Nottingham Trent University

Department of Computer Science

Speech Emotion Recognition in Virtual Reality Exposure Therapy Using Wav2Vec2-based Self-Supervised Learning

Major Project COMP40311 Coursework

Prakhar Kochhar
N1297262

Supervised by Dr. Arif Rahman

Department of Computer Science
School of Science & Technology
Nottingham Trent University

I hereby declare that this dissertation is all my own work, except as indicated in the text:

Signature 

Date 28 / 08 / 2025

I hereby declare that I have all necessary rights and consents to publicly distribute this dissertation (if needed) via the Nottingham Trent University's archive service.

Abstract

This report develops a speech emotion recognition (SER) pipeline targeted at real-time use in Virtual Reality Exposure Therapy (VRET). We build on a self-supervised acoustic encoder (Wav2Vec 2.0) [Baevski et al., 2020], fine-tuned end-to-end on acted speech corpora and then expanded to a heterogeneous merged dataset (NOR). The system ingests 16 kHz mono waveforms and uses class-weighted cross-entropy with on-the-fly waveform augmentations (additive noise, pitch/time perturbations, temporal shift, band-pass) to improve robustness to microphone and room variability. We adopt deployment-oriented inference with short overlapping windows and probability aggregation to support low-latency scene [Chang et al., 2020] adaptation in VRET.

On homogeneous acted datasets the model approaches ceiling performance: **99.8%** accuracy on TESS [Dupuis and Pichora-Fuller, 2010] and **97.6%** accuracy on a 5-class RAVDESS split (angry, neutral, happy, sad, fear; *disgust* excluded) [Livingstone and Russo, 2018a]. On the larger, acoustically diverse NOR corpus, the model attains **82.19%** accuracy and **weighted-F1 $\approx$ 0.822**. Row-normalised confusion matrices show that residual errors concentrate among negative-valence emotions (*sad*/*fear*/*disgust*), while *neutral* and *angry* remain comparatively stable. Despite this, one-vs-rest ROC curves yield micro-AUC $\approx$ **0.969**, indicating strong probability-space separability and suggesting that calibrated class-specific thresholds could further improve operating points under domain shift. Classical baselines (SVM/MLP with MFCCs; CNN on mel-spectrograms) provide competitive references on acted speech but trail the transformer under cross-corpus conditions; MFCC feature-importance and unsupervised clustering analyses corroborate the observed confusion structure.

The implementation emphasises reproducibility (speaker-disjoint splits; per-epoch logs; regenerated metrics and plots) and produces a compact artefact suitable for streaming inference. We also propose a VRET integration architecture that connects headset audio capture to the SER module and a scene controller via lightweight APIs, with temporal smoothing and hysteresis to stabilise actuation. Overall, the results indicate that self-supervised acoustic representations, combined with pragmatic augmentation and thresholding, provide a viable foundation for therapist-supervised, real-time affect adaptation in virtual exposure scenarios.

Acknowledgements

I would like to express my sincere gratitude to **Dr Arif Rahman** for his guidance, constructive feedback, and steady encouragement throughout this project. His advice helped shape the research questions, sharpen the experimental design, and strengthen the final write-up. I am also grateful to my second marker and the examination panel for their insightful comments, which improved the clarity and rigour of the work.

My thanks go to the members of our research group and classmates for helpful discussions, code reviews, and practical suggestions during implementation and evaluation.

This project relied on the generosity of the wider research community. I gratefully acknowledge the curators and contributors of the TESS and RAVDESS datasets, and the sources that underpin the merged NOR corpus used in this study. I also thank the open-source projects that made the work possible, including PyTorch, Hugging Face Transformers, Librosa, scikit-learn, Matplotlib, and the wider Python ecosystem, as well as Kaggle for compute and storage resources used during training.

Finally, I am deeply thankful to my family and friends for their patience, encouragement, and unwavering support. Their belief in me made the long experiments and late edits not only possible but enjoyable.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Virtual Reality (VR) has transformed from a technological novelty into a powerful tool across diverse domains, including mental health treatment. One of the most impactful applications of VR in psychology is Virtual Reality Exposure Therapy (VRET), a method that immerses individuals in computer-generated environments to help them confront and gradually reduce their fears or anxieties in a controlled and safe setting. A seminal study by Barbara Rothbaum et al. (1995) was among the first to empirically validate the use of VRET for treating acrophobia, laying the foundation for decades of VR-assisted psychological therapies [Rothbaum et al., 1995].

According to a WHO report, more than 264 million people worldwide suffer from anxiety disorders, which are ranked as the sixth most common contributor to global disability [Organization et al., 2017].

Since its clinical validation in the 1990s, VRET has expanded to treat a wide range of anxiety-related disorders, including phobias, PTSD, and social anxiety.

Despite these advancements, one of the lasting limitations of self-guided VRET systems is their inability to perceive and adapt to the emotional state of the user in real time [Rahman et al., 2023]. This limitation is critical, especially when emotional responses vary widely between individuals and even between sessions.

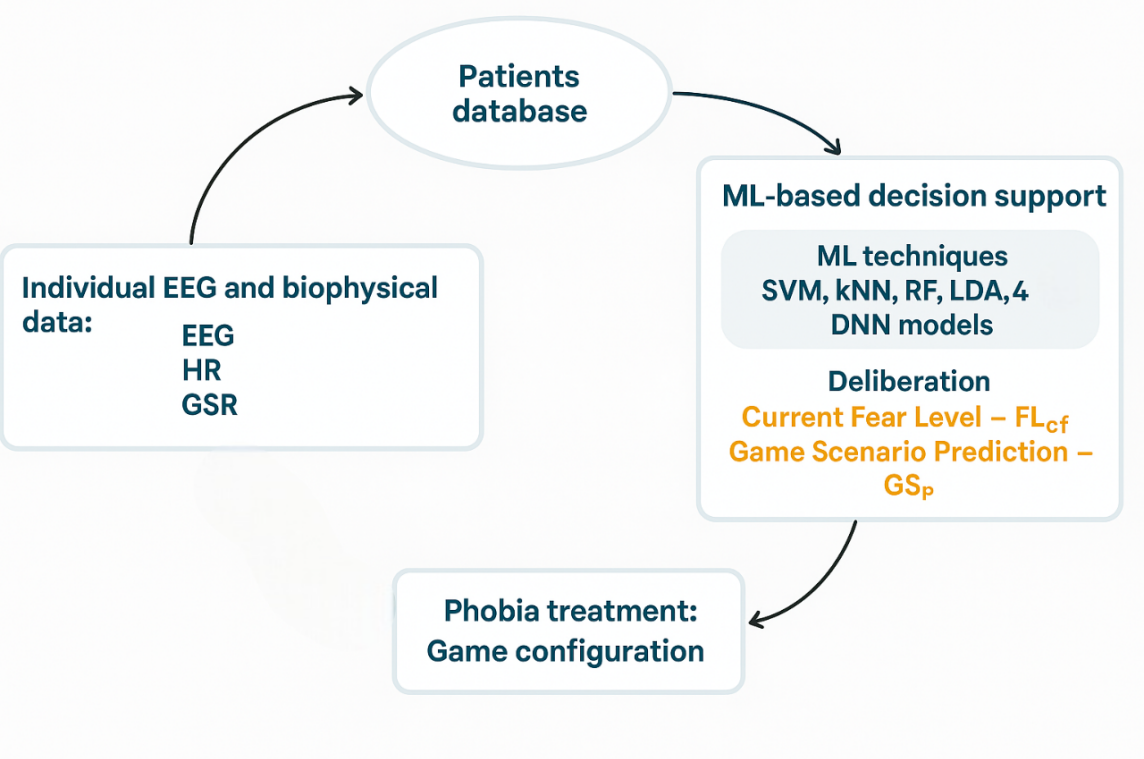


Figure 1.1: Machine learning-powered support system for treating phobias.

Source: Adapted from Bălan et al. (2020).

Recent studies have shown that monitoring physiological signals such as electrodermal activity (EDA), heart rate variability (HRV), and galvanic skin response (GSR) during VRET sessions can greatly enhance understanding of a patient's affective state and personalize the therapeutic process [Šalkevicius et al., 2019]. However, many of these approaches are based on wearable biosensors, which may not always be comfortable, accessible, or feasible in consumer-facing therapy systems.

