-impaired (N=10)	
Impact of SER	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Adjusting Teaching Based on Negative Emotion?	4.5	0.60	4.3	0.67	4.7	0.48

Impact of Positive Feedback on Class Continuity?	4.6	0.59	4.4	0.70	4.8	0.42
Recommend SER?	4.6	0.59	4.4	0.70	4.8	0.42

Here, sentiment analysis is also used to understand the user's opinion on the impact of the SER system on student learning and engagement. The qualitative data collected from users was converted into numerical data. Table 4.24 displays the polarity of user opinions regarding the impact of SER on student engagement and learning.

Table 4.24. Impact of SER - Sentiment analysis of qualitative data

Educators with Hearing Impairment		**Educators without Hearing Impairment**	
Does emotion Recognition have an Impact on student Engagement?	**Polarity**	**Does emotion Recognition have an Impact on student Engagement?**	**Polarity**
yes	1	Yes	1
Yes	1	Yes	1
Yes	1	yes	1
Yes	1	yes	1
Unlikely	0	Yes	1
yes, very much	1	Yes, of course.	1
Yes	1	Yes	1
Yes	1	yes	1
Yes, could change approach to teaching positively	1	yes	1
yes	1	yes	1

The mean and standard deviation of the numerical qualitative data were computed for deeper insights, as shown in Table 4.25. A mean score of 0.9, 0.9 and 1 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) were obtained, respectively. A mean score of 1, or close to 1, indicates that most users provided positive feedback regarding the impact of the SER system on student learning and engagement. Most users rated the system as very impactful in their teaching activities. A standard deviation of 0.22, 0.31 and 0 for all respondents (n=20), respondents with hearing impairment (n=10), and respondents without hearing impairment (n=10) were obtained, respectively. A standard deviation of 0.3 or 0 indicates that the feedback ratings are closely clustered around the mean. This suggests there is relatively low variability in the user feedback, meaning that most users have similar opinions about the system's impact being very positive. The low standard deviation suggests this satisfaction is consistent across users, with few outliers or differing opinions.

Table 4.25. Impact of SER - Evaluation of the qualitative data

Category	All respondents (N=20)		Respondents with hearing-impaired (N=10)		Respondents without hearing-impaired (N=10)	
Impact	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**	**Mean**	**Standard Deviation**
Emotion Recognition Impact on student Engagement?	0.95	0.22	0.9	0.31	1	0

The analysis indicates that the SER system effectively detected students' emotions from verbal feedback during usability testing. This was observed in both separate and combined data from educators with and without hearing impairments. Participants from both groups reported high satisfaction, citing a user-friendly interface, intuitive controls, clear visuals, and a replay function that enhanced instructional flexibility and confidence. Quantitative and qualitative feedback consistently reflected the system's value in improving the effectiveness and efficiency of online teaching. To gain further insight, two independent t-tests were conducted using selected mean values. The first compared feature-related feedback, and the second examined perceived impact. The resulting p-values (0.291 and 0.5) were above the 0.05 significance level, indicating no statistically significant differences between groups. These findings support the system's inclusive design and suggest that educators, regardless of hearing ability, shared similarly positive perceptions of its usability and usefulness.

The high average scores observed in the responses reinforce participants' positive perception of the system, though they may also reflect some degree of social desirability bias. Since the study involved experienced educators selected through purposive sampling, it is likely they found the system relevant and easy to integrate into their familiar online teaching environments. To encourage honest feedback and minimise potential bias, the survey was conducted anonymously, and participants were assured of full confidentiality. Responses would not affect their relationship with the researcher or any associated institution. The survey design included both rating scales and open-ended questions, allowing participants to elaborate on their experiences and share suggestions—many of which were offered. While some level of response bias cannot be entirely ruled out, appropriate steps were taken to support honest, reflective feedback.

For real-time evaluation, student responses during the usability testing sessions were observed live as participants interacted through the online platform. Students were provided with guidelines to help express the emotions as positive (happy or neutral) or negative emotions (angry, sad or fear) while giving the verbal feedback. The system's output was manually recorded by the researcher in an Excel sheet. Ground truth labelling was based on immediate observations during the session and aligned with predefined emotional guidelines given to participants. When students expressed emotions in line with the given instructions, the system demonstrated approximately 90% alignment with the manual observations. This observation-based approach offered a meaningful and practical assessment of the system's real-time performance in an educational setting, reinforcing its applicability and effectiveness in live teaching environments.

4.6 Educator Use of SER Feedback in Online Teaching

As highlighted in Section 1.2.3, emotional engagement, as conveyed through students' vocal cues, is a critical driver of adaptive teaching, particularly in online settings where visual and non-verbal cues are limited. Section 1.2.4 further outlines the issue of an "emotional gap" in virtual classrooms. Educators find it challenging in sensing students' engagement or their level of understanding during online sessions. This lack of emotional insight made it harder to self-regulate their teaching approach. A challenge that is especially pronounced for late-deafened educators who struggle to interpret vocal emotional cues due to their hearing loss. Late-deafened educators generally rely on body language, lip reading, and facial expressions in face-to-face classes to understand student reactions—cues that are often absent or limited in virtual environments.

The SER system is meant for the emotional awareness based on student feedback and adjust teaching approaches. The system displayed real-time emotional output which is either positive or negative. This is based on students' verbal responses. When negative feedback was detected, educators can slow down on their teaching speed, re-explain the content, or provide additional examples to support understanding. When positive feedback is displayed, it is meant to act as a reinforcement that the content was well received, encouraging the educator to proceed or to seek further input from another student. This was especially helpful for late-deafened educators, who found the emotional cues valuable in maintaining engagement and adjusting their teaching dynamically, despite being unable to hear the students' tone or verbal nuances directly.

Though there are no direct measures of academic performance, the SER system is meant to improve the educator’s ability to respond appropriately to student emotional cues, resulting in more interactive and engaging sessions that closely resembled the responsiveness of in-person teaching environments. Thus, the SER system not only enhances emotional awareness but also empowers the educators to make timely, emotion-informed instructional decisions, helping bridge the emotional disconnect often found in online teaching environments.

This signifies the completion of stages 4 and 5 of the HCI research process for SER, as depicted in Figure 3.2. The corresponding sections in Figure 4.56 have been shaded to illustrate this.

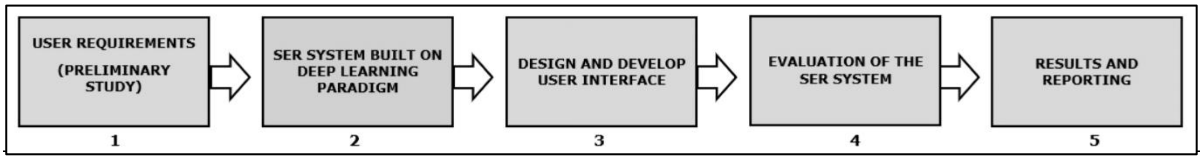

USER REQUIREMENTS (PRELIMINARY STUDY)
SER SYSTEM BUILT ON DEEP LEARNING PARADIGM
DESIGN AND DEVELOP USER INTERFACE
EVALUATION OF THE SER SYSTEM
RESULTS AND REPORTING
1
2
3
4
5

Figure 4.56 Updated process of the HCI research for the SER system

The aims of evaluating our developed SER system, highlighted in section 4.5.1, have been achieved. They are given below:

1) To assess the effectiveness and efficiency of the developed SER system.
2) To obtain feedback on the perceived impact of the SER system on educators with and without hearing impairment.

The findings and analyses of the survey results clearly show that the last objective which is RO4 has also been addressed.

Chapter 5

5. Discussion and Conclusion

5.1 Achievement of Research Objectives

This study achieved each research objective as described below –

RO1 - To identify the key speech features and feature engineering approaches that enhance the efficiency and accuracy of emotion recognition in a deep learning-based real-world SER system.

To achieve the first research objective (RO1), as stated above, an extensive literature review was conducted first. A systematic approach was adopted to conduct the review. A search query combining relevant such as "speech emotion recognition," "real-time emotion detection," "hybrid CNN models," and "accessibility in SER systems" were used to explore key databases such as Google Scholar, IEEE Xplore, and Scopus. Inclusion criteria were applied to focus on peer-reviewed articles, conference papers, and other credible sources for machine learning-based and deep learning-based approaches. This review aimed to establish a foundational understanding of SER fundamentals and identify pathways for advancing the application of SER in real-world scenarios. It facilitated the selection of effective features for SER - ZCR, RMS, Chroma-STFT, Mel Spectrogram, and MFCCs. Each of these features with their complementary strengths, addressed different aspects of the speech signal that were crucial for capturing emotional nuances. The selected speech features effectively distinguish emotions based on energy and articulation patterns. ZCR highlights voiced emotions like anger and happiness with frequent sign changes, while neutral and sadness show reduced ZCR due to subdued articulation. RMS indicates loudness, with higher values for energetic emotions (anger, happiness), moderate levels for neutral, and lower values for sadness. Chroma-STFT captures tonal shifts, showing dynamic variations for anger and happiness, flat monotones for sadness, stable intonation for neutral, and variable patterns for fear. Mel Spectrograms map frequencies perceptually, revealing high-frequency energy for anger and happiness, smooth patterns for sadness, balanced contours for neutral, and oscillations for fear. MFCCs capture vocal traits, reflecting forceful articulation in anger, rapid variation in happiness, slower delivery in sadness, balanced tones in neutral, and rapid or hesitant shifts in fear. These features group emotions as high-energy (anger, happiness), low-energy (sadness, fear), or moderate (neutral), enabling efficient visualization and analysis. Utilizing multiple audio features, rather than relying on a single one, integrated diverse sound characteristics into a single training sample. This approach resulted in a more comprehensive representation of the audio, which enhanced the performance of SER models by improving emotional detection accuracy and generalization. Furthermore, the network configurations of the three hybrid architectures—Model A (averaging fusion), Model B (parallel merging), and Model C (sequential integration)—were specifically designed to leverage the fusion of these acoustic features. Each deep learning model combined three types of CNNs (1D, 2D, and 3D) using distinct fusion techniques to maximize the strengths of the selected audio features.

RO2 - To compare and assess the effectiveness of a hybrid CNN architecture integrating 1D, 2D, and 3D convolutional layers through three selected fusion techniques, identifying the approach with the highest accuracy and efficiency and adapting it to build a real-world SER application.

To achieve the second research objective (RO2), as stated above, three hybrid CNN architectures—Model A (averaging fusion), Model B (parallel merging), and Model C (sequential integration)—were developed and compared. These architectures combined 1D, 2D, and 3D convolutional layers to fuse acoustic features through distinct techniques. The first two architectures, Models A and B, are multi-stream CNNs. Model A fuses streams by averaging decisions, while Model B concatenates the outputs from the multi-streams for further processing. The third architecture, Model C, combines 3D, 2D, and 1D CNNs sequentially, the selected three hybrid architectures were tested across five datasets, which were semi-natural and acted datasets. The five selected datasets were the Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) and Database of Elicited Mood in Speech (DEMoS), Toronto emotional speech set (TESS), Ryerson Audio-Visual *Database* of Emotional Speech and Song (RAVDESS) and Berlin Emotional *Database* (EMODB). The hybrid merge fusion (Model B) approach gave an accuracy of 70% (IEMOCAP), 69% (DEMoS), 73% (EMODB), and 45% (RAVDESS), and 100% (TESS) and hybrid sequential fusion (Model C) an accuracy of 63% (IEMOCAP), 58% (DEMoS), 80%(EMODB), and 53%(RAVDESS), and 98% (TESS). The hybrid average fusion approach consistently outperformed the other methods across all datasets, achieving the highest accuracy, with scores of 82% (IEMOCAP), 91% (DEMoS, EMODB, and RAVDESS), and 100% (TESS). Experimental results showed that Model A, employing averaging fusion, achieved the highest accuracy, validating its suitability for real-world SER applications.

RO3 - To design and develop a GUI that integrates SER-derived emotional feedback and displays emotions to late-deafened educators accurately and in real time.

To achieve the third research objective (RO3), as stated above, a graphical user interface (GUI) was designed and developed to display the SER-derived emotional feedback. The GUI for the SER system was designed adhering to Nielsen's 10 usability heuristics to ensure an intuitive and accessible experience for late-deafened educators. The design of the GUI adhered to each of Nielsen's 10 usability heuristics. The principle of visibility was ensured by providing a clear title page, intuitive options (e.g., record, stop), and constant updates on the system status, such as detecting emotions or displaying results. Familiar icons, logical information flow, and real-world representations enhanced the connection between the system and the user. The single web page interface allowed educators to refresh the home page after each emotion displayed, ensuring user control and freedom. Consistency was maintained through standardized labels and easy navigation. Flexible visual cues and accessible design elements improved efficiency, while error prevention was addressed through clear instructions and real-time error messages, such as "Error processing the audio," prompting users to re-record. The interface followed minimalist design principles, using high-contrast fonts and well-organized layouts to ensure readability and clarity. Help and documentation were provided upfront to guide educators, enabling them to make personalized decisions based on detected emotions. By focusing on Nielsen's heuristic principles, the GUI for the SER system was designed to be efficient, intuitive, and accessible. The GUI displayed the SER-derived emotional feedback, enabling late-deafened educators to interpret real-time student emotions effectively, thereby enhancing the system's usability and adaptability.

RO4 - To evaluate the GUI-integrated SER system for its effectiveness, efficiency and perceived impact on late-deafened educators for an effective engagement in online teaching.

The SER system was tested in real-time and assessed for usability and user experience, targeting educators with hearing impairments (late-deafened) and those without hearing impairments. They evaluated the SER system for its usability, user experience, and effectiveness in addressing challenges related to discerning student emotions during online classes. Participants were selected based on specific criteria such as teaching experience, teaching online, and hearing condition (purposive sampling), and the testing involved system usage, followed by data collection through surveys and questionnaires. Quantitative data were analysed using descriptive statistics, while qualitative data underwent content and sentiment analysis. The late-deafened educators highlighted the web interface as visually appealing and intuitive, which significantly enhanced their overall user experience, making the system easy to interact with during testing. These features were identified as instrumental in being effective. The system was perceived as generally efficient in detecting real-time emotions from students' verbal feedback, categorizing them as positive or negative. It was deemed reliable for use in online teaching. However, some instances of misinterpretation were noted, often due to students' lack of emotional expressiveness, particularly with emotions like happiness. Suggestions from educators included incorporating multimodal emotion detection methods, such as facial recognition and text analysis combined with voice input, to improve accuracy.

Overall, educators agreed that the system had a positive impact on teaching strategies by enabling them to adjust instruction based on detected negative emotions, while the identification of positive emotions improved their confidence and morale, fostering continuity and engagement in online classes. The system's ability to detect emotions also influenced educators' perception of student understanding and engagement, enhancing their teaching effectiveness.

Quantitative and qualitative feedback reflected consistently high satisfaction, with key features such as a user-friendly interface, intuitive controls, effective visual representation, and A satisfactory replay function contributes to its acceptance. While individual experiences varied slightly, the consensus among late-deafened educators underscored the SER system's efficiency and effectiveness in improving the adaptability and confidence of educators in online teaching, ultimately enhancing a more inclusive and impactful learning environment.

The following brief overview of the chapters highlights how each contributed to fulfilling the objectives of the research.

In Chapter 1, the background of this research was outlined. It began by addressing educators' challenges in understanding student emotions during online classes, with a particular focus on late-deafened educators. A real-world SER system based on a deep learning approach was proposed as a potential solution. An HCI-focused perspective was also introduced, highlighting the importance of fostering a strong connection between humans and technology. This connection referred explicitly to the interaction between the SER system and late-deafened educators, who play a pivotal role in shaping the system's design by evaluating its effectiveness, efficiency, and user satisfaction. This foundation led to a literature review covered in Chapter 2.

The first section of the review in Chapter 2 examined the fundamentals of speech emotion recognition (SER), systematically analysing the most commonly used and effective speech features for emotion detection. It also explored feature engineering techniques employed across various studies. The performance of proposed technologies was assessed by reviewing accuracy metrics reported on diverse datasets over time, offering insights into the evolution and limitations of existing approaches. This comprehensive analysis identified critical gaps, particularly the lack of SER systems that accurately detect emotions in dynamic, real-time environments. The second section expanded the review to HCI strategies, focusing on the challenges faced by late-deafened educators. While universities predominantly concentrate on supporting disabled students through targeted policies, limited attention has been given to faculty with disabilities—especially those who are hard of hearing—in terms of policies, technology, and application support. The challenges faced by late-deafened educators have also been underrepresented in HCI literature. For these educators, conducting virtual classes can be particularly challenging, as student responses during feedback may go unheard, negatively impacting the teaching and learning process. These identified gaps formed the basis of this research, which aimed at developing an SER system capable of accurately detecting emotions in real-time, specifically to support late-deafened educators.

Chapter 3 outlined the comprehensive methodology adopted to address the limitations of existing SER systems and achieve the research objectives. It began with a preliminary study employing a mixed-method approach to assess the feasibility and potential benefits of SER systems for educators, combining qualitative insights with quantitative analysis. This was followed by the development and innovative comparison of three hybrid CNN architectures—Model A (averaging fusion), Model B (parallel merging), and Model C (sequential integration)—each designed to leverage acoustic feature fusion through distinct techniques. The detailed network configurations of these models highlight their combination of 1D, 2D, and 3D CNNs to improve emotion detection accuracy via speech. The chapter also covered the design and development of a user interface tailored for real-time emotion detection, focusing on accessibility for late-deafened educators to ensure inclusivity. Lastly, the methodology for evaluating the usability and user experience of the developed SER system was thoroughly detailed. Usability testing was conducted in two distinct phases: pilot and main phases.

Following the usability testing, a survey was conducted to gather additional insights on participants' overall satisfaction, perceived benefits, and any challenges faced using the system. The methodology and procedures for both phases were meticulously described for evaluating the SER system's usability and perceived impact.

Chapter 4 presented the results, which were divided into four sections: 1) results from the preliminary study, 2) results from the comparison of three selected hybrid models, 3) results of integrating the selected SER model with the graphical user interface, and 4) results after evaluating the developed SER system for its effectiveness, efficiency, and perceived impact. The findings from the preliminary study highlighted the critical role of understanding student emotions during classes to enhance teaching quality, applicable to both traditional and online formats. Educators, regardless of hearing ability, reported shared uncertainty in accurately gauging student emotions in online sessions, which potentially negatively impacted teaching and learning outcomes. This underscored the necessity of an automatic SER system that incorporated a clear visual representation of speech-based emotions to reduce reliance on facial emotion recognition. The developed SER system, which utilized a hybrid CNN architecture with averaging fusion for high accuracy, provided functionality for the real-time detection and display of student emotions through a user-friendly interface. By identifying emotions such as happiness, sadness, fear, and neutrality, the system enabled late-deafened educators to gauge student engagement and comprehension, addressing their unique needs and enhancing inclusive teaching environments. This real-time application of the system was evaluated with educators with and without hearing impairments. The results revealed a substantial positive impact on perceived teaching effectiveness and efficiency, with high usability, satisfaction, and confidence levels reported. Features such as a user-friendly interface, intuitive controls, effective visual representation, and a replay option enabled better customization of instruction. Despite minor variations in individual experiences, the SER system was widely recognized as a valuable tool for improving online teaching, with strong recommendations for its use during classes.

5.2 Contributions

This study made several key contributions which have been categorized under the subheadings: Theoretical/Research Contributions; Technical/System Development Contributions; Practical Usability and Evaluation; and Societal/Inclusive.

Theoretical / Research Contributions

This study made several key contributions. It assesses the necessity and potential impact of a SER system in online teaching, focusing on late-deafened educators—a perspective that has been underexplored. It presents a novel exploration of how emotion-aware systems can enhance instructional communication for educators with acquired hearing loss, a group often overlooked in HCI and accessibility research. The study also advances knowledge on hybrid CNN architectures for SER by systematically comparing three fusion approaches—averaging, parallel merging, and sequential integration—across different CNN stream types (1D, 2D, 3D)—to identify the most accurate model, which was then applied in a real-world setting. This comparative evaluation contributes to the theoretical understanding of multi-stream fusion strategies in emotion recognition from speech.

Technical / System Development Contributions

The research developed a real-time SER system that integrates the best-performing hybrid CNN architecture, combining 1D, 2D, and 3D convolutional layers. This fusion model achieved high accuracy in detecting discrete emotions from speech signals. The system includes a user-friendly graphical interface that displays emotion feedback in real-time, designed with accessibility features tailored to late-deafened educators. The integration of backend CNN models with a front-end HCI interface demonstrates a technically novel and functional SER application for real-world use. Evaluation of the system's usability and user experience with educators, both with and without hearing impairments, demonstrated its positive impact on teaching effectiveness, specifically benefiting late-deafened educators in online teaching environments.

Practical Usability and Evaluation

This research successfully developed and evaluated an SER system tailored for late-deafened educators, addressing key challenges in online teaching. A usability evaluation involving educators with and without hearing impairments was conducted to assess the system's real-world applicability. The study reported high user satisfaction, with participants acknowledging improvements in their ability to monitor student emotional responses and adapt teaching strategies accordingly. The system was shown to enhance engagement and comprehension in online classes, especially for educators who may otherwise face barriers in perceiving verbal emotional cues.

Societal / Inclusive Impact

In light of society's increasing emphasis on inclusivity, this research highlighted the importance of supporting late-deafened educators in their teaching and training activities. By addressing these gaps, the study contributed by developing a solution to improve the teaching experiences of late-deafened educators and encouraging inclusivity in online teaching. Ultimately, the study encourages broader adoption of emotion-aware assistive tools to support diverse educator needs in increasingly digital educational ecosystems.

5.3 Limitations

Despite its success, the research has certain limitations. The SER system primarily focused on basic human emotions (Happy, sad, fear and neutral). As discussed in Chapter 4, Section 4.4.2, to support late-deafened educators, the emotions were grouped as positive (happy, neutral) or negative (sad, fear) to align with common student emotional states. This categorization simplifies feedback, making it actionable while reducing computational complexity. By focusing on fewer emotions, the SER system ensures accurate and meaningful real-time feedback without overwhelming users. However, by focusing only on a subset of universal emotions, the study limits its ability to capture the full range of student affect during online learning.

Another limitation of this study is the reliance on publicly available emotional speech datasets such as TESS, EMO-DB and RAVDESS, which contain acted emotional expressions. While these datasets are widely used in SER research due to their clear emotional labels and balanced class distributions, they do not fully reflect the variability and subtlety of natural student speech. To improve realism, the study also incorporated two semi-natural datasets such as IEMOCAP and DEMoS, which include more spontaneous expressions recorded in controlled but less scripted settings. However, even with these additions, the model could still face challenges in fully capturing the spontaneous, low-intensity, and informal speech patterns typical of students in real online learning environments, particularly when influenced by background noise or unclear articulation.

Further, some of the datasets used in this study like TESS and EMO-DB primarily consist of recordings from adult actors with Western accents, which limits the demographic and linguistic diversity of the speech samples. This may introduce bias in real-world applications where students come from varied cultural, linguistic, and age backgrounds, potentially affecting the model's generalizability across global or multilingual educational settings.

Also, there were instances of emotion misclassification by the SER system. Educators both with and without hearing felt the system was generally effective in identifying the real-time emotions of student verbal feedback as positive or negative. This makes the system reliable to use. However, they noted instances where emotions were not correctly interpreted, often due to students' lack of expressiveness with certain emotions, like happiness. For instance, educators noted that students with shy or mumbled low voices were sometimes misclassified as expressing negative emotions, even when they were happy. Similarly, positive feedback was occasionally misinterpreted as negative, especially when students spoke with high pitch but conveyed angry content, or when pauses and tone variations affected the system's detection accuracy. While several participants reported no misclassification issues, the examples shared by some have been incorporated to reflect the system's current limitations. While these misclassifications were not seen as severely disruptive, they did highlight challenges in recognizing subtle, low-intensity, or ambiguous emotional expressions. Interestingly, even though a negative emotion was incorrectly displayed, it prompts the educator to re-engage or clarify with students, ultimately enhancing attentiveness. Since the SER system is intended as a supportive aid rather than a decision-making tool, such misclassifications were not detrimental but served as useful cues for maintaining classroom engagement. These real-world user experiences provide clearer context for the system's limitations and future improvement areas, such as multimodal integration. Non-verbal cues such as facial expressions and textual feedback were not incorporated, potentially limiting the system's adaptability. This absence of comprehensive multimodal integration restricts the system's ability to provide a holistic understanding of emotional feedback for late-deafened educators. Context-aware modeling remains another critical limitation, as current systems often fail to incorporate situational factors, speaker traits, and conversational dynamics, which are vital for accurately interpreting emotional expressions in real-world interactions.

While the real-time evaluation provided valuable insights, audio recordings were not stored, and annotations were conducted live rather than through post-session review. Ground truth labelling was based on immediate observations during the session and aligned with predefined emotional guidelines given to participants. Although external annotations or self-reports were not incorporated at this stage, the evaluation method was effective in assessing system response within real-use conditions. Another limitation of the study was not incorporating multi-rater annotations or self-assessment data to enhance the validation process.

5.4 Future Work

Future work could explore enhancing the current SER model, which has achieved an overall accuracy of 92% on the combined dataset used in this study (including DEMoS, RAVDESS, CREMA-D, EMO-DB, and TESS), with the aim of improving its classification performance further using advanced architectures and optimization strategies. Additionally, emotional categories could be expanded to include more nuanced and education-specific states such as confusion, frustration, boredom, and satisfaction. Capturing these emotions would provide a deeper understanding of student engagement and learning difficulties. This would require the development or adaptation of speech datasets with appropriate annotations or the incorporation of multimodal data sources (e.g., facial expressions or behavioral cues) to improve detection accuracy and context awareness in real-world online teaching environments.

Future work could also focus on retraining or fine-tuning the SER model using more demographically diverse speech datasets, particularly those including younger speakers and a variety of regional or non-native English accents. This would improve the system's generalizability especially when applied in multicultural and multilingual learning environments.

Multimodal approach can be adopted by integrating textual analysis and facial expression recognition with the current SER system. This integration would improve the accuracy of emotion detection and provide a richer and more holistic understanding of student engagement by analysing multiple behavioural and communicative cues. For example, incorporating facial expression recognition could capture non-verbal emotional signals. At the same time, textual analysis of student's written responses or chat inputs could provide deeper insights into their sentiments and comprehension levels. Exploring cultural and contextual variations in emotional expression is particularly critical, as emotions can be displayed and interpreted differently across cultures, which could impact the system's effectiveness in global contexts. Hybrid architectures that combine CNNs with recurrent models like Long Short-Term Memory (LSTM) networks can further improve temporal emotion modeling. Expanding robust, diverse datasets and leveraging high-performance hardware to handle real-time speech, including non-native accents, could enable significant progress. Exploring alternative approaches such as unsupervised methods like auto-encoders can also provide valuable insights.

Furthermore, discussions with experts (late deafened educators who participated in the usability testing) have highlighted the potential for the SER system to be applied in other areas, such as supporting individuals with autism. For students on the autism spectrum who may struggle with traditional emotional expression or interpretation, the system could be adapted to provide tailored feedback to educators, helping them better understand and address the unique emotional and learning needs of these students. This broader applicability highlights the potential of the SER system to not only transform online education but also contribute to inclusive practices in diverse educational and therapeutic settings.

Future research could also expand on the current evaluation approach of multi-rater annotation, or participant self-reports to provide an even more comprehensive ground truth for real-time validation. This would further strengthen the robustness of system evaluation in live educational settings.

References

1 W. Li, Y. Zhang, and Y. Fu, “Speech Emotion Recognition in E-learning System Based on Affective Computing,” in *Natural Computation, 2007. ICNC 2007. Third International Conference on Volume: 5*, 2007, doi:10.1109/ICNC.2007.677.

2 T. S. Ashwin and R. M. R. Guddeti, “Unobtrusive Behavioral Analysis of Students in Classroom Environment Using Non-Verbal Cues,” *IEEE Access*, vol. 7, pp. 150693–150709, 2019, doi: 10.1109/ACCESS.2019.2947519.

3 Y. Zhou and X. Tao, “A Framework of Online Learning and Experiment System Based on Affective Computing,” in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Dec. 2020. doi: 10.1145/3453187.3453405

4 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. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

5 Singh, V., & Thurman, A. (2019). How Many Ways Can We Define Online Learning? A Systematic Literature Review of Definitions of Online Learning (1988-2018). American Journal of Distance Education, 33(4), 289–306. https://doi.org/10.1080/08923647.2019.1663082.

6 C. L. Svihus, “Online teaching in higher education during the COVID-19 pandemic,” *Educ Inf Technol (Dordr)*, vol. 29, no. 3, pp. 3175–3193, Feb. 2024, doi: 10.1007/s10639-023-11971-7.

7 A. P. Cavalcanti *et al.*, “Automatic feedback in online learning environments: A systematic literature review,” *Computers and Education: Artificial Intelligence*, vol. 2, 2021. doi: 10.1016/j.caeai.2021.100027

8. UNESCO IESALC, “COVID-19 and higher education: Today and tomorrow. Impact analysis, policy responses, and recommendations,” *UNESCO IESALC*. [Online]. Available: https://www.right-to-education.org/sites/right-to-education.org/files/resource-attachments/UNESCO_IESALC_Covid-19%20and%20higher%20education_2020_en.pdf.

9 L. Mishra, T. Gupta, and A. Shree, “Online teaching-learning in higher education during lockdown period of COVID-19 pandemic,” *International Journal of Educational Research Open*, vol. 1, Jan. 2020, doi: 10.1016/j.ijedro.2020.100012.

10 S. J. Daniel, “Education and the COVID-19 pandemic,” *Prospects (Paris)*, vol. 49, no. 1–2, pp. 91–96, Oct. 2020, doi: 10.1007/s11125-020-09464-3

11 E. Alqurashi, “Predicting student satisfaction and perceived learning within online learning environments,” *Distance Education*, vol. 40, no. 1, pp. 133–148, Jan. 2019, doi:10.1080/01587919.2018.1553562.

12 https://dictionary.apa.org/emotion

13 R. Pekrun, T. Goetz, W. Titz, and R. P. Perry, “Academic emotions in students’ self-regulated learning and achievement: A program of qualitative and quantitative research,” 2002, *Lawrence Erlbaum Associates Inc.* doi: 10.1207/S15326985EP3702_4

14 S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, “Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations,” *Information Fusion,* vol. 102, 2024. doi: 10.1016/j.inffus.2023.102019

15 D. Dupre *et al.*, “Oudjat: A configurable and usable annotation tool for the study of facial expressions of emotion,” *International Journal of Human Computer Studies*, vol. 83, pp. 51–61, Jul. 2015, doi: 10.1016/j.ijhcs.2015.05.010

16 Robert Plutchik, “The Nature of Emotions,” *American Scientist*, vol. 89, no. 4, 2001, available at: http://www.jstor.org/stable/27857503

17 L. Cen, F. Wu, Z. L. Yu, and F. Hu, “A Real-Time Speech Emotion Recognition System and its Application in Online Learning,” in *Emotions, Technology, Design, and Learning*, Elsevier, 2015, pp. 27–46. doi: 10.1016/B978-0-12-801856-9.00002-5.

18 https://nationaldeafcenter.org/wp-content/uploads/2022/11/Research-Brief-Late-Deafened.pdf

19 M. B. Akçay and K. Oğuz, “Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers,” Jan. 01, 2020, *Elsevier B.V.* https://doi.org/10.1016/j.specom.2019.12.001

20 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. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124

21 K. Sarmah, S. Gogoi, H. C. Das, B. Patir, and J. Sarma, “A State-of-arts Review of Deep Learning Techniques for Speech Emotion Recognition,” *Journal of Electrical Systems*, vol. 20, no. 7, 2024. doi: 10.52783/jes.3745

22 Dennis Galletta and Ping Zhang, *Human-computer Interaction and Management Information Systems: Applications – Applications of Human-computer Interaction in Management Information Systems*. M.E. Sharpe, 2006

23 F. Karray, M. Alemzadeh, J. A. Saleh, and M. N. Arab, “Human-Computer Interaction: Overview on State of the Art,” *International Journal on Smart Sensing and Intelligent Systems*, 2008, doi: 10.21307/ijssis-2017-283

24 J. Abascal and C. Nicolle, “Moving towards inclusive design guidelines for socially and ethically aware HCI,” *Interact Comput*, vol. 17, no. 5, pp. 484–505, 2005, doi: 10.1016/j.intcom.2005.03.002

25 A. Oulasvirta and K. Hornbæk, “HCI research as problem-solving,” in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2016, pp. 4956–4967. doi: 10.1145/2858036.2858283

26 Hossain, M. I. (2023). *Software Development Life Cycle (SDLC) Methodologies for Information Systems Project Management*. www.ijfmr.com

27 M. Shah Fahad, A. Ranjan, J. Yadav, and A. Deepak, “A survey of speech emotion recognition in natural environment,” *Digital Signal Processing*, vol. 110, 2021. doi: 10.1016/j.dsp.2020.102951.

28 L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *J Big Data*, vol. 8, no. 1, Dec. 2021, doi: 10.1186/s40537-021-00444-8.

29 R. Tidwell, “The ‘invisible’ faculty member: The university professor with a hearing disability,” Kluwer Academic Publishers, 2004. [Online]. Available: https://www.jstor.org/stable/4151539?seq=1&cid=pdf-reference#references_tab_contents.

30 D. H. Smith and J. F. Andrews, “Deaf and hard of hearing faculty in higher education: enhancing access, equity, policy, and practice,” *Disabil Soc*, vol. 30, no. 10, pp. 1521–1536, Nov. 2015, doi: 10.1080/09687599.2015.1113160

31 G. V. Subba Reddy, K. Srivatsav, M. Ganganagari, and Y. Bsalasurya, “CNN-Enhanced Speech Emotion Recognition with Recommendations,” in *1st International Conference on Electronics, Computing, Communication and Control Technology, ICECCC 2024*, Institute of Electrical and Electronics Engineers Inc., 2024. doi: 10.1109/ICECCC61767.2024.10593896.

32 S. Ramakrishnan, ”Recognition of Emotion from Speech: A Review, Speech Enhancement, Modeling and Recognition- Algorithms and Applications”, 2012, ISBN: 978-953-51-0291-5, InTech, Available from: http://www.intechopen.com/books/speech-enhancement-modeling-and-recognition-algorithmsand-applications/recognition-of-emotion-from-speech-a-review.

33 Robert Plutchik, “The Nature of Emotions,” *American Scientist*, vol. 89, no. 4, 2001, available at: http://www.jstor.org/stable/27857503

34 G. H. Mohmad and R. Delhibabu, “Speech Databases, speech features and classifiers in speech emotion recognition : A Review,” *IEEE Access*, 2024, doi: 10.1109/ACCESS.2024.3476960.

35 Mustaqeem and S. Kwon, “Optimal feature selection based speech emotion recognition using two-stream deep convolutional neural network,” *International Journal of Intelligent Systems*, vol. 36, no. 9, pp. 5116–5135, Sep. 2021, doi: 10.1002/int.22505.

36 T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, “A Comprehensive Review of Speech Emotion Recognition Systems,” 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2021.3068045.

37 P. Ekman, S. L. David, R. Matsumoto, H. Oster, E. L. Rosenberg, and K. R. Scherer, “Facial Expression and Emotion,” *American Psychological Association*, Apr. 1993, available at: https://sanlab.psych.ucla.edu/wp-content/uploads/sites/31/2016/03/Ekman-American_Psychologist_1993.pdf

38 M. D. Pell, S. Paulmann, C. Dara, A. Alasseri, and S. A. Kotz, “Factors in the recognition of vocally expressed emotions: A comparison of four languages,” *J Phon*, vol. 37, no. 4, pp. 417–435, Oct. 2009, doi: 10.1016/j.wocn.2009.07.005.

39 J. A. Russell and A. Mehrabian, “Evidence for a three-factor theory of emotions,” *J Res Pers*, vol. 11, no. 3, pp. 273–294, 1977, doi: 10.1016/0092-6566(77)90037-X

40 Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, “A survey of affect recognition methods: Audio, visual, and spontaneous expressions,” *IEEE Trans Pattern Anal Mach Intell*, vol. 31, no. 1, pp. 39–58, 2009, doi: 10.1109/TPAMI.2008.52.

41 I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN COMPUT. SCI*, vol. 2, 2021, *Springer*. doi: 10.1007/s42979-021-00815-1

42 E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, “A review on speech emotion recognition using deep learning and attention mechanism,” May 02, 2021, *MDPI AG*. doi: 10.3390/electronics10101163

43 B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, “Deep learning techniques for speech emotion recognition, from databases to models,” Feb. 02, 2021, *MDPI AG*. doi: 10.3390/s21041249

44 A. Hashem, M. Arif, and M. Alghamdi, “Speech emotion recognition approaches: A systematic review,” *Speech Communication*, vol. 154, 2023. doi: 10.1016/j.specom.2023.102974

45 M. Swain, A. Routray, and P. Kabisatpathy, “Databases, features and classifiers for speech emotion recognition: a review,” *Int J Speech Technol*, vol. 21, no. 1, pp. 93–120, Mar. 2018, doi: 10.1007/s10772-018-9491

46 J. de Lope and M. Graña, “An ongoing review of speech emotion recognition,” *Neurocomputing*, vol. 528, pp. 1–11, Apr. 2023, doi: 10.1016/j.neucom.2023.01.002.