Table 1.1: Comparison of Emotion Recognition Modalities

Modality	**Sensors Used**	**Invasiveness**	**Comfort**	**Real-Time Capable?**
EEG	Headset	High	Low	Yes
GSR/EDA	Wristband	Medium	Medium	Yes
SER	Microphone	Low	High	Yes

This motivates the integration of Speech Emotion Recognition (SER) as an alternative or complementary channel for emotion detection, one that is passive, continuous, and

already naturally present in human interaction.

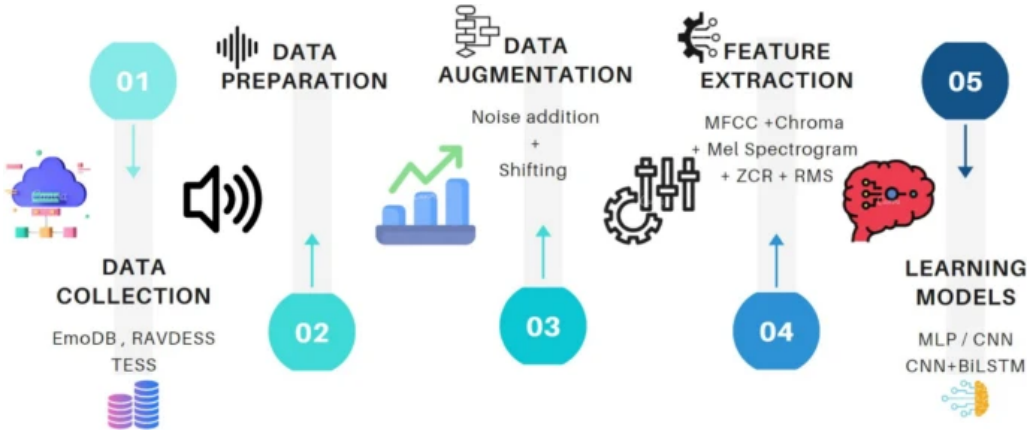


Figure 1.2: Structural Blueprint of the SER Approach.

Source: Adapted from [Barhoumi and BenAyed, 2024]

To enable accurate detection of emotional states from speech in real time, Convolutional Neural Networks (CNNs) and deep CNN architectures have emerged as the most effective techniques in recent SER research [Issa et al., 2020]. These models are capable of learning complex feature hierarchies from audio representations such as mel-spectrograms or (Mel-Frequency Cepstral Coefficients) MFCCs, eliminating the need for extensive manual feature engineering. Deep CNNs, when trained on emotional speech datasets, have been shown to outperform traditional classifiers by capturing both spectral and temporal characteristics of emotion-rich audio [Zhao et al., 2019]. Moreover, advances in hybrid architectures, such as CNN-LSTM and attention-based models, further improve classification by modeling long-range temporal dependencies in speech.

This project focuses on applying deep learning-based SER models, particularly deep CNNs, within a VRET system to recognize and classify emotional states from speech data in real-time. The goal is to enable emotion-adaptive VR therapy that can modulate exposure intensity, alter scene dynamics, or offer supportive interventions based on the user's detected emotional state. Using CNN-based speech emotion recognition to leverage the power of speech emotion recognition, this research aims to contribute to the building of autonomously adaptive VRET systems powered by emotional intelligence.

To bridge this gap, Speech Emotion Recognition (SER) has emerged as a non-invasive, real-time solution for emotional state monitoring. By analyzing vocal attributes such

as pitch, tone, and prosody, SER systems can detect emotional states such as stress, fear, or calmness [Cao et al., 2017], providing essential feedback for dynamically adapting therapeutic interventions. This becomes especially relevant in exposure-based therapy, where the intensity of virtual stimuli must align with the user's ability to tolerate distress [Graham et al., 2025].

Emotions	Pitch	Intensity	Speaking rate	Voice quality
Anger	abrupt on stress	much higher	marginally faster	breathy, chest
Disgust	wide, downward inflections	lower	very much faster	grumble chest tone
Fear	wide, normal	lower	much faster	irregular voicing
Happiness	much wider, upward inflections	higher	faster/slower	breathy, blaring tone
Joy	high mean, wide range	higher	faster	breathy; blaring timbre
Sadness	slightly narrower	downward inflections	lower	resonant

Figure 1.3: Overview of Key Acoustic Changes Identified Across Different Emotional States.

Source: Adapted from [Khalil et al., 2019]

Despite the growing use of VRET for the treatment of anxiety disorders, a significant limitation persists in self-guided systems: the absence of real-time affect-sensitive mechanisms. This deficiency hampers the system's ability to personalize therapeutic experiences based on the user's current emotional state. Studies have highlighted that while self-guided VRET can effectively reduce anxiety symptoms, the lack of adaptive feedback mechanisms can limit its efficacy compared to therapist-led interventions [Premkumar et al., 2024].

Although this project focuses primarily on public speaking anxiety (PSA) as a case study, the proposed Speech Emotion Recognition (SER) framework is designed to be adaptable

for various anxiety disorders. The underlying architecture can be extended to address conditions such as social anxiety, specific phobias, and generalized anxiety disorder. Research indicates that VRET has been successfully applied to a variety of anxiety-related conditions [Bălan et al., 2020], suggesting the potential for a greater applicability of adaptive SER-integrated VRET systems.

Table 1.2: Summary of VRET Applications for Various Anxiety Disorders

Disorder Treated	**Study/Year**	**Sample Size**	**VRET Outcome**
PTSD	[Beidel et al., 2019]	92	Reduced symptoms
Acrophobia	[Diemer et al., 2016]	40	Triggered fear response
Dental Phobia	[Raghav et al., 2016]	30	Anxiety score decreased
Public Speaking	[Reeves et al., 2022]	508	Reductions in PSA%

The implementation of the SER system involves training deep Convolutional Neural Networks (CNNs) on established emotional speech datasets, such as RAVDESS and EmoDB [Wani et al., 2021]. These datasets provide a diverse range of emotional expressions, enabling the model to learn intricate patterns associated with various affective states. The trained model will be integrated into a simulated VRET environment to assess its performance in real-time emotion classification and its impact on therapy personalization.

Table 1.3: Widely Used Datasets in Speech Emotion Recognition Research.

No.	**Database**	**Language**	**Type**	**Size**	**Emotions**
1	Berlin EMO-DB Akinpelu and Viriri [2024]; Burkhardt et al. [2005]	German	Acted	7 emotions, 10 speakers (5 male, 5 female)	Neutral, sadness, fear, boredom, happiness, disgust, anger

No.	Database	Language	Type	Size	Emotions
2	SAVEE Jackson and Haq [2014]	English	Acted	7 emotions, 4 speakers (male)	Surprise, anger, fear, disgust, sadness, happiness, neutral
3	RAVDESS Livingstone and Russo [2018a]	English	Acted	24 speakers (12 female, 12 male), 8 emotions	Calm, happy, sad, angry, fearful, surprise, disgust, neutral
4	RECOLA Ringeval et al. [2013]	French	Natural	7 hours, 46 speakers (19 males, 27 females)	agreement, rapport, dominance, valence and arousal
5	SEMAINE McKeown et al. [2011]	English, Greek, Hebrew	Natural	959 conversations, 150 speakers	power, valence, expectation, activation, overall emotional intensity
6	eNTERFACE'05 Martin et al. [2006]	English	Elicited	1116 videos, 42 speakers	Surprise, sadness, happiness, fear, anger, disgust
7	IEMOCAP Busso et al. [2008]	English	Elicited	Five sessions where each session includes the conversation between one male and one female	Anger, happiness, sadness, frustration, neutral

No.	Database	Language	Type	Size	Emotions
8	FAU Aibo Batliner et al. [2008]	German	Natural	9 hours of speech, 51 children talking to robot dog Aibo	Bored, joyful, helpless, anger, reprimanding, surprised, neutral, rest, emphatic, motherse, touchy
9	BAUM-1 Zhalehpour et al. [2016]	Turkish	Natural and Acted	1222 videos, 31 speakers	Anger, disgust, fear, bothered, sadness, surprise, concentration, contempt, being thoughful, unsure, happiness, boredom, interest
10	Persian Emotional Speech Esmaileyan and Marvi [2014]	Persian	Simulated	748 utterances, 33 speakers	Happiness, sadness, surprise, fear, anger, boredom, neutral, disgust
11	Assamese Emotion Speech Dataset Kandali et al. [2008]	Assamese	Simulated	140 utterances, 30 speakers	Fear, anger, happiness, surprise, sadness, disgust, neutral
12	Chinese Emotion Speech Dataset Tao et al. [2006]	Chinese	Simulated	1500 utterances	Anger, fear, happiness, neutral