47 S. Zhang, R. Liu, X. Tao, and X. Zhao, “Deep Cross-Corpus Speech Emotion Recognition: Recent Advances and Perspectives,” Nov. 29, 2021, *Frontiers Media S.A.* doi: 10.3389/fnbot.2021.784514.

48 B. M. Nema and A. A. Abdul-Kareem, “Preprocessing signal for Speech Emotion Recognition,” *Al-Mustansiriyah Journal of Science*, vol. 28, no. 3, pp. 157–165, Jul. 2018, doi: 10.23851/mjs.v28i3.48

49 Turgut Özseven, “2nd International Symposium on Multidisciplinary Studies and Innovative Technologies : ISMSIT 2018 : proceedings : 19-21 October 2018, Kızılcahamam/Ankara/Turkey,” IEEE, 2018

50 C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, “Iterative feature normalization scheme for automatic emotion detection from speech,” *IEEE Trans Affect Comput*, vol. 4, no. 4, pp. 386–397, 2013, doi: 10.1109/T-AFFC.2013.26

51 T. J. Sefara, “The Effects of Normalisation Methods on Speech Emotion Recognition,” in *Proceedings - 2019 International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019. doi: 10.1109/IMITEC45504.2019.9015895

52 J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, “Spectral and cepstral audio noise reduction techniques in speech emotion recognition,” in *MM 2016 - Proceedings of the 2016 ACM Multimedia Conference*, Association for Computing Machinery, Inc, Oct. 2016, pp. 670–674. doi: 10.1145/2964284.2967306

53 M. Yu, Q. Wang, J. Xu, and J. Zhang, “Improving Speech Emotion Recognition with Dynamic Features and Temporal Correlations,” in *2024 11th International Conference on Machine Intelligence Theory and Applications (MiTA)*, IEEE, Jul. 2024, pp. 1–8. doi: 10.1109/MiTA60795.2024.10751729

54 S. Kuchibhotla, H. D. Vankayalapati, R. S. Vaddi, and K. R. Anne, “A comparative analysis of classifiers in emotion recognition through acoustic features,” *Int J Speech Technol*, vol. 17, no. 4, pp. 401–408, Oct. 2014, doi: 10.1007/s10772-014-9239-3

55 Harshawardhan S. Kumbhar and Sheetal U. Bhandari, “Speech Emotion Recognition using MFCC features and LSTM network,” in *2019 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA)*, IEEE, 2019

56 A. Koduru, H. B. Valiveti, and A. K. Budati, “Feature extraction algorithms to improve the speech emotion recognition rate,” *Int J Speech Technol*, vol. 23, no. 1, pp. 45–55, Mar. 2020, doi: 10.1007/s10772-020-09672-4.

57 M. J. Al-Dujaili and A. Ebrahimi-Moghadam, “Speech Emotion Recognition: A Comprehensive Survey,” *Wireless Personal Communications*, vol. 129, no. 4, 2023. doi: 10.1007/s11277-023-10244-3

58 A. Jacob, “Speech emotion recognition based on minimal voice quality features,” in *ICCSP 2016*, IEEE Inc., Nov. 2016, pp. 886–890. doi: 10.1109/ICCSP.2016.7754275 available at: https://ieeexplore.ieee.org/document/7754275

59 G. Zhou, J. H. L. Hansen, and J. F. Kaiser, “Classification of Speech under Stress Based on Features Derived from the Nonlinear Teager Energy Operator,” in *Conference: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference onVolume: 1*, doi: 10.1109/ICASSP.1998.674489

60 Mustaqeem and S. Kwon, “A CNN-assisted enhanced audio signal processing for speech emotion recognition,” *Sensors (Switzerland)*, vol. 20, no. 1, Jan. 2020, doi: 10.3390/s20010183

61 R. Rani and M. K. Ramaiya, “Detection of Emotions from Speech using Deep Learning Techniques and Traditional Techniques: A Survey,” in *2nd International Conference on Automation, Computing and Renewable Systems, ICACRS 2023 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 1202–1209. doi: 10.1109/ICACRS58579.2023.10404716.

62 Y. Xu and R. Goodacre, “On Splitting Training and Validation Set: A Comparative Study of Cross-Validation, Bootstrap and Systematic Sampling for Estimating the Generalization Performance of Supervised Learning,” *J Anal Test*, vol. 2, no. 3, pp. 249–262, Jul. 2018, doi: 10.1007/s41664-018-0068-2

63 H. W. B. S. and J. L. Huan Wan1, “A Novel Gaussian Mixture Model for Classification,” in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) Bari, Italy. October 6-9, 2019*, IEEE.

64 D. T. G. Z. P. C. C. and T. L. Shuiyang Mao, “REVISITING HIDDEN MARKOV MODELS FOR SPEECH EMOTION RECOGNITION,” in *ICASSP 2019*, Institute of Electrical and Electronics Engineers, 2018, p. 465

65 C. Yu, Q. Tian, F. Cheng, and S. Zhang, “Speech Emotion Recognition Using Support Vector Machines,” *CCIS*, vol. 152, pp. 215–220, 2011

66 M. Granik, V. Mesyura, “Fake News Detection Using Naive Bayes Classifier,” in *UKRCON : 2017 IEEE first Ukraine Conference on Electrical and Computer Engineering : conference proceedings : May 29-June 2, 2017, Kyiv, Ukraine*, IEEE, 2017

67 Atreyee Khan and Uttam Kumar Roy, “Emotion Recognition Using Prosodic and Spectral Features of Speech and Naïve Bayes Classifier,” in *Proceedings of the 2017 International Conference on Wireless*

Communications, Signal Processing and Networking (WiSPNET) : 22-24 March 2017, Chennai, India, IEEE, 2017.DOI**:** 10.1109/WiSPNET.2017.8299916

68 Z. Yao, Z. Wang, W. Liu, Y. Liu, and J. Pan, "Speech emotion recognition using fusion of three multi-task learning-based classifiers: HSF-DNN, MS-CNN and LLD-RNN," *Speech Commun*, vol. 120, pp. 11–19, Jun. 2020, doi: 10.1016/j.specom.2020.03.005

69 V. H. Phung and E. J. Rhee, "A High-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Applied Sciences (Switzerland)*, vol. 9, no. 21, Nov. 2019, doi: 10.3390/app9214500

70 S. Zhang, X. Tao, Y. Chuang, and X. Zhao, "Learning deep multimodal affective features for spontaneous speech emotion recognition," *Speech Commun*, vol. 127, pp. 73–81, Mar. 2021, doi: 10.1016/j.specom.2020.12.009

71 A. Amjad, L. Khan, N. Ashraf, M. B. Mahmood, and H. T. Chang, "Recognizing Semi-Natural and Spontaneous Speech Emotions Using Deep Neural Networks," *IEEE Access*, vol. 10, pp. 37149–37163, 2022, doi: 10.1109/ACCESS.2022.3163712

72 S. Zhang, X. Zhao, and Q. Tian, "Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM," *IEEE Trans Affect Comput*, vol. 13, no. 2, pp. 680–688, 2022, doi: 10.1109/TAFFC.2019.2947464

73 M. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," *Expert Syst Appl*, vol. 218, May 2023, doi: 10.1016/j.eswa.2023.119633

74 C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation Functions: Comparison of trends in Practice and Research for Deep Learning," *arXiv:1811.03378v1*, Nov. 2018, doi: https://doi.org/10.48550/arXiv.1811.03378

75 S. Ullah, Q. A. Sahib, Faizullah, S. Ullahh, I. U. Haq, and I. Ullah, "Speech Emotion Recognition Using Deep Neural Networks," in *2022 International Conference on IT and Industrial Technologies, ICIT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICIT56493.2022.9989197

76 H. Wang and B. Raj, "On the Origin of Deep Learning," *ArXiv*, Feb. 2017, doi: https://doi.org/10.48550/arXiv.1702.07800

77 G. Wen, H. Li, J. Huang, D. Li, and E. Xun, "Random Deep Belief Networks for Recognizing Emotions from Speech Signals," *Comput Intell Neurosci*, vol. 2017, 2017, doi: 10.1155/2017/1945630

78 Z. Zhao, Y. Zheng, Z. Zhang, H. Wang, Y. Zhao, and C. Li, "Exploring spatio-temporal representations by integrating attention-based bidirectional-LSTM-RNNs and FCNs for speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, International Speech Communication Association, 2018, pp. 272–276. doi: 10.21437/Interspeech.2018-1477

79 N. Hajarolasvadi and H. Demirel, "3D CNN-based speech emotion recognition using k-means clustering and spectrograms," *Entropy*, vol. 21, no. 5, May 2019, doi: 10.3390/e21050479.

80 A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, “Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network,” in *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Mar. 2017. doi: 10.1109/PlatCon.2017.7883728.

81 R. B. Lanjewar, S. Mathurkar, and N. Patel, “Implementation and comparison of speech emotion recognition system using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) techniques,” in *Procedia Computer Science*, Elsevier B.V., 2015, pp. 50–57. doi: 10.1016/j.procs.2015.04.226

82 L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, “Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction,” *Inf Sci (N Y)*, vol. 509, pp. 150–163, Jan. 2020, doi: 10.1016/j.ins.2019.09.005

83 S. Langari, H. Marvi, and M. Zahedi, “Efficient speech emotion recognition using modified feature extraction,” *Inform Med Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100424

84 M. B. Er, “A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features,” *IEEE Access*, vol. 8, pp. 221640–221653, 2020, doi: 10.1109/ACCESS.2020.3043201

85 J. Ancilin and A. Milton, “Improved speech emotion recognition with Mel frequency magnitude coefficient,” *Applied Acoustics*, vol. 179, Aug. 2021, doi: 10.1016/j.apacoust.2021.108046.

86 S. Yildirim, Y. Kaya, and F. Kılıç, “A modified feature selection method based on metaheuristic algorithms for speech emotion recognition,” *Applied Acoustics*, vol. 173, Feb. 2021, doi: 10.1016/j.apacoust.2020.107721

87 T. J. Sefara, “The Effects of Normalisation Methods on Speech Emotion Recognition,” in *Proceedings - 2019 International Multidisciplinary Information Technology and Engineering Conference, IMITEC 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019. doi: 10.1109/IMITEC45504.2019.9015895.

88 Sagar K. Bhakre and Arti Bang, “Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier,” in *2016 International Conference on Advances in Computing, Communications and Informatics : Jaipur, India*, IEEE, p. 148. Available at : https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7732408

89 M. Swain, S. Sahoo, A. Routray, P. Kabisatpathy, and J. N. Kundu, “Study of feature combination using HMM and SVM for multilingual Odiya speech emotion recognition,” *Int J Speech Technol*, vol. 18, no. 3, pp. 387–393, Sep. 2015, doi: 10.1007/s10772-015-9275-7

90 A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, “Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network,” *Applied Sciences (Switzerland)*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13084750

91 Mustaqeem, M. Sajjad, and S. Kwon, “Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM,” *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405

92 Mustaqeem and S. Kwon, “CLSTM: Deep feature-based speech emotion recognition using the hierarchical convlstm network,” *Mathematics*, vol. 8, no. 12, pp. 1–19, Dec. 2020, doi: 10.3390/math8122133.

93 Mustaqeem and S. Kwon, “MLT-DNet: Speech emotion recognition using 1D dilated CNN based on multi-learning trick approach,” *Expert Syst Appl*, vol. 167, Apr. 2021, doi: 10.1016/j.eswa.2020.114177

94 O. Atila and A. Şengür, “Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition,” *Applied Acoustics*, vol. 182, Nov. 2021, doi: 10.1016/j.apacoust.2021.108260.

95 S. Ntalampiras, “Speech emotion recognition via learning analogies,” *Pattern Recognit Lett*, vol. 144, pp. 21–26, Apr. 2021, doi: 10.1016/j.patrec.2021.01.018

96 L. Pepino, P. Riera, and L. Ferrer, “Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings,” *arXiv:2104.03502v1*, Apr. 2021, doi: 10.48550/arXiv.2104.03502

97 S. Li, X. Xing, W. Fan, B. Cai, P. Fordson, and X. Xu, “Spatiotemporal and frequential cascaded attention networks for speech emotion recognition,” *Neurocomputing*, vol. 448, pp. 238–248, Aug. 2021, doi: 10.1016/j.neucom.2021.02.094

98 K. Mao, Y. Wang, L. Ren, J. Zhang, J. Qiu, and G. Dai, “Multi-branch feature learning based speech emotion recognition using SCAR-NET,” *Conn Sci*, vol. 35, no. 1, 2023, doi: 10.1080/09540091.2023.2189217

99 L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, “Speech emotion recognition based on DNN-decision tree SVM model,” *Speech Commun*, vol. 115, pp. 29–37, Dec. 2019, doi: 10.1016/j.specom.2019.10.004

100 T. M. Wani, T. S. Gunawan, S. A. A. Qadri, H. Mansor, M. Kartiwi, and N. Ismail, “Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks,” in *Proceedings - 2020 6th International Conference on Wireless and Telematics, ICWT 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020. doi: 10.1109/ICWT50448.2020.9243622

101 C. Wang, Y. Ren, N. Zhang, F. Cui, and S. Luo, “Speech emotion recognition based on multi-feature and multi-lingual fusion,” *Multimed Tools Appl*, vol. 81, no. 4, pp. 4897–4907, Feb. 2022, doi: 10.1007/s11042-021-10553-4

102 B. Z. Mansouri, H. R. Ghaffary, and A. Harimi, “Speech Emotion Recognition using Sub-Band Spectrogram fusion and Deep Convolutional Neural Network transfer learning,” *Research Square*, Dec. 16, 2022. doi: 10.21203/rs.3.rs-2369713/v1

103 A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, “LIGHT-SERNET: A LIGHTWEIGHT FULLY CONVOLUTIONAL NEURAL NETWORK FOR SPEECH EMOTION RECOGNITION,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, IEEE Inc., 2022, pp. 6912–6916. doi: 10.1109/ICASSP43922.2022.9746679

104 H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, “SPEECH EMOTION RECOGNITION WITH CO-ATTENTION BASED MULTI-LEVEL ACOUSTIC INFORMATION,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 7367–7371. doi: 10.1109/ICASSP43922.2022.9747095

105 J. Zhao, X. Mao, and L. Chen, “Speech emotion recognition using deep 1D & 2D CNN LSTM networks,” *Biomed Signal Process Control*, vol. 47, pp. 312–323, Jan. 2019, doi: 10.1016/j.bspc.2018.08.035

106 Z. Gao, W. Xiao, W. Zhou, and Z. Yang, “FPGA Implementation of CNN-LSTM Classifier in Speech Emotion Recognition System,” in *2023 International Conference on High Performance Big Data and Intelligent Systems, HDIS 2023*, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 47–52. doi: 10.1109/HDIS60872.2023.10499604

107 Z. Huijuan, Y. Ning, and W. Ruchuan, “Coarse-to-Fine Speech Emotion Recognition Based on Multi-Task Learning,” *J Signal Process Syst*, vol. 93, no. 2–3, pp. 299–308, Mar. 2021, doi: 10.1007/s11265-020-01538-x

108 Jianyou Wang, Michael Xue, Ryan Culhane, Enmao Diao, Jie Ding and Vahid Tarokh, “Speech Emotion Recognition with Dual-sequence LSTM Architecture,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, p. 9304. doi: 10.1109/ICASSP40776.2020.9054629

109 X. Wang, M. Wang, W. Qi, W. Su, X. Wang, and H. Zhou, “A novel end-to-end speech emotion recognition network with stacked transformer layers,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6289–6293. doi: 10.1109/ICASSP39728.2021.9414314

110 S. Latif, R. Rana, S. Khalifa, R. Jurdak, and J. Epps, “Direct Modelling of Speech Emotion from Raw Speech,” *ArXiv*, Apr. 2019, doi: 10.48550/arXiv.1904.03833

111 Q. Chen and G. Huang, “A novel dual attention-based BLSTM with hybrid features in speech emotion recognition,” *Eng Appl Artif Intell*, vol. 102, Jun. 2021, doi: 10.1016/j.engappai.2021.104277

112 Z. Zhao *et al.*, “Combining a parallel 2D CNN with a self-attention Dilated Residual Network for CTC-based discrete speech emotion recognition,” *Neural Networks*, vol. 141, pp. 52–60, Sep. 2021, doi: 10.1016/j.neunet.2021.03.013

113 M. M. Hassan, M. G. R. Alam, M. Z. Uddin, S. Huda, A. Almogren, and G. Fortino, “Human emotion recognition using deep belief network architecture,” *Information Fusion*, vol. 51, pp. 10–18, Nov. 2019, doi: 10.1016/j.inffus.2018.10.009

114 M. J. Hasan, M. S. Hossain, S. M. N. Hassan, M. Al-Amin, M. N. Rahaman, and M. A. Pranjol, “Bengali Speech Emotion Recognition: A hybrid approach using B-LSTM,” in *4th International Conference on Electrical, Computer and Telecommunication Engineering, ICECTE 2022*, Institute of Electrical and Electronics Engineers Inc., 2022. doi: 10.1109/ICECTE57896.2022.10114510

115 I. Shahin, A. B. Nassif, and S. Hamsa, “Emotion Recognition Using Hybrid Gaussian Mixture Model and Deep Neural Network,” *IEEE Access*, vol. 7, pp. 26777–26787, 2019, doi: 10.1109/ACCESS.2019.2901352

116 S. T. Rajamani, K. T. Rajamani, A. Mallol-Ragolta, S. Liu, and B. Schuller, “A novel attention-based gated recurrent unit and its efficacy in speech emotion recognition,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6294–6298. doi: 10.1109/ICASSP39728.2021.9414489

117 K. S. Chintalapudi, I. A. K. Patan, H. V. Sontineni, V. S. K. Muvvala, S. V. Gangashetty, and A. K. Dubey, "Speech Emotion Recognition Using Deep Learning," in *2023 International Conference on Computer Communication and Informatics, ICCCI 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCCI56745.2023.10128612.

118 H. Meng, T. Yan, F. Yuan, and H. Wei, "Speech Emotion Recognition from 3D Log-Mel Spectrograms with Deep Learning Network," *IEEE Access*, vol. 7, pp. 125868–125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

119 C. J. Tanis, "The seven principles of online learning: Feedback from faculty and alumni on its importance for teaching and learning," Research in Learning Technology, vol. 28, 2020, doi: 10.25304/rlt.v28.2319.

120 G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, Jan. 2020, doi: 10.1016/j.apacoust.2019.107020

121 M. B. Er, "A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features," *IEEE Access*, vol. 8, pp. 221640–221653, 2020, doi: 10.1109/ACCESS.2020.3043201

122 C. Murad, C. Munteanu, B. R. Cowan, and L. Clark, "Revolution or Evolution? Speech Interaction and HCI Design Guidelines," *IEEE Pervasive Comput*, vol. 18, no. 2, pp. 33–45, Apr. 2019, doi: 10.1109/MPRV.2019.2906991

123 World Health Organization, "Hearing loss," *World Health Organization*. [Online]. Available: https://www.who.int/health-topics/hearing-loss#tab=tab_1.

124 S. Garg, C. Deshmukh, M. Singh, A. Borle, and B. Wilson, "Challenges of the deaf and hearing impaired in the masked world of COVID-19," Jan. 01, 2021, *Wolters Kluwer Medknow Publications*. doi: 10.4103/ijcm.IJCM_581_20

125 C. Rivas-Costa, L. Anido-Rifón, M. J. Fernández-Iglesias, M. A. Gómez-Carballa, S. Valladares-Rodríguez, and R. Soto-Barreiros, "An Accessible Platform for People With Disabilities," *Int J Hum Comput Interact*, vol. 30, no. 6, pp. 480–494, 2014, doi: 10.1080/10447318.2014.888503

126 W. Farhan and J. Razmak, "A comparative study of an assistive e-learning interface among students with and without visual and hearing impairments," *Disabil Rehabil Assist Technol*, vol. 17, no. 4, pp. 431–441, 2022, doi: 10.1080/17483107.2020.1786733

127 M. Seita, "Designing automatic speech recognition technologies to improve accessibility for deaf and hard-of-hearing people in small group meetings," in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, Apr. 2020. doi: 10.1145/3334480.3375039

128 B. Baglama, M. Haksiz, and H. Uzunboylu, "Technologies used in education of hearing impaired individuals," *International Journal of Emerging Technologies in Learning*, vol. 13, no. 9, pp. 53–63, 2018, doi: 10.3991/ijet.v13i09.8303

129 Amandeep and Williamjeet, "Tools and Techniques of Assistive Technology for Hearing Impaired People," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019*, [IEEE], 2019.**DOI:** 10.1109/COMITCon.2019.8862454

130 H. Kim, H. Hwang, S. Gwak, J. Yoon, and K. Park, “Improving communication and promoting social inclusion for hearing-impaired users: Usability evaluation and design recommendations for assistive mobile applications,” *PLoS One*, vol. 19, no. 7 July, Jul. 2024, doi: 10.1371/journal.pone.0305726

131 A. Flores Ramones and M. S. del-Rio-Guerra, “Recent Developments in Haptic Devices Designed for Hearing-Impaired People: A Literature Review,” Mar. 01, 2023, *MDPI*. doi: 10.3390/s23062968

132 S. Serafin, A. Adjorlu, and L. M. Percy-Smith, “A Review of Virtual Reality for Individuals with Hearing Impairments,” Apr. 01, 2023, *MDPI*. doi: 10.3390/mti7040036

133 M. Palanisamy, R. Mohanraj, A. Karthikeyan, and E. Mohanraj, “SIGNEASE: AI-Driven American Sign Language Interpretation System,” in *2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS)*, IEEE, Dec. 2024, pp. 1670–1675. doi: 10.1109/ICICNIS64247.2024.10823126

134 G. Kbar, A. Bhatia, M. H. Abidi, and I. Alsharawy, “Assistive technologies for hearing, and speaking impaired people: a survey,” Jan. 02, 2017, *Taylor and Francis Ltd*. doi: 10.3109/17483107.2015.1129456

135 S. Pargaonkar, “A Comprehensive Research Analysis of Software Development Life Cycle (SDLC) Agile & Waterfall Model Advantages, Disadvantages, and Application Suitability in Software Quality Engineering,” *International Journal of Scientific and Research Publications*, vol. 13, no. 8, pp. 120–124, Aug. 2023, doi: 10.29322/ijsrp.13.08.2023.p14015.

136 Gurianov, D. A., Myshenkov, K. S., & Terekhov, V. I. (2023). Software Development Methodologies: Analysis and Classification. Proceedings of the 2023 5th International Youth Conference on Radio Electronics, Electrical and Power Engineering, REEPE 2023. https://doi.org/10.1109/REEPE57272.2023.10086852

137 U. Kuter and C. Yilmaz, “Survey Methods: Questionnaires and Interviews Choosing Human-Computer Interaction (HCI) Appropriate Research Methods Survey Methods: Questionnaires and Interviews,” 2014. [Online]. Available: https://www.researchgate.net/publication/267366565

138 J. Linåker, S. M. Sulaman, M. Höst, and R. Maiani De Mello, “Guidelines for Conducting Surveys in Software Engineering v. 1.1,” 2015. Available on: https://www.researchgate.net/publication/276062061_Guidelines_for_Conducting_Surveys_in_Software_Engineering#fullTextFileContent.

139 J. Kjeldskov and C. Graham, “LNCS 2795 - A Review of Mobile HCI Research Methods,” in *LNCS*, Springer-Verlag, 2003, pp. 317–335. doi: 10.1007/978-3-540-45233-1_23.

140 Neetij and R. Bikash Thapa, “A Study on Purposive Sampling Method in Research,” 2004. [Online]. Available:https://www.academia.edu/28087388/A_STUDY_ON_PURPOSIVE_SAMPLING_METHOD_IN_RESEARCH.

141 P. Victor and T. Redondo, “Purposive Sampling in the Analysis of Count Data.” Available on: https://www.researchgate.net/publication/312060163_Purposive_sampling_in_the_analysis_of_count_data_paolo#fullTextFileContent

142 D. L. Lima, R. De Souza Santos, G. P. Garcia, S. S. Da Silva, C. Franca, and L. F. Capretz, “Software Testing and Code Refactoring: A Survey with Practitioners,” in *Proceedings - 2023 IEEE International*

Conference on Software Maintenance and Evolution, ICSME 2023, IEEE Inc., 2023, pp. 500–507. doi: 10.1109/ICSME58846.2023.00064.

143 H. Ames, C. Glenton, and S. Lewin, "Purposive sampling in a qualitative evidence synthesis: A worked example from a synthesis on parental perceptions of vaccination communication," *BMC Med Res Methodol*, vol. 19, no. 1, Jan. 2019, doi: 10.1186/s12874-019-0665-4.

144 J. Lazar, J. H. Feng, and H. Hochheiser, "Working with research participants with disabilities," in *Research Methods in Human Computer Interaction*, Elsevier, 2017, pp. 493–522. doi: 10.1016/b978-0-12-805390-4.00016-9

145 A. B. Hardi, E. Simorangkir, I. Hutagaol, W. J. W. Saputra, and Sunardi, "User Experience Analysis on Mobile Banking Applications with System Usability Scale and Usability Testing," in *2023 IEEE 9th International Conference on Computing, Engineering and Design, ICCED 2023*, IEEE Inc., 2023. doi: 10.1109/ICCED60214.2023.10424703

146 B. Mike Kuniavsky, *Observing the User Experience: A Practitioner's Guide to User Research*. Morgan Kaufmann, 2003. Avaliable at: http://www.orangecone.com/Macromedia_SWU-eGovBP_F_chapter_7.pdf

147 V. Nasr and M. Zahabi, "Usability Evaluation Methods of Indoor Navigation Apps for People with Disabilities: A Scoping Review," in *Proceedings of the 2022 IEEE International Conference on Human-Machine Systems, ICHMS 2022*, IEEE Inc., 2022. doi: 10.1109/ICHMS56717.2022.9980809

148 Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359. https://doi.org/10.1007/s10579-008-9076-6.

149 E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. W. Schuller, "DEMoS: an Italian emotional speech corpus: Elicitation methods, machine learning, and perception," *Lang Resour Eval*, vol. 54, no. 2, pp. 341–383, Jun. 2020, doi: 10.1007/s10579-019-09450-y.

150 M. M. M. Islam, M. A. Kabir, A. Sheikh, M. Saiduzzaman, A. Hafid, and S. Abdullah, "Enhancing Speech Emotion Recognition Using Deep Convolutional Neural Networks," in *2024 9th International Conference on Machine Learning Technologies (ICMLT)*, New York, NY, USA: ACM, May 2024, pp. 95–100. doi: 10.1145/3674029.3674045

151 F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology*, 2005, pp. 1517–1520. doi: 10.21437/interspeech.2005-446.

152 M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–18, Nov. 2020, doi: 10.3390/s20216008

153 G. Kbar, A. Bhatia, M. H. Abidi, and I. Alsharawy, "Assistive technologies for hearing, and speaking impaired people: a survey," Jan. 02, 2017, *Taylor and Francis Ltd*. doi: 10.3109/17483107.2015.1129456

154 D. Luo, Y. Zou, and D. Huang, "Investigation on joint representation learning for robust feature extraction in speech emotion recognition," in *Proceedings of the Annual Conference of the International Speech*

Communication Association, INTERSPEECH, International Speech Communication Association, 2018, pp. 152–156. doi: 10.21437/Interspeech.2018-1832

155 K. Zvarevashe and O. Olugbara, "Ensemble learning of hybrid acoustic features for speech emotion recognition," *Algorithms*, vol. 13, no. 3, Mar. 2020, doi: 10.3390/a13030070

156 A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network," in *2017 International Conference on Platform Technology and Service, PlatCon 2017 - Proceedings*, Institute of Electrical and Electronics Engineers Inc., Mar. 2017. doi: 10.1109/PlatCon.2017.7883728

157 A. Bapa, O. Bandgar, A. Ekapure, and J. Sisodia, "Respiratory disorder classification based on lung auscultation using MFCC, Mel Spectrogram and Chroma STFT," in *2023 International Conference on Artificial Intelligence and Applications, ICAIA 2023 and Alliance Technology Conference, ATCON-1 2023 - Proceeding*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICAIA57370.2023.10169299

158 D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed Signal Process Control*, vol. 59, May 2020, doi: 10.1016/j.bspc.2020.101894

159 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. doi: 10.1109/RADIOELEK.2019.8733432.

160 S. Zhang, X. Tao, Y. Chuang, and X. Zhao, "Learning deep multimodal affective features for spontaneous speech emotion recognition," *Speech Commun*, vol. 127, pp. 73–81, Mar. 2021, doi: 10.1016/j.specom.2020.12.009

161 A. Shamsaldin, P. Fattah, T. Rashid, and N. Al-Salihi, "A Study of The Convolutional Neural Networks Applications," *UKH Journal of Science and Engineering*, vol. 3, no. 2, pp. 31–40, Dec. 2019, doi: 10.25079/ukhjse.v3n2y2019.pp31-40

162 D. Bhatt *et al.*, "Cnn variants for computer vision: History, architecture, application, challenges and future scope," Oct. 01, 2021, *MDPI.* doi: 10.3390/electronics10202470

163 S. T. Seydi, M. Hasanlou, and M. Amani, "A new end-to-end multi-dimensional CNN framework for land cover/land use change detection in multi-source remote sensing datasets," *Remote Sens (Basel)*, vol. 12, no. 12, Jun. 2020, doi: 10.3390/rs12122010

164 H. Tang, Y. Li, Z. Huang, L. Zhang, and W. Xie, "Fusion of Multidimensional CNN and Handcrafted Features for Small-Sample Hyperspectral Image Classification," *Remote Sens (Basel)*, vol. 14, no. 15, Aug. 2022, doi: 10.3390/rs14153796

165 A. Elboushaki, R. Hannane, K. Afdel, and L. Koutti, "MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences," *Expert Syst Appl*, vol. 139, Jan. 2020, doi: 10.1016/j.eswa.2019.112829

166 J. Liu, T. Wang, A. Skidmore, Y. Sun, P. Jia, and K. Zhang, "Integrated 1D, 2D, and 3D CNNs Enable Robust and Efficient Land Cover Classification from Hyperspectral Imagery," *Remote Sens (Basel)*, vol. 15, no. 19, Oct. 2023, doi: 10.3390/rs15194797

167 J. A. Aghamaleki and V. Ashkani Chenarlogh, “Multi-stream CNN for facial expression recognition in limited training data,” *Multimed Tools Appl*, vol. 78, no. 16, pp. 22861–22882, Aug. 2019, doi: 10.1007/s11042-019-7530-7

168 A.Ross, A.K. Jain and K Nandakumar, *Chapter 4 SCORE LEVEL FUSION*. Springer, Boston, MA, 2006. doi: https://doi.org/10.1007/0-387-33123-9_4.

169 R. Geetha, T. Thilagam, and T. Padmavathy, “Effective offline handwritten text recognition model based on a sequence-to-sequence approach with CNN–RNN networks,” Sep. 01, 2021, *Springer Science and Business Media Deutschland GmbH*. doi: 10.1007/s00521-020-05556-5

170 J. R. Lewis and J. Sauro, “Usability and User Experience: Design and Evaluation,” in *Handbook of Human Factors and Ergonomics*, wiley, 2021, pp. 972–1015. doi: 10.1002/9781119636113.ch38

171 A. Rana and S. Kumar Dubey, “Analytical Roadmap to Usability Definitions and Decompositions,” 2010. [Online]. Available: https://www.researchgate.net/publication/282848700.

172 E. R. Ro, K. O. An, A. J. Kim, S. U. Jang, E. J. Kim, and S. D. Eun, “Usability Study to Promote Co-Creation among People with Disabilities, Developers, and Makers with a Focus on the Assistive Technology Open Platform in Korea,” *IEEE Access*, vol. 12, pp. 39016–39027, 2024, doi: 10.1109/ACCESS.2023.3345036.

173 A. Valerian, H. B. Santoso, M. Schrepp and G. Guarddin, "Usability Evaluation and Development of a University Staff Website," *2018 Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, 2018, pp. 1-6, doi: 10.1109/IAC.2018.8780456

174 D. Aerlangga, R. M. Arsy, G. Sunardy, and T. Prasandy, “User Experience Analysis Using Usability Testing on Library and Knowledge Center BINUS University with SmartPLS,” in *2022 7th International Conference on Informatics and Computing, ICIC 2022*, IEEE Inc., 2022. doi: 10.1109/ICIC56845.2022.10006983

175 A. Y. Karoma, T. J. Ichsan, M. A. Dewi, and S. G. Rabiha, “User Experience Analysis on JIBAS Computer Based Exam Application Using Usability Testing Method,” in *2023 IEEE 9th International Conference on Computing, Engineering and Design, ICCED 2023*, Institute of Electrical and Electronics Engineers Inc., 2023. doi: 10.1109/ICCED60214.2023.10425732

176 J. Lazar, J. H. Feng, and H. Hochheiser, “Surveys,” in *Research Methods in Human Computer Interaction*, 2017, pp. 105–133. doi: 10.1016/B978-0-12-805390-4.00005-4

177 H. Taherdoost, “What are Different Research Approaches? Comprehensive Review of Qualitative, Quantitative, and Mixed Method Research, Their Applications, Types, and Limitations,” *Journal of Management Science & Engineering Research*, vol. 5, no. 1, pp. 53–63, Apr. 2022, doi: 10.30564/jmser.v5i1.4538

178 A. J. Kleinheksel, N. Rockich-Winston, H. Tawfik, and T. R. Wyatt, “Demystifying content analysis,” *Am J Pharm Educ*, vol. 84, no. 1, pp. 127–137, Jan. 2020. DOI: 10.5688/ajpe7113

179 M. Vaismoradi, J. Jones, H. Turunen, and S. Snelgrove, “Theme development in qualitative content analysis and thematic analysis,” *J Nurs Educ Pract*, vol. 6, no. 5, Jan. 2016, doi: 10.5430/jnep.v6n5p100

180 L. McKenna, I. Brooks, and R. Vanderheide, “Graduate entry nurses’ initial perspectives on nursing: Content analysis of open-ended survey questions,” *Nurse Educ Today*, vol. 49, pp. 22–26, Feb. 2017, doi: 10.1016/j.nedt.2016.11.004

181 S. Ulfa, R. Bringula, C. Kurniawan, and M. Fadhli, “Student Feedback on Online Learning by Using Sentiment Analysis: A Literature Review,” in *Proceedings - 2020 6th International Conference on Education and Technology, ICET 2020*, IEEE Inc., Oct. 2020, pp. 53–58. doi: 10.1109/ICET51153.2020.9276578

182 M. Wankhade, A. C. S. Rao, and C. Kulkarni, “A survey on sentiment analysis methods, applications, and challenges,” *Artif Intell Rev*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022, doi: 10.1007/s10462-022-10144-1

183 W. Medhat, A. Hassan, and H. Korashy, “Sentiment analysis algorithms and applications: A survey,” *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011

184 K. Chitra, T. Madhumitha, D. Sivabalaselvamani, P. Keerthana, M. Pyingkodi, and S. G. Dharani, “Sentiment Analysis on Online Education During Covid Pandemic,” in *4th International Conference on Inventive Research in Computing Applications, ICIRCA 2022 - Proceedings*, IEEE Inc., 2022, pp. 1043–1048. doi: 10.1109/ICIRCA54612.2022.9985634

185 R. D. Tan *et al.*, “LMS Content Evaluation System with Sentiment Analysis Using Lexicon-Based Approach,” in *2022 10th International Conference on Information and Education Technology, ICIET 2022*, IEEE Inc., 2022, pp. 93–98. doi: 10.1109/ICIET55102.2022.9778976

186 N. Ismail, G. Kinchin, and J.-A. Edwards, “Pilot Study, Does It Really Matter? Learning Lessons from Conducting a Pilot Study for a Qualitative PhD Thesis,” *Int J Soc Sci Res*, vol. 6, no. 1, p. 1, Nov. 2017, doi: 10.5296/ijssr.v6i1.11720

187 E. Rodero, “Intonation and emotion: Influence of pitch levels and contour type on creating emotions,” *Journal of Voice*, vol. 25, no. 1, Jan. 2011, doi: 10.1016/j.jvoice.2010.02.002.

188 Gudrun Klasmeyer and Walter F Sendlmeier, *voice and emotional states*. Singular Thomson Learning, 2000. Available at: https://www.researchgate.net/publication/301852949_Voice_and_Emotional_States

189 R. Jürgens, A. Grass, M. Drolet, and J. Fischer, “Effect of Acting Experience on Emotion Expression and Recognition in Voice: Non-Actors Provide Better Stimuli than Expected,” *J Nonverbal Behav*, vol. 39, no. 3, pp. 195–214, Sep. 2015, doi: 10.1007/s10919-015-0209-

190 Python Software Foundation, “About Python,” *Python.org*. [Online]. Available: https://www.python.org/about/.

191 N. Ansari, S. Awari, S. H. Movva, A. Parthasarathy, and S. Poriwade, “GesSpy: ML Driven Real Time Sign Language Detection,” Institute of Electrical and Electronics Engineers (IEEE), Sep. 2024, pp. 912–917. doi: 10.1109/icipcn63822.2024.00157

192 Adithya Sanjeev Byalpi and Anush, “Alexa based Real-Time Attendance System,” in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)*, IEEE, 2018. doi: 10.1109/CESYS.2018.8724006

193 Pandas, “Pandas: Python Data Analysis Library,” *Pandas*. [Online]. Available: https://pandas.pydata.org/.

194 Librosa, “Librosa: audio and music processing in Python,” *Librosa*. [Online]. Available: https://librosa.org/doc/main/.