No.	Database	Language	Type	Size	Emotions
13	SAFE (Situation Analysis in a Fictional and Emotional Database) Clavel et al. [2006]	English	Elicited	4724 segments of speech	Positive, negative, neutral

Source: Adapted from Wani et al., 2021

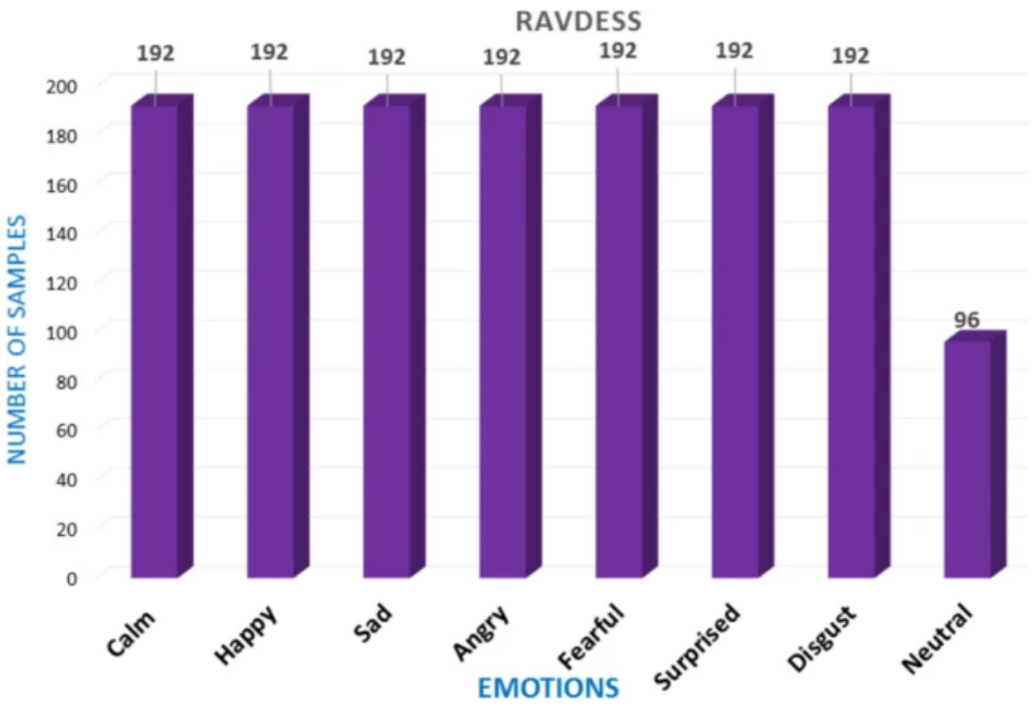


Figure 1.4: RAVDESS emotion categories.

Source: Adapted from [Barhoumi and BenAyed, 2024]

Integrating SER into self-guided VRET systems holds significant promise for enhancing the scalability and accessibility of mental health interventions. By enabling real-time emotional state monitoring, these systems can dynamically adjust therapeutic content, providing personalized experiences without the constant need for therapist supervision.

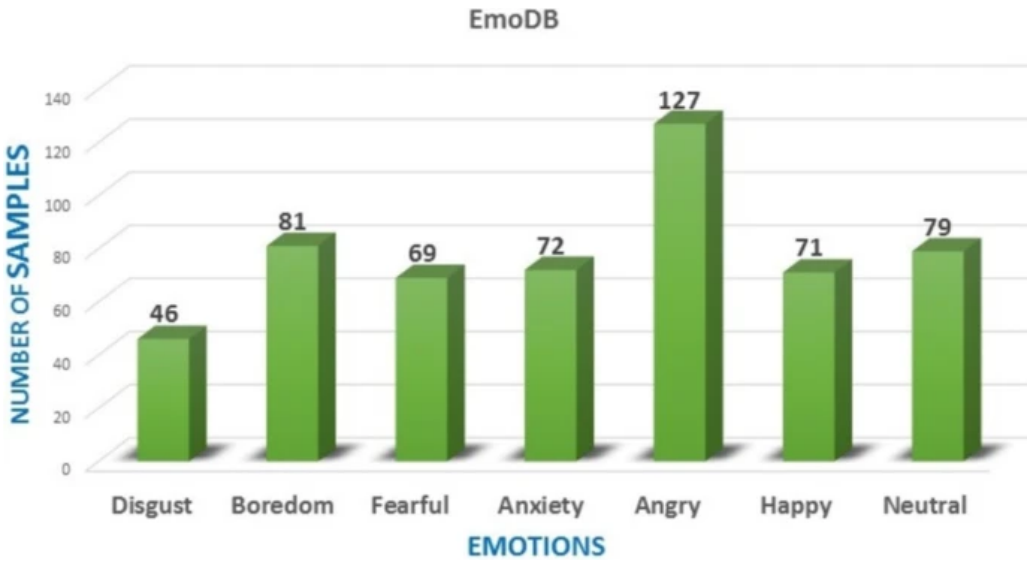


Figure 1.5: EmoDB emotion categories.

Source: Adapted from [Barhoumi and BenAyed, 2024]

Such advancements are particularly valuable in contexts with limited access to mental health professionals, offering a cost-effective solution to address the growing demand for mental health services [Hildebrand et al., 2022].

> *Scope note.* While the Motivation surveys CNN-based SER, the implemented system in this report uses *self-supervised* encoders (Wav2Vec2) as the primary approach; CNNs are maintained as baselines for comparison.

Why Self-Supervised Speech Encoders (SSL)? While CNNs on mel-spectrograms have been effective for SER, they rely on hand-crafted front-ends and often degrade under domain shift (speaker/microphone/room). Recent *self-supervised* speech models—especially **wav2vec 2.0**—learn general-purpose acoustic representations directly from raw waveforms using large-scale unlabelled audio, and then need only a small supervised head for downstream tasks. In emotion recognition, this has yielded stronger accuracy and cross-corpus generalisation than MFCC/CNN baselines, with a simpler, feature-light pipeline [Baevski et al., 2020; Chen et al., 2022; Hsu et al., 2021]. Accordingly, this dissertation adopts an *SSL-first design*: a Wav2Vec2 encoder with a lightweight classification head, retaining MFCC/CNN pipelines purely as baselines. This choice also supports the *low-latency, streaming* inference required by VRET—using short, overlap-

ping windows to trigger timely adaptive changes in the VR scene (see Aims & Objectives and Research Questions).

1.2 Aims and Objectives

1.2.1 Aim

To design, develop, and evaluate a *low-latency* Speech Emotion Recognition (SER) module built on **self-supervised speech encoders** (e.g., Wav2Vec2 transformers) operating directly on raw 16 kHz waveforms, and to integrate this module into **Virtual Reality Exposure Therapy** (VRET) for Public Speaking Anxiety (PSA). The goal is to enable real-time emotion detection that can drive *adaptive, personalised* adjustments within the VR scenario. Classical MFCC/CNN pipelines are implemented only as *baselines for comparison*, not as the primary approach.

1.2.2 Objectives

1. **Literature synthesis:** Review VRET for anxiety (with emphasis on PSA), SER, *self-supervised* speech learning (wav2vec 2.0 family), and emotion-adaptive systems to identify gaps and best practices.

2. **Data curation and protocol:** Select and harmonise suitable speech–emotion corpora (e.g., RAVDESS, TESS, CREMA-D, a heterogeneous corpus), define a consistent label set relevant to PSA, enforce *speaker-disjoint* splits, and address class imbalance.

3. **SSL model design:** Implement a Wav2Vec2-based pipeline with a lightweight classification head on top of frozen/partially unfrozen encoder layers; compare against *classical baselines* (e.g., MFCC+SVM/CNN) to quantify the benefit of SSL.

4. **Robustness via augmentation:** Employ on-the-fly audio perturbations (additive noise, reverberation/impulse responses, mild time-stretch/pitch-shift) and ablate their effect on generalisation.

5. **Cross-corpus generalisation:** Evaluate transfer across datasets and recording conditions (speakers, microphones, rooms), including train-on-one/test-on-another experiments to measure domain shift.

6. **Streaming inference engineering:** Propose a windowed/overlapping inference loop (sliding window + stride) to achieve *sub-second* end-to-end latency suitable for real-time VRET.

7. **VRET integration:** Propose an interface that maps detected emotions to adaptive VR controls (e.g., audience size/noise, prompts, task difficulty) and demonstrate closed-loop behaviour in a simulated PSA scenario.