195 Scikit-learn, “Scikit-learn: machine learning in Python,” *Scikit-learn*. [Online]. Available: https://scikit-learn.org/stable/.

196 TensorFlow, “TensorFlow,” *TensorFlow*. [Online]. Available: https://www.tensorflow.org/

197 S. Dinesh Kumar, D. Rajasekar, and S. Sharan Prasad, “SENTIMENT ANALYSIS USING RECURRENT NEURAL NETWORKS,” in *National Conference on Recent Advancements in Communication, Electronics and Signal Processing-RACES’20*, 2020. [Online]. Available: https://www.irjet.net/archives/V7/i8/Velammal/NCRACES-41.pdf

198 H. Taherdoost, “What are Different Research Approaches? Comprehensive Review of Qualitative, Quantitative, and Mixed Method Research, Their Applications, Types, and Limitations,” *Journal of Management Science & Engineering Research*, vol. 5, no. 1, pp. 53–63, Apr. 2022, doi: 10.30564/jmser.v5i1.4538

199 G. Marshall and L. Jonker, “An introduction to descriptive statistics: A review and practical guide,” *Radiography*, vol. 16, no. 4, 2010. doi: 10.1016/j.radi.2010.01.001

200 M. J. Fisher and A. P. Marshall, “Understanding descriptive statistics,” *Australian Critical Care*, vol. 22, no. 2, pp. 93–97, May 2009, doi: 10.1016/j.aucc.2008.11.003

201 S. Jaggi, “Descriptive Statistics and Exploratory Data Aanalysis.”

202 V. Braun and V. Clarke, “Using thematic analysis in psychology,” *Qual Res Psychol*, vol. 3, no. 2, pp. 77–101, 2006, doi: 10.1191/1478088706qp063oa.

203 M. E. Kiger and L. Varpio, “Thematic analysis of qualitative data: AMEE Guide No. 131,” *Med Teach*, vol. 42, no. 8, pp. 846–854, Aug. 2020, doi: 10.1080/0142159X.2020.1755030

204 R. Mohammed, J. Rawashdeh, and M. Abdullah, “Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results,” in *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, Institute of Electrical and Electronics Engineers Inc., Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556

205 Hammoumi Oussama, Benmarrakchi Fatimaezzahra, Ouherrou Nihal, Kafi Jamal, and Hore Ali El, “Emotion Recognition in E-learning Systems,” in *2018 6th International Conference on Multimedia Computing and Systems : 10-12 May 2018, Rabat, Morocco*, IEEE, 2018, p. 87. doi: 10.1109/ICMCS.2018.8525872

206 B. L. Fredrickson, “The Role of Positive Emotions in Positive Psychology The Broaden-and-Build Theory of Positive Emotions,” *American Psychologist,* vol. 56, no.3, pp. 218-226, 2001. doi: 10.1037//0003-066x.56.3.218

207 Y. Niu, “A breakthrough in Speech emotion recognition using Deep Retinal Convolution Neural Networks,” *arXiv:1707.09917*, 2017, doi: 10.48550/arXiv.1707.09917

208 T. Anvarjon, Mustaqeem, and S. Kwon, “Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features,” *Sensors (Switzerland)*, vol. 20, no. 18, pp. 1–16, Sep. 2020, doi: 10.3390/s20185212

209 E. V. Svinndal, C. Jensen, and M. B. Rise, “Working life trajectories with hearing impairment,” *Disabil Rehabil*, vol. 42, no. 2, pp. 190–200, Jan. 2020, doi: 10.1080/09638288.2018.1495273

210 World Health Organization, “World report on hearing,” *Ear Science Institute Australia*. [Online]. Available: https://www.earscience.org.au/wp-content/uploads/World-Report-on-Hearing.pdf.

211 A. F. Rahmawati, T. Wahyuningrum, A. C. Wardhana, A. Septiari, and L. Afuan, “User Experience Evaluation Using Integration of Remote Usability Testing and Usability Evaluation Questionnaire Method,” in *Proceedings - 2022 IEEE International Conference on Cybernetics and Computational Intelligence, CyberneticsCom 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 40–45. doi: 10.1109/CyberneticsCom55287.2022.9865664

212 B. O. Olusanya, A. C. Davis, and H. J. Hoffman, “Hearing loss grades and the international classification of functioning, disability and health,” Oct. 01, 2019, World Health Organization. doi: 10.2471/BLT.19.230367

213 P. Ghosh, T. S. Chingtham, M. K. Ghose, and R. Dutta, “A Review of Modern HCI Challenges for Betterment of Differently Abled People,” *International Journal of Computational Intelligence Research*, vol. 9, no. 2, pp. 79–88, Dec. 2013, doi: 10.37622/ijcir/9.2.2013.79-88

214 A. Beha and H. Hasanbegović, “LABOR CAPACITY OF DEAF WORKERS ON THE WORKPLACE: QUALITATIVE ANALYSIS OF THE ATTITUDES OF DEAF WORKERS AND THEIR CO-WORKERS WITHOUT HEARING IMPAIRMENT,” *Human Research in Rehabilitation*, vol. 9, no. 2, pp. 40–47, 2019, doi: 10.21554/hrr.091906

215 S. E. Kramer, “Hearing impairment, work, and vocational enablement,” in *International Journal of Audiology*, Nov. 2008. doi: 10.1080/14992020802310887

216 E. V. Svinndal, C. Jensen, and M. B. Rise, “Employees with hearing impairment. A qualitative study exploring managers’ experiences,” *Disabil Rehabil*, vol. 42, no. 13, pp. 1855–1862, Jun. 2020, doi: 10.1080/09638288.2018.1541101.

217 Abascal, J., & Nicolle, C. (2005). Moving towards inclusive design guidelines for socially and ethically aware HCI. Interacting with Computers, 17(5), 484–505. https://doi.org/10.1016/j.intcom.2005.03.002.

Appendices

Appendix 1 Table 2.1

Table 2.1 Summary of literature review of the machine learning techniques for SER.

SNO	AUTHOR/ REFERENCE	DATASET/s	EMOTIONS CONSIDERED	FEATURES USED	CLASSIFEIR	CONTRIBUTION	LIMITATIONS
1.	Rahul B.et al.[81]	Berlin emotional DB (Emo-DB)	Angry, Happy, sad, surprised, neutral, and fearful	Spectral components of MFCC, speech wavelet features and pitch	GMM and K-NN	The two classifiers K-NN, GMM have been implemented to identify the six emotions. GMM has worked well for angry emotion with 92% accuracy, and K-NN technique detects the 'happy' emotion with 90% rate.	The two classifiers have been independently implemented, fusion of the two techniques has not been considered for improving emotion recognition
2.	M.Swain et. al. [89]	Multilingual database	Happy disgust, sad, fear, neutral and surprise	Log power, Mel frequency cepstral coefficients (MFCC), Delta MFCC, Double delta MFCC, log frequency power coefficients, and linear predictive cepstral coefficients	HMM and SVM	MFCC with SVM is more efficient. 78.81% accuracy using HMM and 82.41% with SVM was obtained for speaker independent signal for 'Sambalpuri' language	Hierarchical classification of emotions has not been considered for overall enhancement of the system performance
3.	Sagar K. et al. [88]	Self-created database	Angry, Happy, sad, and fearful	pitch, zero-crossing rate (ZCR), MFCC, and energy were extracted.	Naïve Bayes Classifier	MFCC, pitch and energy features are used to train the classifier. 76% for sad, 77% neutral, 78% for happy, and 81% for angry have been achieved.	Features like voice quality and prosodic have not been considered, which can help in enhancing the performance.
4.	F. Noroozi et al. [87]	SAVEE	Happy, sad, surprise, fear, and disgust.	Prosodic features	Random forest and Decision Tree	A Random Forest with decision tree approach is adopted. 78% recognition rate.	Improves accuracy but time-consuming method
5.	Koduru et al. [56]	RAVDESS	Anger, Happiness, Sadness, and Neutral	MFCC, DWT, pitch, energy and ZCR	SVM, Decision tree, and LDA.	Results that Decision tree gives 85%, SVM gives 70% and LDA gives 65% accuracy for RAVDESS	The proposed model gives good accuracy and efficiency but unfortunately tested with only one dataset.
6	Chen et. Al [82]	CASIA and EMODB	angry, fear, happy, neutral, sad, and surprise	non-personalized and personalized features	Random forest	Compared to backpropagation neural networks and random forests, TLFMRF demonstrated a	Intelligent optimization can be used so that robots are able to sense human emotion, and

						1.39%–7.64% and 4.06%–4.30% higher recognition rate, respectively.	people can communicate with robots more smoothly.
7	Langari et al. [83]	EMO-DB, SAVEE, PDREC	Sad, Neutral, Happy, Fear, Disgust, Bored, Angry.	Time–Frequency features	SVM	The conducted experiments have given an accuracy of 97.57%, 80%, 91.46% for the datasets EMO-DB, SAVEE and PDREC respectively	A potential integration of deep features with the extracted time features, alongside the adoption of deep learning classifiers for constructing a speaker-independent SER system would be ideal.
8.	(Bilal, 2020) [84]	RAVDESS, EMO-DB, and IEMOCAP	Angry, Happy, Neutral and Sad	Acoustic and deep features	SVM	RAVDESS, EMO-DB, and IEMOCAP—yielded accuracy rates of 79.41%, 90.21%, and 85.37%, respectively.	New techniques to identify optimal feature should be considered.
9.	Yildirim et al. [86]	EMO-DB, IEMOCAP	Angry, Boredom, Disgust, Fear, Happy, Sad and Neutral.	INTER-SPEECH 2010 paralinguistic	SVM, KNN, Tree Bagger	With EMO-DB dataset, 87.66% and 87.20 recognition rate were attained using modified feature selection with Cuckoo Search and NSGA-II respectively. On the other hand, an accuracy of 69.30% and 68.32% was obtained using SVM classifier with IEMOCAP dataset	More datasets and better intelligent algorithms can be used for its applicability to real-world applications.
10.	Ancilin & Milton [85]	Berlin, RAVDESS, SAVEE, EMOVO, eNTERFACE and Urdu databases	Anger, Happy, Sad, Neutral, Disgust, Surprise, Fear	MFCC, Log Frequency Power Coefficient (LFPC), and Linear Prediction Cepstral Coefficient (LPCC)	multiclass SVM	Accuracies of 81.50% for Berlin, 64.31% for RAVDESS, 75.63% for SAVEE, 73.30% for EMOVO, 56.41% for eNTERFACE, and 95.25% for Urdu were achieved.	Feature selection, feature fusion and multiple classification schemes not adapted.

Appendix 2 Table 2.2

Table 2. 2 Summary of literature review of the DL techniques for SER.

SNO	AUTHOR/ REFERENCE	DATASET/s	EMOTIONS CONSIDERED	FEATURES USED	CLASSIFEIR	CONTRIBUTION	LIMITATIONS
1.	Mohammad Mehedi et al. (2019) [113]	DEAP	Angry, happy and neutral	MFCC	DBN and FGSVM	Fusion of classifiers with DBN and FGSVM resulted in 89.53% accuracy	Lack of an ensemble of classifiers to increase generalizability and robustness of SER.
2.	Linhui Sun et al. [99]	Chinese Academy Sciences Emotional corpus	Anger, happy, sad, fear, and surprise	Deep bottleneck features	SVM, DNN-SVM, DNN-decision tree SVM	A multi-layer classifier is used which accuracy achieved of 75.83%	Unable to differentiate the emotions fear and sad.
3.	H.S. Kumbhar et al. (2019) [55]	RAVDESS	Happy, sad, calm, angry, surprise, fearful, disgust	MFCC	LSTM	Uses MFCC and LSTM to obtain 80.81%	There is a loss of 67.21%, which needs to be taken care of and a combination of features should be used.
4.	Zhao et al., (2019) [105]	IEMOCAP, EmoDB	Angry, Excited, Frustrated, Happy, Neutral, Sad	Raw speech and log-mel spectrograms, respectively	CNN and LSTM	The 2D CNN-LSTM achieved 95.33% and 95.89% accuracy of speaker-dependent and speaker-independent experiments respectively, on Berlin EmoDB, surpassing traditional approaches like Deep Belief Networks (DBNs) and CNNs, which achieved 91.6% and 92.9%, respectively. Similarly, on the IEMOCAP dataset, it achieved 89.16% and 52.14% accuracy, significantly outperforming the DBN and CNN, which achieved 73.78% and 40.02%, respectively.	Challenging in explaining how the networks recognize the emotions.
5	S. Latif et. Al. (2020) [110]	IEMO CAP and MSP-IMPROV	Angry, happy, neutral, and sad	Raw speech	multi-layer CNN stacked on an LSTM	Proposed model achieved 60.23% by directly using raw speech.	Using raw speech signals as inputs direct have resulted in low accuracy.

6.	Shahin et al., [115]	Emirati-accented speech and SUSAS	Neutrality, happiness, sadness, disgust, anger, and fear'	MFCC	A hybrid of a cascaded Gaussian mixture model and deep neural network (GMM-DNN)	The sequential GMM-DNN classier performance accuracy shown to be 83.97%, which is higher than support vector machines (SVMs) and multilayer perceptron (MLP) classifiers, which show accuracy of 80.33% and 69.78% respectively.	The hybrid GMM-DNN classifier’s performance was validated using two different emotional databases under both normal and noisy speaking conditions. The dominant signal mask generated by the hybrid classifier enhances system performance, particularly in noisy environments.
7.	Mustaqeem et al. (2020) [91]	IEMOCAP, EmoDB and RAVDESS	Happy, sad, angry and neutral	Discriminative and salient features from spectrograms of speech signal	RBFN, CNN and BILSTM	77% accuracy has been achieved using RAVDEES and 72.25% with IEMOCAP and EmoDB.	computational complex.
8.	TM Wani et al. (2020) [100]	SAVEE	Angry, happy, and neutral	Spectrograms	CNN and DSCNN	87.7% and 79.4% accuracy has been obtained resp.	Model has not been tested with more datasets to check the effectiveness.
9.	D. Chen et. Al. (2020) [108]	IEMOCAP	Happy, neutral, angry, and sad	MFCC features, and Mel-spectrograms	DSLSTM with two mel-spectrograms simultaneously	Weighted accuracy of 72.7% and an unweighted accuracy of 73.3%—a 6% improvement over current state-of-the-art unimodal models	It should be tested on more databases
10.	Mustaqeem et. Al. (2020) [92]	IEMOCAP and RAVDESS	Happy, neutral, angry, and sad	Raw speech data	CNN with ConvLSTM	It gave 75% accuracy on IEMOCAP and 80% on RAVDESS.	For applicability to real time applications, the model needs to be validated with other datasets as well and verify the accuracy.
11.	Yao et al., 2020 [68]	IEMOCAP	angry, happy, neutral, and sad	Frame-level low-level descriptors (LLDs), segment level Mel-spectrograms (MS), and utterance-level outputs of high-level statistical	Fusion of three classifiers, namely DNN, CNN and RNN are used.	Proposed fusion model gave weighted accuracy 57.1% and unweighted accuracy 58.3%, which were significantly higher than for each individual classifier.	Feature-level fusion framework should be developed by jointly optimizing the sub-classifiers.

				functions (HSFs)			
12.	Z.Huijuan et al. (2020) [107]	IEMOCAP	Happy, angry, frustrated, and excited	3D log-Mel spectrum	CNN blocks with RNN attention module	A novel top-down hierarchal approach has been proposed. F1mscore of 0.4673	Lacking a cross-corpus emotion recognition.
13.	X. Wang et. Al. (2020) [109]	IEMOCAP	Happy, neutral, angry, and sad	End-to-end architecture	Stacked transformer layers (STLs)	20% improvement compared to the previous model	Model is not tested on range on public datasets
14.	Mustaqeem et. Al. (2020) [93]	IEMOCAP and EMO-DB	Angry, happy, neutral, and sad	Residual blocks with a skip connection (RBSC) module, in order to find a correlation, the emotional cues, and the sequence learning (Seq_L) module, to learn the long term contextual dependencies in the input features	lightweight dilated CNN architecture that implements the multi-learning trick (MLT) approach.	Evaluated on IEMOCAP and EMO-DB datasets and obtained a high recognition accuracy, which were 73% and 90%.	The suggested model does not utilize various real-time amenities, for example, speaker recognition and identification.
15.	Mustaqeem et al., 2021 [35]	EMO-DB, SAVEE, and RAVDESS	Angry, sadness, happiness, neutral, disgust, fear, and boredom	Spectrum, spectrogram and high-level discriminative features	Two-stream CNN model	High accuracy of 95%, 82%, and 85% respectively on the EMO-DB, SAVEE, and RAVDESS speech datasets	This model needs to be evaluated with huge and natural datasets for its applicability in real world applications.
16.	Atila, O et al., 2021 [94]	SAVEE, RAVDESS and RML	Angry, disgusted, fearful, happy, sad, and surprised	Spectrogram, MFCC and cochleagram and a signal processing method namely windowed fractal dimension	A novel attention guided 3D convolutional neural networks (CNN)-long short-term memory (LSTM) model has been proposed which gives an improved accuracy over the published works for selected datasets	Average accuracy of 87.5%, 93.2%, 96.18% and 93.71% was obtained for SAVEE, RML and RAVDESS respectively.	Newer datasets must be used to evaluate the model.
17.	Zhang et al., 2021 [70]	AFEW5.0 and BAUM-1	Anger, joy, sadness, fear, disgust, surprise	Feature representations from 1D, 2D, and 3D CNN networks are fused for the final classification, surpassing	Raw waveform modeling, time-frequency Mel-spectrogram modeling, and temporal-	This method gives a performance accuracy of 35.77%, and 44.06% on the AFEW5.0 database. BAUM-1 respectively.	Larger natural datasets have not been used to evaluate the model for its usage in real world applications.