8. **Evaluation and reporting:** Report Accuracy, weighted-F1, and confusion matrices; when relevant, include ROC/AUC and latency/throughput metrics for the streaming setting.

9. **Reproducibility & ethics:** Provide seeds, configs, and code; use only public datasets with no PII (Personally Identifiable Information); document risks, limitations, and alignment with institutional research ethics.

1.3 Research Questions

To align the SSL-based design (Wav2Vec 2.0 [Baevski et al., 2020]) with the intended VRET deployment, we pose the following questions and specify how each will be evaluated using TESS [Dupuis and Pichora-Fuller, 2010], RAVDESS (5-class: angry, neutral, happy, sad, fear) [Livingstone and Russo, 2018a], and the merged NOR corpus.

Table 1.4: Research questions, motivation, and planned evaluation.

No.	Research Question	Motivation	Planned Evaluation
RQ1	Do self-supervised encoders on raw waveforms (Wav2Vec 2.0) outperform classical MFCC/MLP/CNN pipelines on acted speech?	Establish whether SSL representations provide a measurable gain over feature-engineered baselines for SER on clean, controlled data.	Train/evaluate Wav2Vec 2.0 vs. SVM/MLP (MFCC) and CNN (mel) on **TESS** and **RAVDESS-5**; report Accuracy and Macro/Weighted-F1.
RQ2	How well does a Wav2Vec 2.0 SER model generalise under domain shift (speakers, mics, rooms), and do class-weighted loss and on-the-fly waveform augmentation help?	Quantify robustness on heterogeneous audio; mitigate imbalance and channel variation.	Fine-tune on **NOR** with class-weighted cross-entropy and augmentations; report Accuracy/F1, confusion matrix, and ROC/AUC.
RQ3	What error structure (confusable classes) emerges across datasets?	Identify systematic confusions to inform thresholding/smoothing.	Row-normalised confusion matrices on TESS, RAVDESS-5, and NOR.
RQ4	Do model probabilities exhibit strong ranking (threshold-agnostic performance) on heterogeneous data?	If AUC is high, calibrated thresholds may improve operating points.	One-vs-rest ROC on **NOR**; class-wise and micro-AUC.
RQ5	Are short windowed-inference settings (2–3 s, 50% hop) compatible with low-latency VRET control?	Ensure timely, stable signals for scene adaptation.	Implement overlapping windows with probability aggregation; estimate throughput/latency budget.

1.4 Tasks

To enable a systematic, organized, and effective implementation of the research project, the overall workflow has been divided into distinct tasks and subtasks. Each task indicates a significant milestone on the continuum extending from the initial phases of conceptualization until final submission and evaluation. Subtasks aim to break each major task into workable components, thereby making systematic progress, tracking, and fulfillment possible. This task division improves the effective resource use, supervision, and quality assurance during the project's lifespan. The following table describes an overview of the upcoming tasks along with their respective subtasks, which collectively guide the design of the Speech Emotion Recognition system as part of a Virtual Reality Exposure Therapy (VRET) setup for adaptive treatment of anxiety.

Table 1.5: Detailed Project Task Breakdown

No.	Tasks	Subtasks
1	Define Project Scope	1.1 Identify the problem statement and research objectives. 1.2 Define the scope and boundaries of the project. 1.3 Establish success criteria and measurable goals.
2	Conduct Literature Review	2.1 Search and gather relevant literature on VRET, SER, and anxiety therapy. 2.2 Categorize and summarize key findings. 2.3 Identify research gaps and theoretical frameworks.
3	Formulate Research Questions	3.1 Develop primary and secondary research questions. 3.2 Align questions with project aims and objectives.

No.	Tasks	Subtasks
4	Select and Prepare Datasets	4.1 Choose suitable emotional speech datasets. 4.2 Preprocess and clean the data. 4.3 Handle class imbalances if needed.
5	Feature Extraction	5.1 Extract MFCCs, pitch, energy, and prosodic features. 5.2 Generate spectrograms for CNN-based input.
6	Baseline Model Implementation	6.1 Train traditional ML models. 6.2 Evaluate and record baseline results.
7	Deep Learning Model Development	7.1 Build and train CNN models. 7.2 Experiment with CNN-LSTM and attention mechanisms.
8	Model Evaluation	8.1 Evaluate using accuracy, F1-score, confusion matrix. 8.2 Compare ML and DL performance.
9	VRET System Design	9.1 Design the self-guided VRET flow. 9.2 Define emotional triggers and system responses.
10	Integration Plan	10.1 Develop the pipeline from SER to VRET. 10.2 Define real-time feedback actions.
11	Prototype Development	11.1 Build a basic simulated VRET-SER integration. 11.2 Evaluate latency and real-time processing.

No.	Tasks	Subtasks
12	Review All Deliverables	12.1 Review completeness and quality. 12.2 Cross-check with objectives. 12.3 Prepare documentation.
13	Incorporate Feedback	13.1 Collect feedback from supervisors and peers. 13.2 Apply necessary revisions. 13.3 Finalize the updated report.
14	Final Documentation	14.1 Compile all sections, methods, and results. 14.2 Ensure formatting and completeness.
15	Presentation Preparation	15.1 Create slides and diagrams. 15.2 Summarize findings clearly.
16	Practice Oral Defense	16.1 Conduct mock sessions. 16.2 Prepare for possible questions.
17	Submit	17.1 Submit final report and presentation materials.

1.5 Resources required

To ensure the successful execution of this project—from data collection and preprocessing to model training and prototype integration—a range of technical and academic resources are required. These resources support each phase of the system design, from building a robust Speech Emotion Recognition (SER) model to embedding it in a Virtual Reality Exposure Therapy (VRET) framework. The table below outlines the primary resources required for this research, categorized by type, and describes their specific importance in relation to the project's objectives and implementation strategy.

Table 1.6: Resources Required for the Project

Resource Type	**Resource**	**Importance / Relevance**
Datasets	RAVDESS, EmoDB, TESS	Essential for training and evaluating SER models on various emotional states relevant to anxiety detection. These datasets provide validated samples for classification tasks.
Development Tools	Python, Jupyter Notebook, Google Colab, TensorFlow, Keras, scikit-learn	Core platforms for implementing and training ML and DL models. Google Colab provides cloud GPU access, crucial for handling deep CNN architectures efficiently.
Feature Extraction Tools	Librosa, torchaudio	Required for extracting MFCCs, mel-spectrograms, and other acoustic features needed to train emotion classification models.

Resource Type	**Resource**	**Importance / Relevance**
Virtual Reality Tools	Unity3D (or Unreal Engine) for VRET simulation	Needed to simulate public speaking environments and integrate SER output into the therapy flow for adaptive feedback.
Hardware	High-performance system with GPU / Cloud resources (e.g., Colab Pro, AWS EC2 GPU)	Training deep models (like CNNs and LSTMs) requires significant computational resources; cloud-based GPUs ensure scalability.
Academic Literature	IEEE Xplore, Google Scholar, Scopus	Crucial for understanding previous work on SER, VRET, and adaptive therapy systems. Supports literature review and framing research questions.
Documentation Tools	Overleaf, LaTeX, Mendeley / Zotero	For writing, formatting, and citing academic content, as well as managing project documentation and references.
Visualization Tools	Matplotlib, Seaborn, Tableau (if needed)	To visualize model accuracy, loss curves, confusion matrices, and system architecture for analysis and reporting.

1.6 Project risks

In any project that involves the integration of advanced technologies—such as deep learning-based Speech Emotion Recognition (SER) and Virtual Reality Exposure Therapy (VRET)—the presence of uncertainties, complexities, and external dependencies introduces a wide array of potential risks. These risks can stem from a variety of sources, including technical constraints, data limitations, system integration challenges, and broader ethical and regulatory considerations. As such, identifying, analyzing, and mitigating these risks at an early stage is not just good practice but a critical success factor for achieving the desired project outcomes.

To ensure proactive management of potential obstacles that could affect the successful execution of this project, a comprehensive and structured risk assessment has been conducted. This assessment focuses on pinpointing vulnerabilities across the full life cycle of the system—from conceptualization and model design to deployment and evaluation in a simulated VRET setting. It takes into account both internal challenges (such as model overfitting, dataset quality, and integration complexity) and external factors (such as ethical concerns, stakeholder expectations, and submission timelines).

Each identified risk has been evaluated according to two fundamental dimensions:

1. Probability(Prob.) – the likelihood of the risk occurring given current resources, practices, and dependencies.

2. Impact – the potential magnitude of disruption the risk could cause to the project timeline, deliverables, or quality.

Table 1.7: Project Risks and Mitigation Strategies (Across All Risk Levels)

Risk Description	**Prob.**	**Impact**	**Overall Risk**	**Mitigation Strategy**
Inadequate Dataset	4	5	Very High	Combine diverse datasets; apply augmentation and synthetic data generation.
Emotion Misclassification	4	4	Very High	Fine-tune models on task-specific data; integrate real-time feedback loops.
Data Quality Issues	3	4	High	Apply robust preprocessing techniques; validate and cross-check data from multiple sources.
Integration Complexity	3	3	High	Use modular architecture; perform phased integration and testing.
Privacy Concerns	2	5	Medium	Use anonymized, ethically sourced public datasets only.
Scope Misalignment	2	4	Medium	Conduct scope alignment meetings and confirm with the supervisor.
Tool Compatibility Issues	2	3	Low	Use popular ML frameworks (TensorFlow, PyTorch); fix framework versions.
Presentation Delays	3	2	Low	Allocate time for dry runs and build buffers into deadlines.
Literature Gaps	1	3	Very Low	Use academic alerts; review recent SER and VRET publications weekly.

Table 1.8: Risk Rating Levels and Corresponding Colors

Level	Likelihood / Impact	Description	Color
5	Very Likely / Critical	Very High Risk	
4	Likely / Significant Impact	High Risk	
3	Moderate	Medium Risk	
2	Unlikely / Some Impact	Low Risk	
1	Very Unlikely / Minimal	Minimal Risk	

1.7 Professional issues

The integration of Speech Emotion Recognition (SER) with Virtual Reality Exposure Therapy (VRET) for mental health intervention brings numerous professional, ethical, and legal considerations. Addressing these responsibly is crucial to ensure the system is safe, inclusive, transparent, and ethically sound [Hanna et al., 2024].

1.7.1 Ethical Use of AI in Mental Health Contexts

Issue: SER systems interpret human emotions, which are deeply personal and context-sensitive. Misclassifications in a therapeutic setting could misguide users or worsen emotional distress.

How to Address:

- Clearly communicate that the system is a supportive tool, not a replacement for licensed psychological evaluation.
- Include disclaimers in the interface or documentation.
- Ensure the system is used under the supervision of a trained professional in clinical applications.

1.7.2 Data Privacy and Informed Consent

Issue: Handling emotional audio data requires strict adherence to privacy laws such as the UK GDPR. Improper use or lack of consent may violate participant rights [Addis

and Kutar, 2018].

How to Address:

- Use only publicly available datasets with explicit informed consent.
- Avoid storing or processing personally identifiable information (PII).
- If future data is collected, ensure participants are informed and ethical clearance is obtained.

1.7.3 Bias and Fairness in Emotion Recognition

Issue: Emotion datasets often lack diversity, potentially causing biased classification across different demographics (gender, ethnicity, accent, etc.).

How to Address:

- Use diverse and inclusive datasets.
- Evaluate model performance across subgroups to identify bias.
- Report and mitigate observed disparities using fairness techniques.

1.7.4 Accuracy, Reliability, and Accountability

Issue: Incorrect emotion classification can affect therapy outcomes or user trust. Systems used in sensitive domains must meet high reliability standards.

How to Address:

- Evaluate performance using confusion matrices, F1-score, and real-time response accuracy.
- Document known limitations in reports and user-facing materials.
- Provide manual override or human review in sensitive use cases.

1.7.5 Accessibility and Inclusivity

Issue: Systems that fail to accommodate users with speech impairments, accents, or disabilities may unintentionally exclude or misclassify them.

How to Address:

- Follow inclusive design standards and accessibility guidelines (e.g., WCAG 2.1).
- Train the model on a range of accents and speech types.
- Design VR environments that accommodate different physical and cognitive abilities.

1.7.6 Professional Transparency and Academic Integrity

Issue: Using third-party datasets or models without attribution undermines academic ethics and reproducibility.

How to Address:

- Cite all data sources, libraries, and tools used.
- Document all experimental steps, including data preprocessing and training parameters.
- Where possible, make code and methodology available for reproducibility.

1.7.7 Mental Health Risk Management

Issue: VRET simulations may unintentionally cause discomfort or anxiety, especially in users already prone to such conditions.

How to Address:

- Clearly inform users of possible risks before using the system.
- Allow users to opt-out or pause at any time.
- For clinical applications, involve certified mental health professionals and obtain approval from the ethical board.

1.7.8 Misuse of the Technology

Issue: SER systems could be misapplied in non-therapeutic settings such as surveillance or manipulation, leading to ethical violations.

How to Address:

- Clearly define intended use cases and disclaim unethical usage.
- Avoid supporting or licensing the system for unauthorized applications.
- Engage in responsible AI communication and awareness within the research community.

1.7.9 Sustainability and Maintenance Responsibility

Issue: Model performance may degrade over time due to data drift or outdated assumptions.

How to Address:

- Plan for regular re-evaluation and retraining of models.
- Use version control for datasets and models.
- Include update schedules and maintenance notes in the final documentation.

1.7.10 Final Remarks on Professional Issues

By identifying and proactively addressing these professional issues, this project supports ethical standards in AI, supports responsible research practices, and ensures the development of a user-centric and socially beneficial mental health technology.

1.8 Scope and Terminology

This report focuses on speech emotion recognition (SER) for real-time use in Virtual Reality Exposure Therapy (VRET). We use *Wav2Vec2* [Baevski et al., 2020] as the acoustic

encoder and evaluate on three corpora: **TESS** [Dupuis and Pichora-Fuller, 2010] (7 labels; *pleasant_surprise* merged into *happy*), **RAVDESS** [Livingstone and Russo, 2018b] (5 labels: angry, neutral, fear, happy, sad), and a merged **NOR** corpus (6 labels: angry, disgust, fear, happy, neutral, sad). All audio is processed as 16 kHz mono waveforms; augmentations are *train-only*. Unless otherwise specified, metrics are Accuracy and weighted/macro F1 with speaker-disjoint, stratified splits. We refer to the deployed component as the *SER module* (the model plus policies) and to the neural network itself as the *model*.

1.9 Contributions

1. A robust SER pipeline built on Wav2Vec2 with class-weighted loss and on-the-fly waveform augmentation for channel variability.

2. Transfer from TESS to RAVDESS (**97.6%**) followed by expansion to the heterogeneous NOR corpus (**82.19%** / weighted-F1 $\approx$ **0.822**).

3. A deployment-oriented inference scheme (windowed processing, probability aggregation) and a proposed VRET integration architecture for low-latency scene adaptation.

4. Comprehensive evaluation: per-epoch diagnostics, confusion matrices, ROC/AUC, plus unsupervised clustering and importance of MFCC in RAVDESS.

1.10 Chapter Roadmap

Chapter 3 details the system design and the VRET integration architecture. **Chapter 4** presents the implementation: environment, datasets, preprocessing, augmentation, training, and logging. **Chapter 5** reports results, including training curves, confusion matrices, and ROC/AUC, plus unsupervised analyzes. **Chapter 6** discusses limitations and practical implications for VRET, outlines future work and deployment steps, and finally concludes the report.