				state-of -the arts performances.	spatial dynamic modeling.		
18.	Chen et al., 2021 [111]	IEMOCAP	Happy, Sad, Angry, Natural.	Mel-spectrogram, MFCC and deltas.	Dual attention-BLSTM model	Achieves an average recognition accuracy of 70.29% in unweighted accuracy (UA), showcasing performance enhancements of 2.89% compared to the leading baseline method.	The model needs to be tested with variety of datasets.
19.	Zhao et al. 2021 [112]	IEMOCAP, FAU-AEC	Neutral, Happy, Sad, Angry. Angry, Emphatic, Neutral, Positive, Rest	3D spectrograms	2D CNN	Experimental results indicate achieving a weighted accuracy (WA) of 73.1% and an unweighted accuracy (UA) of 66.3% on IEMOCAP, as well as a UA of 41.1% on the FAU-AEC dataset	The proposed method must be further explored to check it suitability to applications related to speech and also its applicability in real world applications
20.	Ntalampiras 2021 [95]	Emo-DB	Angry, Disgust, Fear, Happy, Natural	Log-mel spectrograms	Siamese Neural Network (SNN)	Good recognition rates were seen only for some classes like - anger, sad-ness and neutrality.	The learning analogies must be language independent for suitability to real-world scenarios
21.	Pepino et al. (2021) [96]	IEMOCAP, RAVDESS	Neutral, Happy, Sad, Angry. Anger, Sad, Calm, Happy, Neutral, Fear, Disgust, Surprise.	Features extracted from pre-trained wav2vec 2.0 models	DNN	Evaluations on the two datasets, IEMOCAP and RAVDESS demonstrated an accuracy of 66.3%, 77.5% respectively. It showed superior performance as compared to results in literature,	It would be appropriate to include more datasets for testing
22.	Li et al. (2021) [97]	IEMOCAP, EMO-DB, ENTERFACE05, SAVEE	Happy, Neutral, Angry, Sad.	Spectrogram	CNN	Accuracy of 80.47%, 83.30%, 75.80%, 56.50% has been shown on IEMOCAP, EMO-DB, ENTERFACE05, SAVEE respectively	Handling variable-length speech based on the proposed method must be looked at.
23.	Amjad et al., 2022 [71]	SAVEE, IEMOCAP, and BAUM-1s	Joy, surprise, anger, disgust, sadness, fear	Raw speech data and augmented mel spectrograms	Deep convolutional neural networks (DCNNs) 1D (Model A) and 2D (Model B) with two layers of long-short-term memory (LSTM).	CNN 1D obtained identification accuracies of 97.19%, 94.09%, and 53.98% on SAVEE, IEMOCAP, and BAUM-1s respectively and CNN 2D achieved identification accuracy of 96.85%, 88.80%,	There is a need for an approach to learn and compare more specific features and for further improvements in terms of accuracy.

						and 48.67% on SAVEE, IEMOCAP, and the BAUM-1s	
24.	Zhang et al., 2022 [72]	AFEW5.0 and BAUM-1	Anger, joy, sadness, fear, disgust, surprise	Deep segment-level features and spectrograms.	Results of multiscale CNN and LSTM are fused for the final emotion classification.	The approach achieved an accuracy of 40.73% on AFEW5.0 and 50.22% on BAUM-1.	Cross-corpus testing must be employed.
25.	Wang et al. 2022 [101]	IEMOCAP, SAVEE, CHEAVD, CASIA.	Happy, Angry, Sad, Neutral. Anger, Disgust, Fear, Happy, Neutral, Sad, Surprise. Anger, Happy, Sad, Worry, Anxiety, Surprise, Disgust, Neutral. Angry, Happy, Sad, Surprise, Neutral	Handcraft + DNN features	Local Attention + RNN	The proposed model gave an accuracy of 72.3%, 81.8%, 55.7%, 63.3% on IEMOCAP, SAVEE, CHEAVD, CASIA datasets respectively.	More multi-feature algorithms can be considered as in this model the fusion worked well. Languages with different dialects and accents must also be taken into consideration.
26.	Zou et al. 2022 [104]	IEMOCAP	Happy, Angry, Sad, Neutral.	MFCC, Spectrograms, Wav2vec features	Co-attention mechanism	Model with dataset IEMOCAP gives an accuracy of 71.05%	More datasets with different languages can be considered to improve the performance of the model further.
27.	Mansouri et al. 2022 [102]	EMO-DB, IEMOCAP, eNTERFACE05	Neutral, Sad, Happy, Anger. Happy, Angry, Sad, Neutral. Anger, Disgust, Anxiety/Fear, Happy, Sad, Bored, Surprise, Neutral.	Spectrograms	DCNN	Gives an accuracy of 95.14%, 74.23%, 89.46% for datasets EMO-DB, IEMOCAP, eNTERFACE05 respectively	An optimization technique with hyper-parameters will result in better performance of the model
28.	Aftab et al. 2022 [103]	IEMOCAP, EMO-DB	Happy, Sad, Angry, Neutral. Happy, Sad, Angry, Neutral, Fear, Disgust, Bored.	MFCC	CNN	Model gives an accuracy of 79.87% for IEMOCAP and 94.21% for EMODB	The proposed method must be tested with many other available datasets for its usability to develop real-world applications.
29.	Hasan et al., [114]	SUST Bengali Emotional Speech Corpus (SUBESCO)	Six primary emotions, anger, disgust, fear, happiness, sadness, and surprise	MFCC, Chroma, and Mel-Spectrogram	Bidirectional Long Short Term Memory (Bi-LSTM)	The hybrid model (Bi-LSTM) achieved an accuracy of 83.33%, outperforming the SVM classifier (81.33%) and the MLP classifier (80.38%).	The model can be trained with a larger trained dataset so that the model can work on the larger range of the audio file's loudness, pitch,

							frequency, audibility, and so on.
30.	Rayhan Ahmed et al. 2023 [73]	TESS, EMO-DB, RAVDESS, SAVEE and CREMAD	happiness, sadness, fearful, surprise, anger, surprise, boredom, neutral etc.,	ZCR, Chromagram, MFCC, RMS, and Log mel spectrogram	The first architecture uses 1D CNN followed by Fully Connected Networks (FCN). In the other two architectures, LSTM-FCN and GRU-FCN layers follow the CNN layer respectively.	They achieve a state-of-the-art (SOTA) weighted average accuracy of 99.46% for TESS, 95.42% for EMO-DB, 95.62% for RAVDESS, 93.22% for SAVEE, and 90.47% for CREMAD datasets	This work could reduce training time for individual models for the ensemble prediction. This could be achieved by selecting optimal features together with attention mechanism
31.	Alluhaidan et al., 2023 [90]	EMO-DB, SAVEE, and RAVDESS	anger, boredom, calm, disgust, fear, happy, neutral, sad, surprise.	MFCCs and time-domain features (MFCCT)	1D CNN	Model demonstrated superior accuracy over baseline methods, achieving 96.6% on EMO-DB, 92.6% on SAVEE, and 91.4% on RAVDESS. This method notably improved accuracy by 10% for EMO-DB, 26% for SAVEE, and 21% for RAVDESS.	Additional datasets should be used. The integration of RNNs with optimal acoustic features, as they offer improved accuracy by capturing high-level acoustic must be considered.
32.	Mao et al., 2023 [98]	EMO-DB, SAVEE, and RAVDESS	anger, boredom, calm, disgust, fear, happy, neutral, sad, surprise.	MFCC from the input Audios followed by multi-branch deep feature	CNN	Extensive experiments demonstrate SCAR-NET's robustness and effectiveness, achieving accuracies of 96.45%, 83.13%, and 89.93% on the EMO-DB, SAVEE, and RAVDESS datasets, respectively	SCAR-NET can be optimized further to enhance its performance and reduce parameter count, as well as assess its effectiveness on a more realistic corpus.
33.	Gao et al [106]	TESS	Neural, Happy, Sad, Angry, Fear, Pleasant Surprise, and Disgust.	Mel Spectrograms	CNN-LSTM	CNN-LSTM neural network model to the Zynq FPGA board from Xilinx after simulation experiments using the TESS dataset achieved a recognition accuracy of 97.86%, which verified the feasibility of the system.	Speech emotions in real time can be classified better by adding a speech acquisition module and a feature extraction module.

Appendix 3 Table 2.3

Table 2.3: Summary of Related work on HCI Solutions for Late-deafened Educators

SNO	AUTHOR/ REFERENCE	CONTRIBUTION	RELEVANCE TO HEARING-IMPAIRED EDUCATORS	LIMITATIONS
Challenges Faced by Hearing-Impaired Educators				
1.	Tidwell [29]	Age-related hearing loss; adaptation strategies	Personal insights from a senior educator; highlights lived challenges	No discussion of technology use in teaching
2.	Smith & Andrews [30]	Institutional support for DHH faculty	Suggests assistive apps, tech, and accommodations	Lacks specific focus on online teaching environment
3.	Deshmukh et al. [124]	Pandemic challenges for hearing-impaired people	Tech-supported inclusion strategies	Minimal focus on education/online teaching
Assistive Technologies and HCI Approaches				
4.	Rivas-Costa et al. [125]	E-inclusion via adaptive platforms	Accessibility tool tested with disabled users	No direct link to late-deafened educators or online teaching
5.	Farhan & Jamil [126]	Inclusive e-learning interface design	Prototypes tested with hearing-impaired students	Teacher-specific needs not deeply explored
6.	Seita [127]	ASR systems usability for DHH individuals	Relevant for improving online communication	No emotion detection or specific focus on educators
Educational and Communication Support Technologies				
7.	Baglama et al. [128]	Language tech and mobile apps for communication	Proposes app development for diverse skill levels	No mention of teaching challenges for educators
8.	Amandeep & Williamjeet [129]	Assistive devices (ALD, AAC, etc.)	Describes tech to aid communication and teaching	Doesn't address online teaching platform integration
9.	Kim et al. [130]	Usability of mobile apps for communication	Inclusive design based on user feedback	Focused on general social tools, not educator needs
Emerging Technologies and Future Directions				
10.	Flores Ramones et al. [131]	Review of haptic assistive devices	Tactile tech supports sensory substitution	Not suited for speech-based teaching contexts
11.	Serafin et al. [132]	VR in assistive tech training	Discusses VR potential in testing/training	No application for emotional feedback or online education
12.	Palanisamy et al. [133]	ASL recognition using deep learning + graph theory	Enhances communication through gesture recognition	No solutions proposed for online teaching

Appendix 4 Ethics Approval Letter

Science & Engineering Research Ethics Committee

University of Nottingham
Malaysia

43500Jalan Broga
Semenyih

Selangor Darul Ehsan
Malaysia

+603 8924 8000
nottingham.edu.my

20 October 2022

Dear Aparna Vyakaranam,

Study title: Speech emotion recognition system (SER) for late deafened educators in online education SEREC reference: AV061022

Thank you for submitting the above study for review by the Science and Engineering Research Ethics Committee (SEREC).

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form and supporting documentation. As your proposal has been found to comply with the requirements for research with human participants as set down by the University of Nottingham, the Committee is able to grant you approval to commence the study.

Conditions of the favourable opinion

If there are any changes or developments in the methods, treatment of data, or debriefing of participants, then you are obliged to seek further ethical approval for these changes using the *Amendment Form*. Furthermore, you are obligated to notify the Committee within 7 calendar days of any adverse effects on participants and any other unforeseen events or near misses that might affect the continued ethical acceptability of the project.

We would remind all researchers of their ethical responsibilities under the University of Nottingham Code of Research Conduct and Research Ethics. If you have any concerns whatsoever during the conduct of your research, then you should consult this Code and/or SEREC.

You should also be aware that supervisors are responsible for staff, student and participant safety during projects. Relevant information can be found on the Safety Office pages of the University website:

www.nottingham.ac.uk/safety

University of
Nottingham in
Malaysia Sdn Bhd
(473520-K)

Responsibility for compliance with the University Data Protection Policy and Guidance also lies with the project supervisor. Ethics Committee approval does not alter, replace or remove those responsibilities, nor does it certify that they have been met.

I would also like to remind all researchers of their responsibilities to provide feedback to participants and participant organisations whenever appropriate, and to publish research for which ethical approval is given in appropriate academic and professional journals.

Yours sincerely,

ShiHui

Dr Cheng Shi Hui
Chair
Science and Engineering Research Ethics Committee

Appendix 5 Questionnaire

To ___________________

This is Aparna Vyakaranam from the School of Computer Science, Nottingham University, Malaysia campus. I am doing a user requirement analysis for my research work. All data obtained will be kept confidential and are accessed by the primary researcher only. Participation, completion and return of the survey are voluntary. You may **withdraw** at any time or decline to participate. Thank you for your time.

MY RESEARCH IDEA

In an online class, student responses help the educator know of the students understanding, learning and involvement in the on-going class. These responses carry emotions. Verbal responses are heard but gauging the emotion through these responses is not conclusive. For educators who are late deafened, even hearing the student responses is difficult. **My research focuses on developing a speech emotion recognition (SER) system producing maximum accuracy for benefit of all especially the late deafened educators**. Such an applications would help in extracting and displaying emotions from students in online education.

In view of this, I would like to seek your feedback through this interview which has the below questions.

We seek just 20 - 30 minutes of your time

The findings from this interview will help to develop a SER system that will be relevant to users. Your contribution would be very helpful in this matter, and we are happy to share these findings with you also.

I am **not looking for any ‘technical’ or ‘Personal’ information, but just your views** / opinions on issues related usage and please rest assured that all data will be kept confidential, and results will be used only in aggregate form.

If you have any concerns or require any clarifications regarding this questionnaire, please feel free to connect either

Aparna @hcxav1@nottingham.edu.my

Or

Dr.Tomas Maul @Tomas.Maul@nottingham.edu.my

Dr.Bavani @Bavani.R@nottingham.edu.my

This questionnaire consists of the following –

- Section A – Questions on respondent’s demographics / Profile
- Section B – Questions on respondents hearing condition (to understand if they have any form of deafness that has developed at a later stage in life)
- Section C – Questions to understand how respondents use online systems.
- Section D – Questions related to feedback and emotion recognition in online systems.

Seeking your help to provide the required information for the questions outlined as below

Section A. Respondent’s Demographics / Profile

Q.1 Your gender: -

Male	Female
○	○

Q.2 Please indicate the age group you belong to -

< 30	31 - 40	41 - 50	51– 60	> 61
○	○	○	○	○

Q.3 Please indicate your teaching experience: -

1-4 years	5-10 years	11-15 years	16-20 years	>20 years
○	○	○	○	○

Q.4 Please indicate which is the area of your teaching?

Social Science / Business	STEM	Others
○	○	○

Q.5 What level students do you teach?

Primary School / Middle school	High School / Pre – uni courses (A levels/STPM etc.)	Higher Education or University	TVET	Others (Corporate training etc.)
○	○	○	○	○

Q.6 Have you received any form of formal training in education?

Pedagogical training	Online tool training	Pedagogical and online tool training	None	Others
○	○	○	○	○

Section B. – Respondent's Hearing condition.

Q.7 Describe your current hearing condition. Please select an option below.

Normal	Mild (may have difficulty hearing soft spoken people)	Moderate (without hearing aids they hear, but may not always understand speech)	Severe (without hearing aids, regular speech is inaudible)	Profound (without hearing aids, unable to hear even very loud sounds)
○	○	○	○	○

Q.8 When conducting a class/session (face-to-face) in the **classroom**, were you able to hear the student vocal feedback with your current hearing condition described in Q.7?

Yes	Yes, with help of an hearing aid	No
○	○	○

Q.9 If you answered 'No' in Q.8, please provide some details on the challenges.

Section C. Usage of online systems.

Q.10 How often have you conducted classes online in the last two (2) years?

Regularly (> 10 times in an year)	5 - 9 times in an year	< 5 times in an year	None
○	○	○	○

Q.11 Please indicate the different tools/platform you have used in online teaching (MS teams / Zoom / Skype etc.)

Q.12 Please indicate the different methods you have used to conduct your online classes (Like ppt / pdf slides, slido, video conferencing etc.)

Sections D. - Feedback and emotion recognition in online systems

Q.13 In a (face-to-face) **classroom**, educators initiate interaction with students by asking some questions. To what extent do you agree emotions carried in the student vocal feedback are helpful in understanding their engagement?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.14 Educators can adjust teaching strategies based on the student vocal response and emotions carried in them as described in Q.13?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.15 In an **online class**, real-time student feedback for understanding student engagement is more critical, as there is no physical interaction?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.16 In an **online class**, how did you take the feedback in real time?

□Chat box S□veys Voc□feedback Other □

Q.17 Were you able to hear the student's voice in vocal feedback?

Yes	Yes, with help of an hearing aid	No
○	○	○

Q.18 If you answered ‘No’ in Q.17, please provide some details on the challenges.

Q.19 In an online class, would it be beneficial if the underlying emotion (like happy or sad or neutral or bored and such) of the student is automatically detected from his/her vocal feedback and displayed using an image?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.20 Understanding the student emotion during online teaching is helpful in adjusting the teaching approach for the educator.

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.21 Please describe any other ways you gauge the emotions of students in online sessions.

We are very grateful for your time in filling up this questionnaire.

Appendix 6 Ethics Approval Letter

Science & Engineering Research Ethics Committee

University of Nottingham
Malaysia

43500Jalan Broga
Semenyih
Selangor Darul Ehsan
Malaysia

+603 8924 8000
nottingham.edu.my

13 May 2024

Dear Aparna Vyakaranam,

Study title: Speech emotion recognition (SER) for late deafened educators in online education SEREC reference: AV200124

Thank you for submitting the above study for review by the Science and Engineering Research Ethics Committee (SEREC).

Confirmation of ethical opinion

On behalf of the Committee, I am pleased to confirm a favourable ethical opinion for the above research on the basis described in the application form and supporting documentation. As your proposal has been found to comply with the requirements for research with human participants as set down by the University of Nottingham, the Committee is able to grant you approval to commence the study.

Conditions of the favourable opinion

If there are any changes or developments in the methods, treatment of data, or debriefing of participants, then you are obliged to seek further ethical approval for these changes using the *Amendment Form*. Furthermore, you are obligated to notify the Committee within 7 calendar days of any adverse effects on participants and any other unforeseen events or near misses that might affect the continued ethical acceptability of the project.

We would remind all researchers of their ethical responsibilities under the University of Nottingham Code of Research Conduct and Research Ethics. If you have any concerns whatsoever during the conduct of your research, then you should consult this Code and/or SEREC.

You should also be aware that supervisors are responsible for staff, student and participant safety during projects. Relevant information can be found on the Safety Office pages of the University website:

www.nottingham.ac.uk/safety

University of
Nottingham in

Malaysia Sdn Bhd
(473520-K)

Responsibility for compliance with the University Data Protection Policy and Guidance also lies with the project supervisor. Ethics Committee approval does not alter, replace or remove those responsibilities, nor does it certify that they have been met.

I would also like to remind all researchers of their responsibilities to provide feedback to participants and participant organisations whenever appropriate, and to publish research for which ethical approval is given in appropriate academic and professional journals.

Yours sincerely,

ShiHui

Dr Cheng Shi Hui
Chair Science and Engineering Research Ethics Committee

Appendix 7 Questionnaire

To ___________________

Good day. I am Aparna Vyakaranam from the School of Computer Science, University of Nottingham, Malaysia campus. I am conducting a usability testing for **speech emotion recognition (SER)** system which is part of my research work. All data obtained will be kept confidential and are accessed by the primary researcher only. Participation, completion and return of the survey are voluntary. You may **withdraw** at any time or decline to participate. Thank you for your time.

MY RESEARCH IDEA

I have developed a speech emotion recognition (SER) system which detects the emotion of the student as positive or negative from their speech signal during their verbal feedback in an online class. This detected emotion can benefit all educators, especially the late deafened in adjusting their teaching approach.

In light of this, I would like to seek your time to read the guidelines (refer to page 2) to use the system (website) and provide feedback through this **questionnaire**. Your responses will help me to understand your views on the accuracy and satisfaction relative to the developed SER system. The insights from this usability testing can help lead to any refinements for an enhanced user experience.

I seek about 30 minutes of your time.

The findings from this testing will help me understand the efficiency and the effectiveness of the developed SER system. Your contribution would be very helpful in this matter, and I will be happy to share these findings with you as well.

I am **not looking for any 'technical' or 'personal' information, but just your views** / opinions on issues related to the usage of the system. Please rest assured that all data will be kept confidential, and results will be used only in aggregate form.

If you have any concerns or require any clarifications regarding this questionnaire, please feel free to contact either

Aparna @hcxav1@nottingham.edu.my

Or

Dr.Tomas Maul @Tomas.Maul@nottingham.edu.my

Dr.Bavani @Bavani.R@nottingham.edu.my

This questionnaire consists of the following –

Section A – Questions on the respondent's demographic / profile

Section B – Questions on the respondent's hearing condition (to understand if they have any form of deafness that has developed at a later stage in life)

Section C – Questions to understand how the respondent uses online systems and the significance of feedback during online sessions.

Section D – Questions related to the effectiveness and efficiency (or evaluation) of the developed Speech Emotion Recognition system.

Section E – Questions related to feedback on the impact and improvement of the developed Speech Emotion Recognition system in Online education.

Seeking your help to provide the required information for the questions outlined as below.

Section A. Respondent's Demographics / Profile

Q.1 Please indicate your gender.

Male	Female
○	○

Q.2 Please indicate the age group you belong to.

< 30	31 - 40	42 - 50	51– 60	> 61

○	○	○	○	○

Q.3 Please indicate the length of your teaching / training experience.

1-4 years	5-10 years	11-15 years	16-20 years	>20 years
○	○	○	○	○

Q.4 Please indicate below, what is the broad area of your teaching/training?

Social Science / Business	STEM	Others
○	○	○

Q.5 Please indicate below, what is the level of the students you teach?

Higher Education or University	TVET (Technical and Vocational Education and Training)	Others (Corporate training etc.)
○	○	○

Section B. – Respondent's Hearing condition.

Q.6 Please select an option below that best describes your current hearing condition.

Normal	Mild (may have difficulty hearing soft spoken people)	Moderate (you can hear without hearing aids, but may not always understand speech)	Severe (without hearing aids, regular speech is inaudible)	Profound (without hearing aids, even very loud sounds are inaudible)
○	○	○	○	○

Q.7 Were you able to hear the student verbal feedback clearly during online classes, with your current hearing condition (as described in Q.6)?

Section C. Usage of online systems and significance of student feedback during online classes.

Q.8 How often have you conducted classes online in the last three (3) years?

> 10 Times	4-10 Times	2-3 Times	1 Time	Never
○	○	○	○	○

Q.9 Please specify the tools/platform (MS teams / Zoom / Skype etc.) you used to conduct the classes online. (Please list the two tools that you used the most)

1.

2.

Q.10 Do you think real-time student feedback during an online class is crucial for student learning and engagement, from the perspective of an educator?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.11 How do you get student feedback in an online class?

Q.12 In case of student's verbal feedback during an online class, were you able to discern the emotion (such as, positive or negative) in the student's voice? What were the challenges you faced in determining the emotion?

Positive feedback would sound happy or neutral. Negative feedback would sound more like sad, upset, fearful or maybe even angry.

Sections D. – Usability testing - effectiveness and efficiency of the speech emotion recognition system

This speech emotion recognition system helps to detect the underlying emotion from the student's speech signal during verbal feedback. It then displays the recognized emotion as positive or negative. Educators can use this system during the online class, for a selected student at a time. In case a positive emotion is detected, the educator can opt to continue the class or ask another student for feedback. In case a negative emotion is detected, he can use that as a cue to adjust his/her teaching approach.
This emotion is not based on content or context, it's purely based on intonation.

Q.13 Do you understand the purpose of the developed speech emotion recognition system?

Yes	Somewhat	Not Sure	Not Quite	No
○	○	○	○	○

Effectiveness of Speech emotion recognition system

Q.14. Is the user interface of the SER system user friendly?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.15 Were the controls for starting and stopping speech recording of the student intuitive?

Yes	Somewhat	Not Sure	Not Quite	No
○	○	○	○	○

Q.16. Did you find the real time visual representation of the detected student emotion in verbal feedback as positive or negative, along with an emoji, effective?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.17. Did you find the feature to re-play to hear the student verbal feedback as many times as you want useful?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Efficiency of Speech emotion recognition system

Q.18. Do you believe the system was able to detect and display the real time emotion (positive/negative) of the student's vocal feedback correctly?

Q.19 Were there any instances where the emotions were not correctly interpreted? If so, please provide some specific comments on your assessment.

Q.20 Do you have any suggestions for enhancing the efficiency and effectiveness of the speech emotion recognition system?

Section E – Feedback on the impact and improvement of Speech Emotion Recognition (SER) system in Online Education.

Q.21 Does a negative emotion detected in student feedback (via SER) in an online class help in adjusting the teaching strategy (i.e. making it more personalized)?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.22. Does a positive emotion detected in student feedback (via SER) in an online class reinforce confidence to continue with the ongoing class?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.23. Would you like to recommend the use of the SER system during online teaching?

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

Q.24. Do you think the SER feature would influence your perception of student engagement or comprehension during online class?

Q.25 Do you have any other feedback or comments?

We are very grateful for your time in filling this questionnaire.

Appendix 8 Guidelines for Educators

Good day. I am Aparna Vyakaranam from the School of Computer Science, University of Nottingham, Malaysia campus. I am conducting a usability testing for **speech emotion recognition (SER)** system which is part of my research work. All data obtained will be kept confidential and are accessed by the primary researcher only. Participation, completion and return of the survey are voluntary. You may **withdraw** at any time or decline to participate. Thank you for your time.

MY RESEARCH IDEA

I have developed a speech emotion recognition (SER) system which detects the emotion of the student as positive or negative from their speech signal during their verbal feedback in an online class. This detected emotion can benefit all educators, especially the late deafened in adjusting their teaching approach.

In light of this, I would like to seek your time to read the guidelines (refer to page 2) to use the system (website) and provide feedback through this **questionnaire**. Your responses will help me to understand your views on the accuracy and satisfaction relative to the developed SER system. The insights from this usability testing can help lead to any refinements for an enhanced user experience.

I seek about 30 minutes of your time.

The findings from this testing will help me understand the efficiency and the effectiveness of the developed SER system. Your contribution would be very helpful in this matter, and I will be happy to share these findings with you as well.

I am **not looking for any 'technical' or 'personal' information, but just your views** / opinions on issues related to the usage of the system. Please rest assured that all data will be kept confidential, and results will be used only in aggregate form.

If you have any concerns or require any clarifications regarding this questionnaire, please feel free to contact either

Aparna @hcxav1@nottingham.edu.my

Or

Dr.Tomas Maul @Tomas.Maul@nottingham.edu.my

Dr.Bavani @Bavani.R@nottingham.edu.my

Guidelines to the respondent (educator):

1. The developed Speech Emotion Recognition (SER) system is integrated into a website which can be accessed via a link [link will be provided on the day of testing]. The screen shot of the website is below in Pg 3.
2. Simultaneously, you need to open any online communication tool, such as MS Teams or Zoom or such [I will be initiating this].
3. AT THIS POINT you will have the SER System and the ZOOM / MS TEAMs open simultaneously on your device.

Next Step

4. Please engage two learners (students) via this selected communications platform (Zoom or MS Team or any other). They should be online as if a virtual classroom is being conducted.

   *** Students must be above 18 years of age*
5. The intent is to capture the student's voice (verbal feedback) via the online tool, which will be simultaneously captured by the SER system, to detect their underlying emotion.
6. Instruct each student to give their verbal feedback using their microphone as below.
   a. Prompt the student to start to speak into the microphone to give verbal feedback. Instruct that each verbal feedback should last for a duration of at least 5 seconds or more.

b. Simultaneously you should click on the 'Record' button on the SER website to capture live voice/audio coming via the microphone (of the online tool).

c. Once the student finishes - Click the 'Stop' button.

d. Click on the 'Show Emotion' button to enable the system to identify and display emotion based on the verbal feedback.

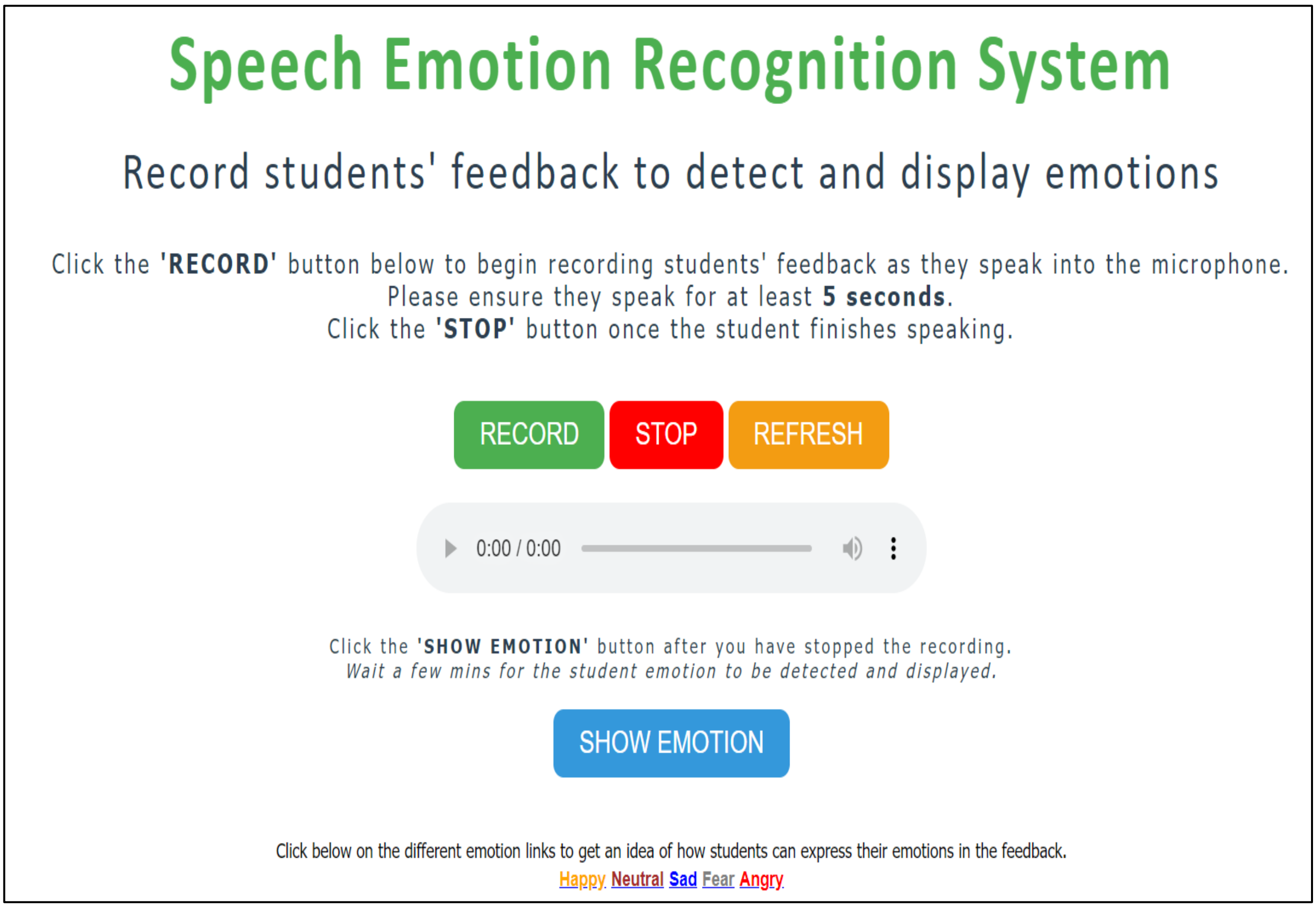
Final Steps

7. A student can mimic any one emotion at a time. They can choose from a set of positive emotions which are - **happy** or **neutral** and negative emotions which are - **sad** or **anger** or **fear**. Below please find information about how different emotions can be mimicked. There is also a script in case a student would like to refer to it. Links to different emotions are also available on the website.
8. Click the 'Refresh' button whenever you need to reload the page to record the next verbal utterance of the student to detect the underlying emotion.
9. Each student needs to give verbal feedback four times by speaking into the microphone. They should be able to mimic 2 positive and 2 negative emotions.

***The detected (underlying) student emotion is solely determined by the speaker's speech signal characteristics and is not influenced by the content or context of the speech.*

After the system testing, you need to help fill up the questionnaire.

Screen shot of the SER website -



Speech Emotion Recognition System

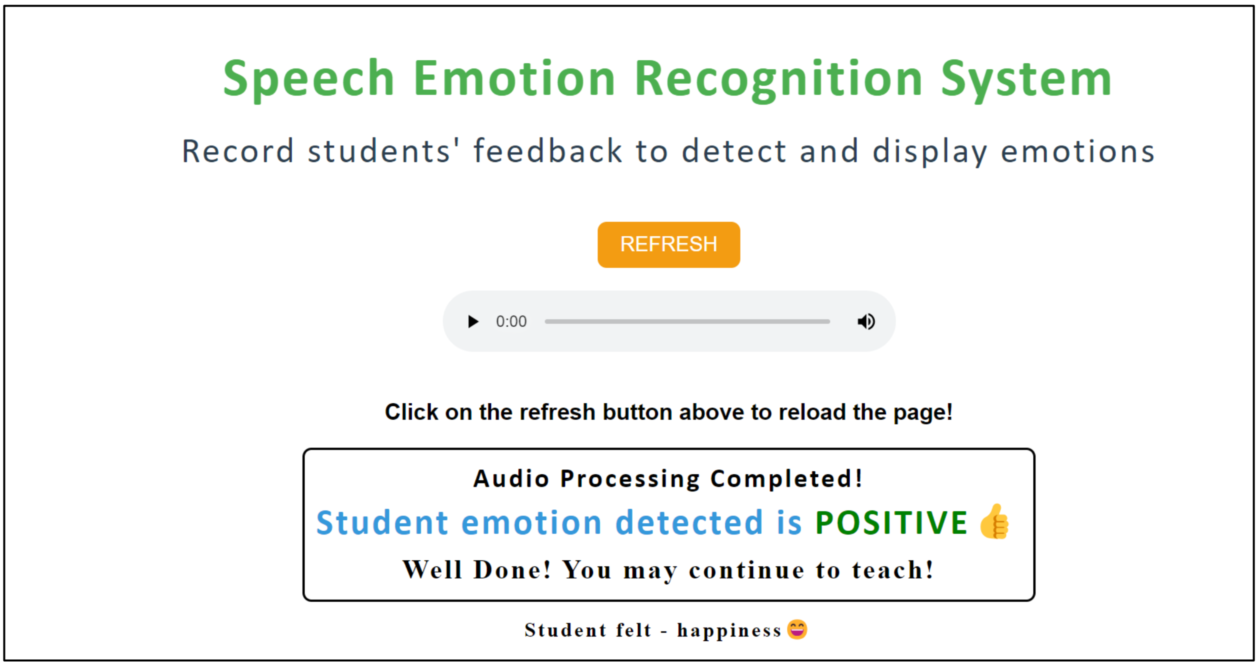
Record students' feedback to detect and display emotions

Click the **'RECORD'** button below to begin recording students' feedback as they speak into the microphone.
Please ensure they speak for at least **5 seconds**.
Click the **'STOP'** button once the student finishes speaking.

Click the **'SHOW EMOTION'** button after you have stopped the recording.
Wait a few mins for the student emotion to be detected and displayed.

Click below on the different emotion links to get an idea of how students can express their emotions in the feedback.
Happy Neutral Sad Fear Angry

If POSITIVE emotion is detected, screen shot below-



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

Click on the refresh button above to reload the page!

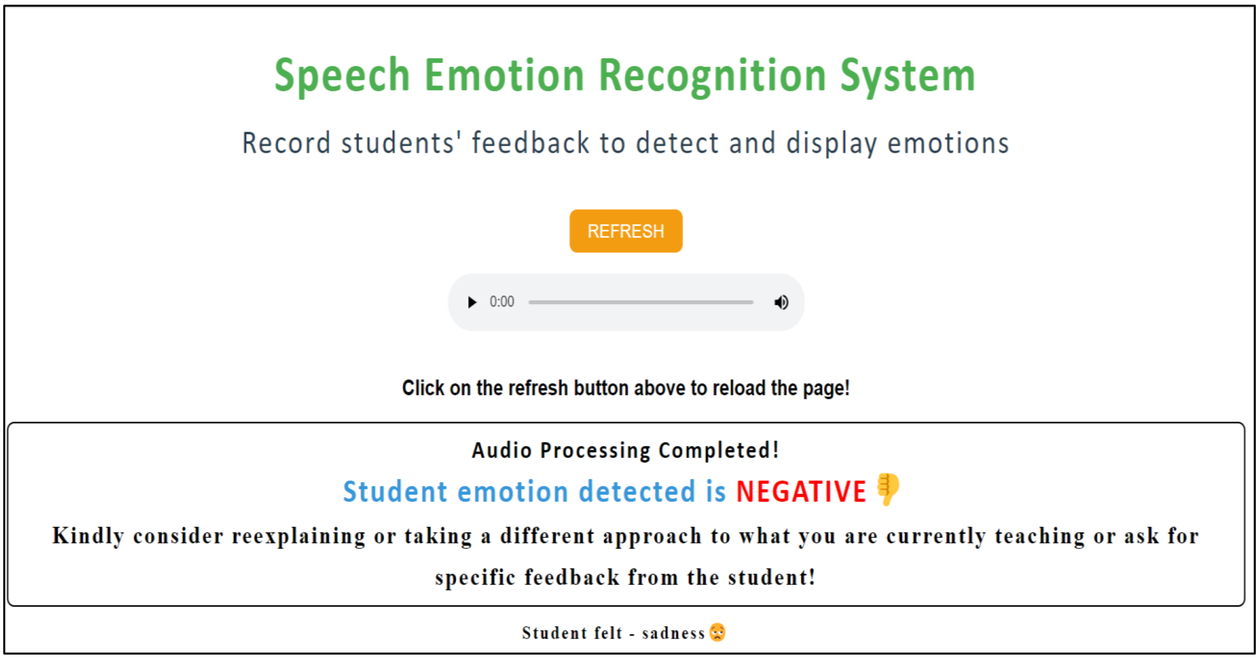
Audio Processing Completed!