1.11 Time plan

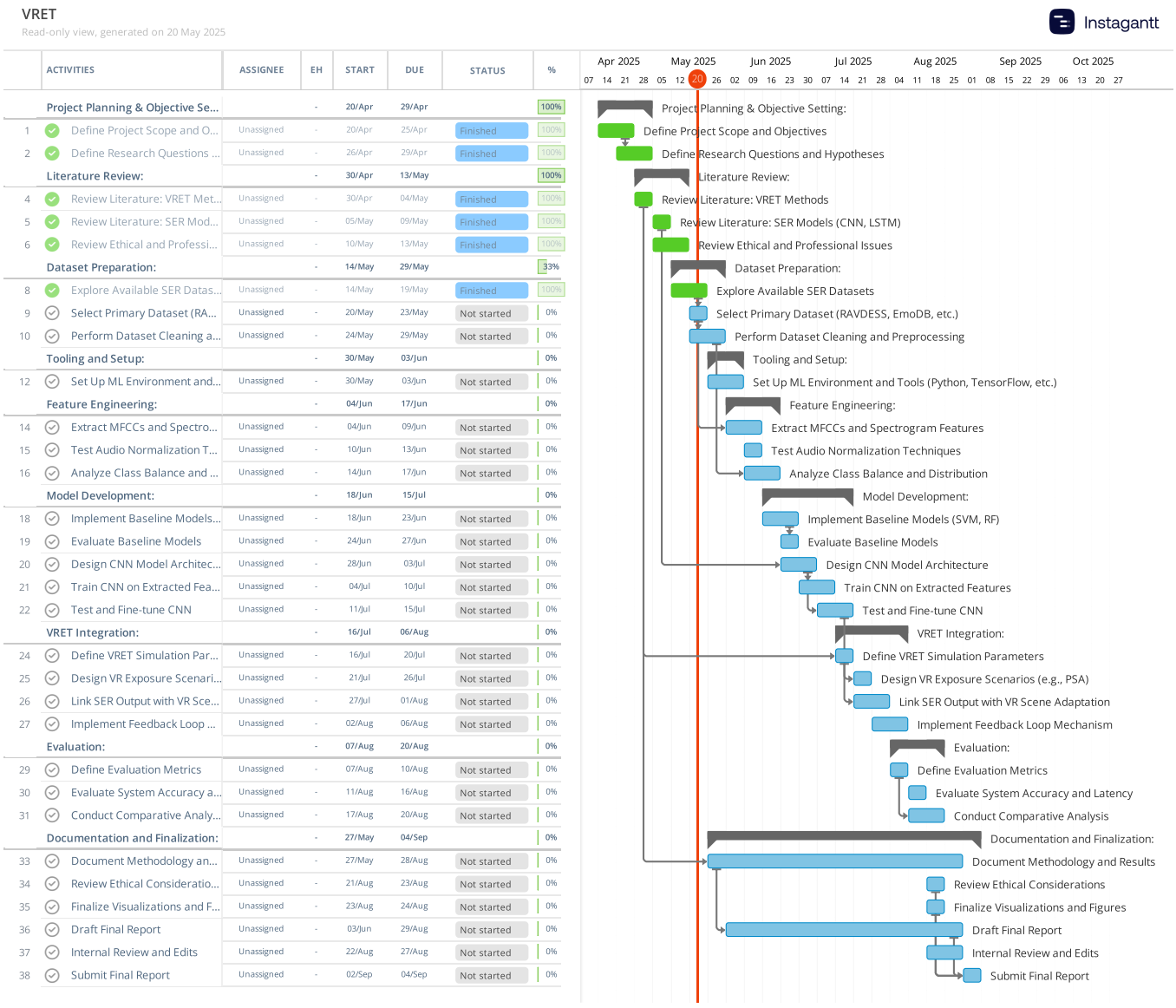


Figure 1.6: Gantt chart illustrating the detailed timeline, tasks, and dependencies for the Speech Emotion Recognition in VRET project.

Chapter 2

Background and Related Work

2.1 Background

Speech Emotion Recognition (SER) is a rapidly growing subfield of affective computing that focuses on identifying and interpreting emotional states through speech signals. This area of research holds immense promise across a range of applications, including intelligent virtual assistants, mental health diagnostics, customer service automation, and, more recently, adaptive Virtual Reality Exposure Therapy (VRET) [Šalkevicius et al., 2019]. While numerous systems have been developed for emotion recognition using facial expressions, relatively few have focused on using speech input, as reported in the literature [Nwe et al., 2003] . Over the past decade, the landscape of SER has dramatically evolved, transitioning from traditional machine learning to advanced deep learning methods. Architectures such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and more recently, Transformer-based models have significantly enhanced emotion classification performance. Deep models are capable of learning hierarchical representations directly from raw audio or spectrograms, thereby reducing the need for extensive feature engineering. For instance, the work by [Lee and Tashev, 2015] on using RNNs for high-level feature representation in SER is considered a seminal deep learning contribution in this domain. Despite architectural advancements, the success of these models hinges largely on the

quality, diversity, and representativeness of the datasets used for training and evaluation. Existing emotional speech corpora such as EMO-DB, RAVDESS, IEMOCAP, TESS, and SAVEE offer a range of acted, elicited, and natural emotional recordings across different languages and speaker demographics. Each dataset comes with its unique characteristics and limitations. For instance, EMO-DB is well-structured and clean but lacks diversity in gender and language, whereas IEMOCAP provides multimodal data but is limited in scale. One of the major criticisms in SER literature is the over-reliance on small, homogeneous datasets, which restricts the generalizability of the models in real-world scenarios [Abbaschian et al., 2021]. In light of limited and often imbalanced data, the role of data augmentation and synthetic data generation has gained considerable traction. Techniques such as pitch shifting, time stretching, noise injection, and SpecAugment have been effective in enhancing dataset variability. More advanced approaches, like using Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), aim to create entirely synthetic speech samples that mimic real emotional expressions. Research by [Sahu et al., 2018a] explored the use of GANs to generate synthetic feature vectors for SER. The study found that incorporating GAN-generated data into training improved the performance of emotion classifiers. A study by [Barhoumi and BenAyed, 2024] demonstrated that applying noise addition and spectrogram shifting techniques enhanced the robustness of SER models across multiple datasets, including TESS, EmoDB, and RAVDESS.

Table 2.1: Planned Literature Focus Areas and Review Objectives

Literature Area	**Focus Level**	**Key Parameters Reviewed**	**Purpose of Review**
Speech Emotion Recognition (SER)	High	Datasets, speech features, SER models (classical and DL-based), emotion categories, evaluation metrics	To establish a foundation and identify trends in emotion recognition via speech
Machine Learning and Deep Learning Techniques for SER	High	CNN, RNN, LSTM, Attention, Transformers, hybrid architectures	To explore which models are most effective for speech-based emotion classification
Feature Engineering in SER	Medium	MFCCs, spectrograms, chroma features, voice quality indicators	To understand best practices and discover novel or underused emotional features
Preprocessing Techniques	Medium	Noise removal, normalization, silence trimming, VAD techniques	To tailor preprocessing pipeline for VRET-based speech input
Data Augmentation and Synthetic Data in SER	Medium	GANs, VAEs, SpecAugment, oversampling strategies	To explore how synthetic data affects model generalization and class balance
Multilingual and Cross-Cultural SER	Low-Medium	Language-dependent vs. independent features, dataset diversity, cross-lingual transfer	To assess model adaptability across different languages and cultures
Ethics in Emotion AI	Low	Fairness, privacy, user consent, bias mitigation frameworks	To build an ethically grounded SER framework for mental health applications
SER in Virtual Reality or Therapeutic Contexts	Medium	Integration with VRET, therapy-specific emotions, real-time feedback systems	To identify prior use cases and inform system design for mental health applications

Table 2.2: Motivation and Research Contributions

Area of Focus	**Why It Matters**	**How This Project Contributes**
Dataset Diversity	Existing datasets are often mono-lingual, acted, or limited in scale, restricting generalization	Incorporates data augmentation and synthetic data to enhance emotional and speaker diversity
Model Architecture	Basic CNNs or RNNs may fail to fully capture emotion patterns in speech	Develops a hybrid CNN-LSTM model with attention for improved emotional representation
Therapeutic Application	SER is underutilized in therapeutic contexts such as anxiety management	Implements emotion recognition in a Virtual Reality Exposure Therapy (VRET) setting
Emotion Coverage	Most systems focus only on basic emotions like anger, joy, or sadness	Expands to include more nuanced emotions relevant to therapy (e.g., nervousness, calmness)
Real-Time Processing	Many SER models are not optimized for real-time or interactive use	Builds a lightweight, responsive model suitable for integration in real-time VR environments
Ethical Design	SER models often lack fairness, bias handling, and ethical safeguards	Emphasizes ethical principles: fairness, transparency, informed consent, and data privacy

Table 2.3: Scope and Focus Criteria for SER Literature Review

Criteria	**Inclusion Focus**	**Not Prioritized**
Application Relevance	Studies related to human-computer interaction, virtual reality, mental health, or real-time applications	Generic emotion recognition not intended for adaptive or real-time scenarios
Model Complexity	Research using ML/DL models such as CNN, LSTM, Transformers, or hybrid approaches	Studies using only shallow or traditional statistical models
Data Characteristics	Multilingual, multimodal, or augmented datasets with labeled emotional speech	Single-language, small-scale datasets with acted-only emotions
Evaluation Criteria	Papers reporting accuracy, UAR, latency, and confusion matrix analysis	Studies using only accuracy with no breakdown of emotion-wise performance
Ethical Transparency	Works discussing bias, fairness, privacy, and informed consent in SER systems	Studies lacking attention to ethical or societal implications

The criteria outlined across Tables 2.1, 2.2, and 2.3 collectively establish a structured and targeted approach for shaping the scope, direction, and impact of this research. Table 2.1 ensures that the literature review prioritizes high-quality, ethically sound, and application-relevant studies. Table 2.2 delineates the depth of exploration in key thematic areas such as model development, data strategies, and domain-specific SER applications ensuring focused and comprehensive coverage. Meanwhile, Table 2.3 highlights forward-looking research avenues that this project aims to investigate, laying the groundwork for both methodological innovation and practical relevance. Together, these criteria serve to justify the relevance, feasibility, and originality of the proposed work within the broader landscape of Speech Emotion Recognition research.