Student emotion detected is POSITIVE 👍

Well Done! You may continue to teach!

Student felt - happiness 😄

And NEGATIVE emotion detected, screen shot below –



Speech Emotion Recognition System

Record students' feedback to detect and display emotions

Click on the refresh button above to reload the page!

Audio Processing Completed!

Student emotion detected is NEGATIVE 👎

Kindly consider reexplaining or taking a different approach to what you are currently teaching or ask for specific feedback from the student!

Student felt - sadness 😢

Appendix 9 Guidelines for Students

Good day. I am Aparna Vyakaranam from the School of Computer Science, University of Nottingham, Malaysia campus. I am conducting a usability testing for **speech emotion recognition (SER)** system which is part of my research work. All data obtained will be kept confidential and are accessed by the primary researcher only. Participation, completion and return of the survey are voluntary. You may **withdraw** at any time or decline to participate. Thank you for your time.

MY RESEARCH IDEA

I have developed a speech emotion recognition (SER) system which detects the emotion of the student as positive or negative from their speech signal during their verbal feedback in an online class. This detected emotion can benefit all educators, especially the late deafened in adjusting their teaching approach.

I seek about 5-10 minutes of your time to participate in this simulated online teaching session.

In light of this, I would like to seek your time to read the guidelines below -

Please click on the google form for participant consent-
https://docs.google.com/forms/d/e/1FAIpQLSfWe1G004fkBpaXfRv1sRScE2_Og-qxFNrZ2ubP-v_T4NcJRw/viewform?usp=sharing

Guidelines to the Student:

1. A zoom link will be shared with you by your educator/instructor/teacher. Click on it to join the simulated online teaching session.

2. The intent is to capture your voice (verbal feedback) via the online teaching tool being used (zoom in this case), to detect the underlying emotion in your voice.
    a. When the educator clicks on the ‘RECORD’ button on the SER website - you have start to be speaking into the microphone to give your verbal feedback. This verbal feedback should last for a duration of at least 5 seconds or more.
    b. You should sound like you are giving feedback for the ongoing online class. Sample scripts are below for your reference (Page 2)
    c. This verbal feedback must have emotions in them like sadness, fear, anger, happiness or neutral. In case you need help with how to emote, please find information below about how different emotions can be mimicked (Page 2). Links to different emotions are also available on the SER website.

3. Each student needs to give verbal feedback four times by speaking into the microphone. They should be able to mimic 2 positive and 2 negative emotions. (**sad, anger and fear fall under negative emotions whereas happy and neutral fall under positive emotions**)

***The detected (underlying) student emotion is solely determined by the speaker’s speech signal characteristics and is not influenced by the content or context of the speech.*

What effects emotion in a voice –

1**. **Tone of voice**:**

- A cheerful and enthusiastic tone may indicate that a student is confident and understands the

material well (POSITIVE).

- A flat or monotone voice might suggest boredom or disengagement (NEGATIVE).
- A hesitant or uncertain tone may signal confusion or lack of understanding (NEGATIVE).

2. **Pitch:**

- Higher pitch voices might indicate excitement or enthusiasm (POSITIVE).
- Lower pitch voices might suggest seriousness or frustration (NEGATIVE).

3. **Volume**:
   - Increased volume might signify strong emotions, such as excitement or frustration (POSITIVE).
   - Decreased volume might suggest shyness or discomfort (NEGATIVE).

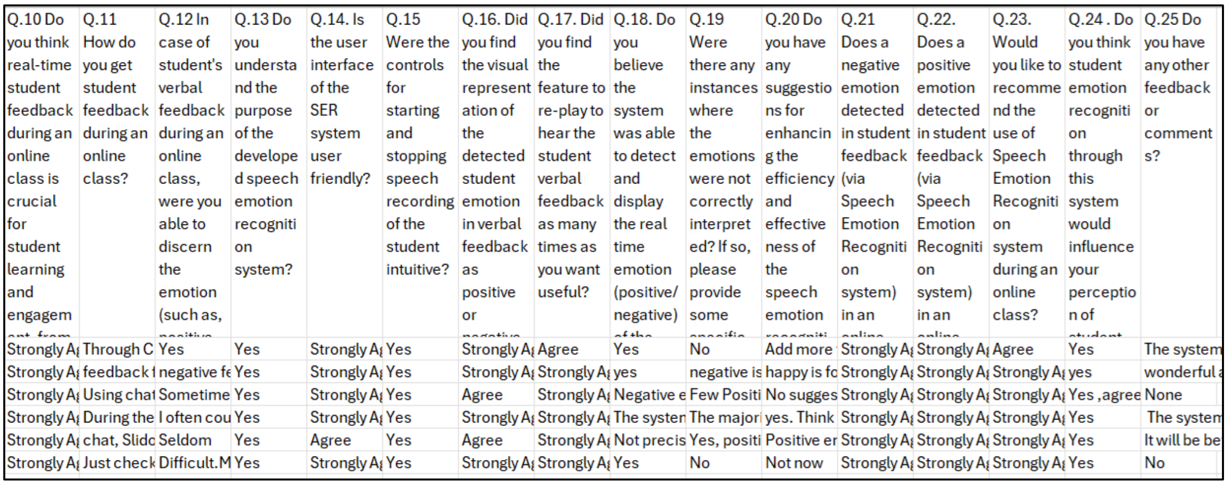
4. **Rate of speech**:
   - Rapid speech may indicate excitement or eagerness to share ideas (POSITIVE).
   - Slower speech may suggest careful consideration or uncertainty (NEGATIVE).

5. **Prosody** (patterns of stress and intonation):
   - Emphasis on certain words or phrases can highlight agreement, disagreement, confusion, or Confidence.

HOW TO SHOW DIFFERENT EMOTIONS	
NEGATIVE EMOTIONS	
Sadness / upset / confused: Lower pitch(volume) a bit. Speak with a slower pace. Use a softer tone. Allow for pauses and a more subdued delivery.	**SAMPLE SCRIPTS** **[Sadness]** "Didn't understand sir. Please explain again. Thank you."
Fear: Speak with a higher pitch. Increase the pace but allow for moments of hesitation. Use a shaky or trembling tone. Convey a sense of urgency. **Anger:** Raise your pitch. Speak with a louder and more forceful tone. Increase your pace slightly. Use assertive and direct language.	**[Fear]** "I'm a bit nervous about the upcoming exam. Could we have a review session to go over challenging topics?" **[Anger]** "I'm frustrated because it seems like some students aren't contributing to group projects. It's impacting the overall team dynamic."

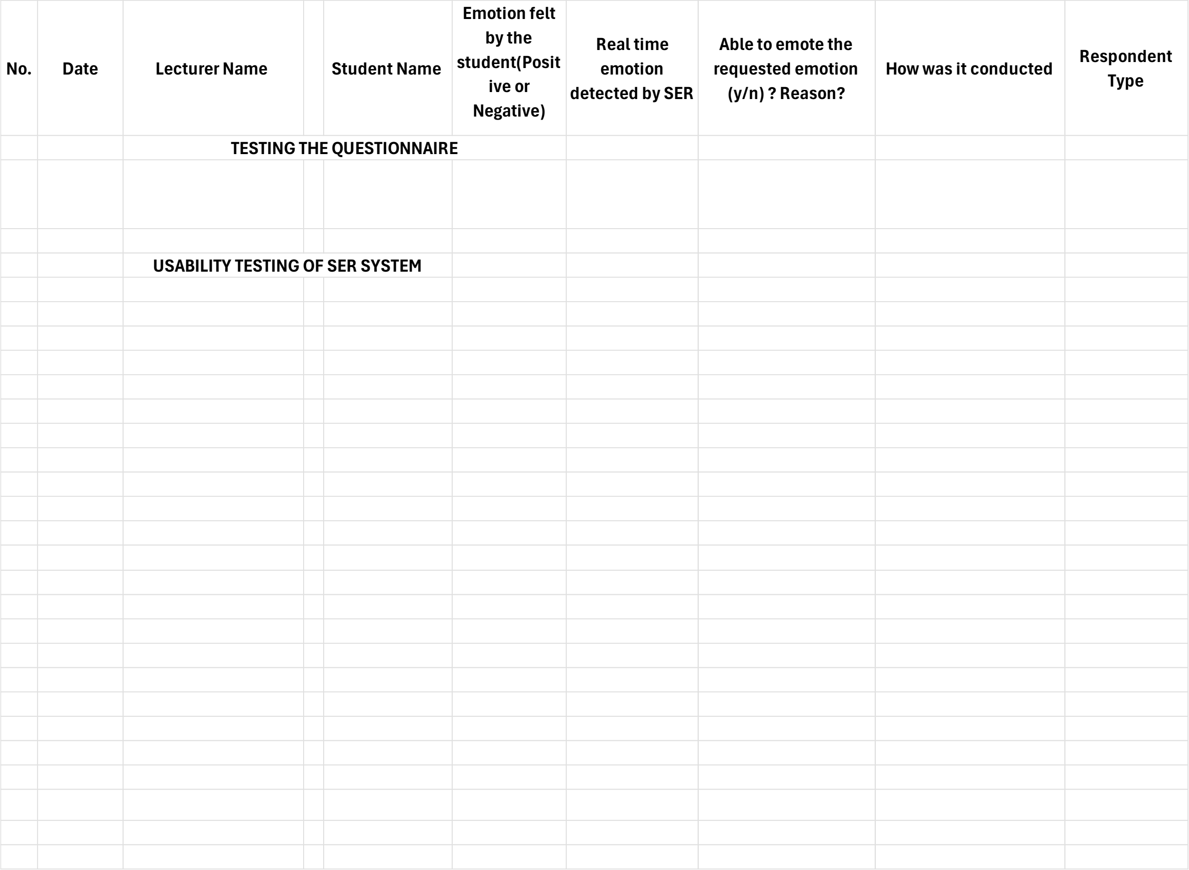
POSITIVE EMOTIONS	
Happiness / Satisfied /clear: Increase pitch(volume) slightly. Speak with a lively and upbeat tone. Use a faster pace. Add laughter or smiles in your speech. **Neutral/Calm:** Maintain a moderate pitch. Speak at a steady pace. Use a clear and composed tone. Avoid extreme variations in pitch or tone.	**[Happiness]** "yes miss all good!" **[Neutral]** "Your explanation was clear and easy to follow. I appreciate your teaching style."

Appendix 10 Sample of the Pilot User Data



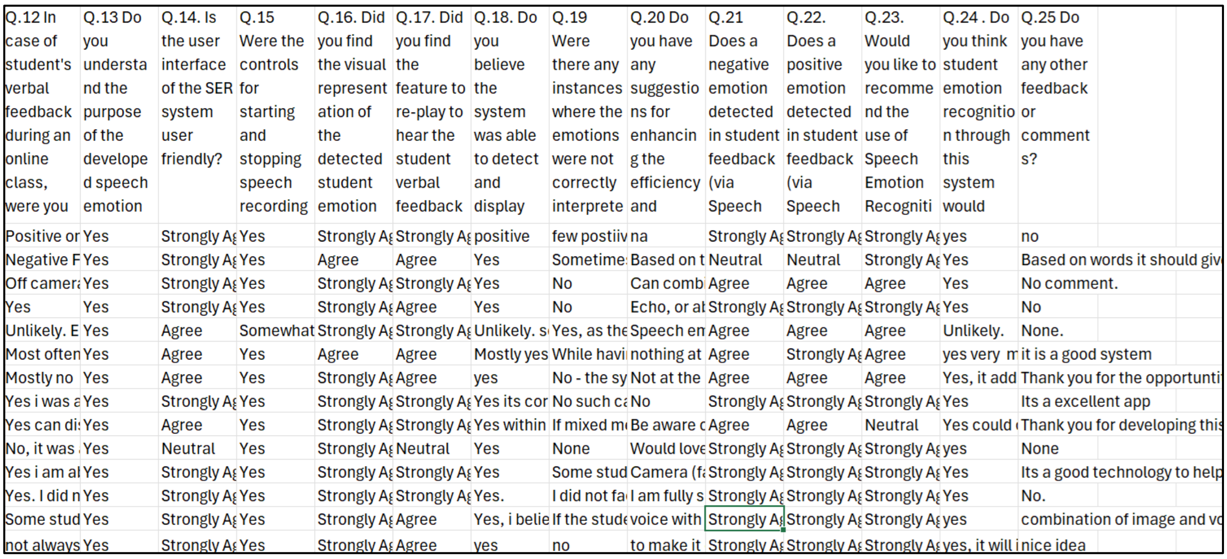
Q.10 Do you think real-time student feedback during an online class is crucial for student learning and engagement from	Q.11 How do you get student feedback during an online class?	Q.12 In case of student's verbal feedback during an online class, were you able to discern the emotion (such as, positive	Q.13 Do you understand the purpose of the developed speech emotion recognition system?	Q.14. Is the user interface of the SER system user friendly?	Q.15 Were the controls for starting and stopping speech recording of the student intuitive?	Q.16. Did you find the visual representation of the detected student emotion in verbal feedback as positive or negative	Q.17. Did you find the feature to re-play to hear the student verbal feedback as many times as you want useful?	Q.18. Do you believe the system was able to detect and display the real time emotion (positive/ negative) of the	Q.19 Were there any instances where the emotions were not correctly interpreted? If so, please provide some specific	Q.20 Do you have any suggestions for enhancing the efficiency and effectiveness of the speech emotion recogniti	Q.21 Does a negative emotion detected in student feedback (via Speech Emotion Recognition system) in an online	Q.22. Does a positive emotion detected in student feedback (via Speech Emotion Recognition system) in an online	Q.23. Would you like to recommend the use of Speech Emotion Recognition system during an online class?	Q.24 . Do you think student emotion recognition through this system would influence your perception of student	Q.25 Do you have any other feedback or comments?
Strongly A	Through C	Yes	Yes	Strongly A	Yes	Strongly A	Agree	Yes	No	Add more	Strongly A	Strongly A	Agree	Yes	The system
Strongly A	feedback t	negative fe	Yes	Strongly A	Yes	Strongly A	Strongly A	yes	negative is	happy is fo	Strongly A	Strongly A	Strongly A	yes	wonderful a
Strongly A	Using chat	Sometime	Yes	Strongly A	Yes	Agree	Strongly A	Negative e	Few Positi	No sugges	Strongly A	Strongly A	Strongly A	Yes ,agree	None
Strongly A	During the	I often cou	Yes	Strongly A	Yes	Strongly A	Strongly A	The systen	The majori	yes. Think	Strongly A	Strongly A	Strongly A	Yes	The system
Strongly A	chat, Slido	Seldom	Yes	Agree	Yes	Agree	Strongly A	Not precis	Yes, positi	Positive er	Strongly A	Strongly A	Strongly A	Yes	It will be be
Strongly A	Just check	Difficult.M	Yes	Strongly A	Yes	Strongly A	Strongly A	Yes	No	Not now	Strongly A	Strongly A	Strongly A	Yes	No

Appendix 11 Template



No.	Date	Lecturer Name		Student Name	Emotion felt by the student(Positive or Negative)	Real time emotion detected by SER	Able to emote the requested emotion (y/n) ? Reason?	How was it conducted	Respondent Type
		TESTING THE QUESTIONNAIRE							
		USABILITY TESTING OF SER SYSTEM							

Appendix 12 Sample User Data



Q.12 In case of student's verbal feedback during an online class, were you	Q.13 Do you understand the purpose of the developed speech emotion	Q.14. Is the user interface of the SER system user friendly?	Q.15 Were the controls for starting and stopping speech recording	Q.16. Did you find the visual representation of the detected student emotion	Q.17. Did you find the feature to re-play to hear the student verbal feedback	Q.18. Do you believe the system was able to detect and display	Q.19 Were there any instances where the emotions were not correctly interprete	Q.20 Do you have any suggestions for enhancing the efficiency and	Q.21 Does a negative emotion detected in student feedback (via Speech	Q.22. Does a positive emotion detected in student feedback (via Speech	Q.23. Would you like to recommend the use of Speech Emotion Recogniti	Q.24 . Do you think student emotion recognition through this system would	Q.25 Do you have any other feedback or comments?
Positive or	Yes	Strongly Ag	Yes	Strongly Ag	Strongly Ag	positive	few postiiv	na	Strongly Ag	Strongly Ag	Strongly Ag	yes	no
Negative F	Yes	Strongly Ag	Yes	Agree	Agree	Yes	Sometime	Based on t	Neutral	Neutral	Strongly Ag	Yes	Based on words it should giv
Off camera	Yes	Strongly Ag	Yes	Strongly Ag	Strongly Ag	Yes	No	Can combi	Agree	Agree	Agree	Yes	No comment.
Yes	Yes	Strongly Ag	Yes	Strongly Ag	Agree	Yes	No	Echo, or al	Strongly Ag	Strongly Ag	Strongly Ag	Yes	No
Unlikely. E	Yes	Agree	Somewhat	Strongly Ag	Strongly Ag	Unlikely. s	Yes, as the	Speech em	Agree	Agree	Agree	Unlikely.	None.
Most often	Yes	Agree	Yes	Agree	Agree	Mostly yes	While havi	nothing at	Agree	Strongly Ag	Agree	yes very m	it is a good system
Mostly no	Yes	Agree	Yes	Strongly Ag	Agree	yes	No - the sy	Not at the	Agree	Agree	Agree	Yes, it add	Thank you for the opportunit
Yes i was a	Yes	Strongly Ag	Yes	Strongly Ag	Strongly Ag	Yes its cor	No such ca	No	Strongly Ag	Strongly Ag	Strongly Ag	Yes	Its a excellent app
Yes can di	Yes	Agree	Yes	Strongly Ag	Strongly Ag	Yes within	If mixed m	Be aware c	Agree	Agree	Neutral	Yes could	Thank you for developing this
No, it was	Yes	Neutral	Yes	Strongly Ag	Neutral	Yes	None	Would love	Strongly Ag	Strongly Ag	Strongly Ag	yes	None
Yes i am al	Yes	Strongly Ag	Yes	Strongly Ag	Strongly Ag	Yes	Some stud	Camera (fa	Strongly Ag	Strongly Ag	Strongly Ag	Yes	Its a good technology to help
Yes. I did n	Yes	Strongly Ag	Yes	Strongly Ag	Strongly Ag	Yes.	I did not fa	I am fully s	Strongly Ag	Strongly Ag	Strongly Ag	Yes	No.
Some stud	Yes	Strongly Ag	Yes	Strongly Ag	Agree	Yes, i belie	If the stude	voice with	Strongly Ag	Strongly Ag	Strongly Ag	yes	combination of image and vo
not always	Yes	Strongly Ag	Yes	Strongly Ag	Agree	yes	no	to make it	Strongly Ag	Strongly Ag	Strongly Ag	yes, it will i	nice idea