2.2 Related Work

The field of Speech Emotion Recognition (SER) has witnessed significant evolution over the past two decades, transitioning from handcrafted feature-based models to sophisticated deep learning architectures capable of extracting nuanced affective cues from human speech [Schuller, 2018]. As interest in emotionally intelligent systems continues to grow, an increasing body of research has emerged exploring various aspects of SER, including feature extraction techniques, classification algorithms, data augmentation strategies, and application-specific frameworks. This section reviews key developments in the SER literature, focusing on contributions that have informed the use of machine learning and deep learning approaches, the design and evaluation of emotional speech datasets, and the integration of SER systems in real-time and therapeutic contexts such as Virtual Reality Exposure Therapy (VRET). By mapping out this landscape, the following review aims to highlight both foundational work and emerging trends that shape the direction of this project.

2.2.1 Overview of Speech Emotion Recognition (SER)

Speech Emotion Recognition (SER) is a critical area within affective computing, focusing on the identification and interpretation of human emotions through speech signals. Its significance spans various domains, including human-computer interaction, virtual assistants, mental health diagnostics, and customer service automation.

One of the earliest efforts in this domain can be traced to the work of J.D. Williamson, who in 1978 patented a speech analyzer system designed to detect pitch and frequency perturbations in an individual's voice to determine their emotional state [Williamson, 1978]. This innovation laid foundational groundwork by emphasizing the potential of vocal signals as indicators of psychological states.

Building on these insights, the pioneering study *Recognizing Emotion in Speech* (1996) explored the intersection of spoken language and emotional expression, proposing methods to detect affective states using acoustic analysis and language models [Dellaert et al., 1996]. This early work underscored the feasibility of emotion-sensitive systems well before the widespread adoption of machine learning.

In a significant leap forward, Trigeorgis et al. [Trigeorgis et al., 2016] introduced an end-to-end deep learning architecture that combines Convolutional Neural Networks (CNNs) with Long Short-Term Memory (LSTM) networks. This approach allows the model to learn emotional representations directly from raw audio waveforms, bypassing the need for handcrafted features. When evaluated on the RECOLA database, the model significantly outperformed traditional signal-processing-based SER methods in predicting spontaneous and natural emotions.

More recently, comprehensive survey studies have highlighted both the advancements and ongoing challenges in SER, including the need for models capable of handling noisy, diverse, and real-world data environments [George and Ilyas, 2024]. These studies reflect the field's continued momentum toward more robust, context-aware, and generalizable emotion recognition systems.

2.2.2 Traditional Machine Learning Approaches in SER

Traditional machine learning techniques have long served as the foundation for early speech emotion recognition (SER) systems. A wide variety of classifiers have been employed in this domain, including Support Vector Machines (SVM), Hidden Markov Models (HMM), Gaussian Mixture Models (GMM), k-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), decision trees, and fuzzy logic classifiers. Each of these models offers distinct advantages and challenges depending on the data distribution, dimensionality, and feature type [Fahad et al., 2021b].

Support Vector Machines (SVMs) became one of the most widely used classifiers [Fahad et al., 2021b; Lee et al., 2011; Stasiak and Rychlicki-Kicior, 2012; Wu et al., 2011] due to their ability to perform well in high-dimensional spaces and their robustness against overfitting. SVMs map features into a higher-dimensional space using kernel functions, making them effective at separating complex emotional patterns. One of their key advantages is that they can perform reliably even without extensive feature selection, making them suitable for early SER pipelines.

Hidden Markov Models (HMMs) and Gaussian Mixture Models (GMMs) were also frequently employed, especially for modeling the temporal and probabilistic structure of speech. HMMs are well-suited for time-series data and were among the first models to account for the dynamic flow of emotions over time [Fahad et al., 2021a]. GMMs, although simpler, modeled the distribution of features for different emotion classes, offering a statistical perspective on emotion classification.

Other traditional classifiers such as k-Nearest Neighbors (k-NN), Decision Trees, and Naïve Bayes were explored for their simplicity and interpretability. However, these models were generally more sensitive to the high dimensionality and noise often present in real-world speech data [Fahad et al., 2021b]. As such, they were typically used in constrained settings or as part of ensemble systems.

A notable area of development within traditional SER was the use of ensemble methods and hierarchical classification [Fahad et al., 2021b; Wu and Liang, 2010]. Ensemble classifiers combined the strengths of multiple base models to reduce prediction variance

and improve robustness, especially when dealing with spontaneous or natural speech. Hierarchical classifiers, structured as decision trees, allowed broader emotion categories to be broken down into finer subcategories, mirroring the way humans perceive and differentiate emotions.

Despite being gradually replaced by deep learning architectures, traditional machine learning approaches remain relevant, particularly in applications where computational resources are limited or where model transparency is essential. They also serve as valuable benchmarks and can be effectively integrated into hybrid systems that combine classical and modern techniques [Fahad et al., 2021b].

2.2.3 Deep Learning Techniques in SER

The emergence of deep learning has led to substantial advancements in Speech Emotion Recognition (SER), enabling the automatic extraction of complex, high-level emotional features directly from raw audio or spectrogram representations. Unlike traditional machine learning methods, deep learning models can learn hierarchical representations, allowing them to capture intricate emotional patterns that are often lost in handcrafted features [LeCun et al., 2015]. Convolutional Neural Networks (CNNs) have been extensively used in SER due to their capability to extract spatial features from spectrograms. These models excel at identifying local patterns such as shifts in pitch, energy, and spectral content—crucial components of emotional speech. Researchers have shown that CNNs, when fed with Mel-spectrograms or log-frequency representations, significantly outperform traditional classifiers in tasks involving emotion classification [Fayek et al., 2017].

Hybrid architectures combining CNNs and LSTMs have gained prominence in SER research. In these models, CNNs first extract spatial features, which are then passed to LSTMs to model temporal dependencies. This fusion of spatial and temporal modeling has been proven to deliver higher classification accuracy across benchmark datasets such as IEMOCAP and RAVDESS [Zhao et al., 2019].

Beyond these, Deep Neural Networks (DNNs) have been applied in SER to map complex

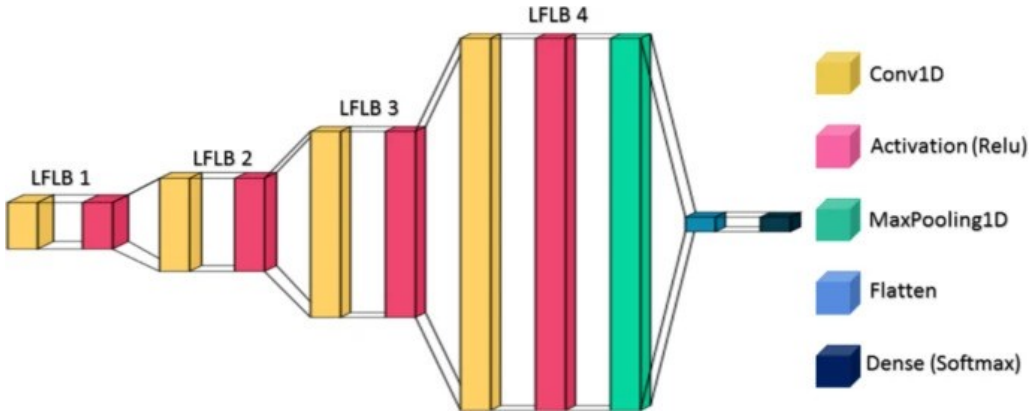


Figure 2.1: CNN model Architecture.

Source: Adapted from [Barhoumi and BenAyed, 2024]

feature representations to emotional labels. For instance, early studies utilized DNNs to perform generalized discriminant analysis, outperforming traditional models such as SVMs [Stuhlsatz et al., 2011]. Hybrid DNN-HMM architectures have also shown improvement over GMM-HMM models when trained on MFCCs, leveraging the strengths of both deep learning and temporal modeling [Li et al., 2013].

The use of soft labels to capture the subjectivity of emotion perception has also been explored. Rather than using hard labels derived from majority voting, models have been trained on probabilistic emotion labels aggregated from multiple annotators, resulting in better generalization [Fayek et al., 2016].

Efforts to reduce latency and simplify SER pipelines have led to frame-based SER approaches using DNNs, where intra-utterance emotional dynamics are captured at the frame level [Han et al., 2014].Such models prioritize efficiency while maintaining high performance, relying less on high-level handcrafted features and more on automatic deep feature learning.

The integration of deep learning into SER has thus marked a pivotal shift in the field, making it possible to model complex emotional expressions with higher accuracy and adaptability. While challenges remain—particularly with respect to dataset size, model interpretability, and real-time application—deep architectures continue to set the benchmark for modern SER systems.

2.2.4 Feature Extraction and Representation

The effectiveness of any Speech Emotion Recognition (SER) system depends heavily on the quality of the features extracted from speech signals. Early SER studies relied on handcrafted features based on knowledge of human speech production and perception. Commonly used features included Mel-Frequency Cepstral Coefficients (MFCCs), pitch, energy, formants, jitter, and shimmer [Eyben et al., 2010]. These low-level descriptors (LLDs) were often paired with statistical functionals to generate fixed-length representations from variable-length utterances [Schuller et al., 2011]. MFCCs, in particular, have remained a staple in SER due to their ability to approximate the human auditory system's response. However, they primarily capture the spectral envelope and may not encode all relevant paralinguistic information [Ververidis and Kotropoulos, 2006]. To address this, researchers have explored other spectral features such as chroma, spectral flux, spectral centroid, and harmonics-to-noise ratio [Akçay and Oğuz, 2020]. In recent years, representations derived from time-frequency analyses such as spectrograms, Mel-spectrograms, and log-Mel filter banks have become the standard input for deep learning models. These representations preserve both temporal and frequency information, making them suitable for architectures like CNNs and RNNs [Badshah et al., 2017]. Mirsamadi et al. [Mirsamadi et al., 2017] showed that raw spectrograms, when used with attention mechanisms, outperformed many traditional handcrafted feature sets.

The openSMILE toolkit has been instrumental in the field by offering a reproducible and standardized way to extract a wide range of acoustic features [Eyben et al., 2010]. The INTERSPEECH Computational Paralinguistics Challenge (ComParE) feature set has also been widely adopted as a benchmark in emotion recognition tasks [Schuller et al., 2013].

Beyond static features, researchers have also begun to explore dynamic representations such as derivatives (delta and delta-delta coefficients) and utterance-level aggregation using attention pooling or statistics like mean and variance [Satt et al., 2017]. These help capture the evolving nature of emotions throughout a speech segment.

As SER models move toward end-to-end learning, there's also a growing interest in learn-

ing feature representations directly from raw waveforms. Models like wav2vec and TRILL extract embeddings from speech that encapsulate both linguistic and emotional content [Baevski et al., 2020; Shor et al., 2020]. These representations often outperform handcrafted features, especially when fine-tuned for emotion recognition tasks.

Overall, the field has shifted from manual feature engineering to data-driven feature learning, with an emphasis on maintaining emotionally relevant information across time and frequency domains. Feature representation remains a critical research area in achieving robust and generalizable SER systems.

2.2.5 Data Augmentation and Synthetic Data Generation

A critical challenge in Speech Emotion Recognition (SER) is the limited availability and imbalance of annotated emotional speech datasets. Emotions are inherently subjective and difficult to label consistently [Fahad et al., 2021b], and natural emotional expressions are less frequently recorded than acted emotions, resulting in skewed class distributions and limited training data. These issues can significantly hinder model generalization and accuracy, particularly in real-world applications where emotion variability is high.

To address this, data augmentation techniques have been widely employed. Simple augmentation strategies such as pitch shifting, time stretching, and adding background noise help simulate variability in speaker and recording conditions [Shankar et al., 2022]. These transformations allow SER models to learn from a broader distribution of input patterns, thereby improving robustness. SpecAugment, a technique originally proposed for automatic speech recognition, has also been adapted for SER. It applies time and frequency masking on spectrogram inputs, encouraging models to rely on more generalizable features [Park et al., 2019].

In addition to these signal-level augmentations, researchers have also investigated synthetic data generation using advanced generative models. Generative Adversarial Networks (GANs) have shown promise in producing artificial emotional speech samples that closely mimic the distribution of real data. For example, GAN-based frameworks have been used to generate feature vectors or raw waveform data for underrepresented emo-

tional classes, contributing to class balance and improved model fairness [Sahu et al., 2018b]. Variational Autoencoders (VAEs) have also been applied in SER for dimensionality reduction and synthetic data creation. Unlike GANs, which use a discriminator-generator framework, VAEs model the data distribution explicitly and are often more stable to train. Both GANs and VAEs enable augmentation without further collection of labeled speech, which is particularly useful for low-resource or multilingual SER scenarios [Latif et al., 2020].

Recent studies have explored combining augmentation techniques with emotion-aware embedding methods to generate high-quality synthetic data. For instance, adversarial autoencoders and emotion-conditioned VAEs can help control the style and content of generated speech to maintain emotional fidelity [Zhao and Yang, 2023]. Some frameworks use emotion vectors derived from real data to guide the synthesis process, thereby preserving class-specific information.

Despite these advancements, challenges remain in evaluating the quality and emotional consistency of synthetic samples. Metrics for assessing the realism and emotional clarity of generated data are still evolving, and the subjective nature of emotion poses validation difficulties. Moreover, synthetic augmentation should be used with caution, as excessive reliance on artificial data might introduce bias or overfitting if the generated samples do not accurately reflect real-world variability.

Nonetheless, data augmentation and synthetic data generation remain indispensable tools in SER, offering practical solutions to the constraints of limited labeled data. Their judicious use can enhance model generalization, improve class balance, and facilitate the development of SER systems that are robust across different speakers, languages, and emotional styles.

2.2.6 Multilingual and Cross-Cultural SER

Most existing Speech Emotion Recognition (SER) systems are trained on monolingual datasets, typically focused on English or other widely spoken languages. This narrow linguistic scope limits the generalizability of these models in global contexts. Emotional

expression is influenced by cultural and linguistic nuances, which may not be adequately captured by models trained on a single-language dataset [Ververidis and Kotropoulos, 2006].

To address this limitation, cross-lingual SER has emerged as a promising area of research. These models aim to transfer emotional knowledge learned from one language to another by leveraging shared acoustic or prosodic features. Language-independent features such as pitch contours, energy dynamics, and temporal variations have been shown to be robust across languages [Latif et al., 2023].

Transfer learning and domain adaptation techniques have also been explored to improve model performance in multilingual settings. For instance, pretraining models on high-resource languages and fine-tuning them on target low-resource languages has demonstrated improved emotion classification accuracy Valiyavalappil Haridas et al. [2022]. Additionally, Heracleous and Yoneyama (2019) proposed a two-pass classification system that integrates language identification and emotion recognition to enhance SER performance in bilingual and multilingual settings [Heracleous and Yoneyama, 2019].

While challenges remain, including limited multilingual emotion datasets and variable annotation standards, multilingual SER is crucial for creating inclusive and widely deployable systems.

2.2.7 Real-Time SER Systems

With the increasing integration of SER into conversational agents, smart assistants, and monitoring tools, the demand for real-time, low-latency emotion recognition systems has intensified. Traditional SER systems, which often involve complex preprocessing and feature extraction steps, are not always suitable for real-time deployment.

Recent efforts have focused on reducing computational overhead while maintaining classification accuracy. Model optimization techniques such as quantization, pruning, and knowledge distillation have been used to create lightweight SER architectures [Kim et al., 2021]. These models can run efficiently on edge devices with limited computational power. Bertero et al. (2016) developed a real-time SER module integrated into an interactive di-

alogue system, emphasizing fast response times and context-awareness for improving user interaction [Bertero et al., 2016]. Frame-level emotion prediction and buffering strategies have also been explored to reduce inference delays without sacrificing performance [Chang et al., 2020].

Ensuring real-time capability often requires a trade-off between model complexity and responsiveness. Therefore, system-level optimization, including efficient feature pipelines and concurrent inference mechanisms, plays a crucial role in enabling real-time SER applications.

2.2.8 SER in Therapeutic and Virtual Reality Applications

The use of SER in therapeutic contexts, particularly in Virtual Reality Exposure Therapy (VRET), represents a growing interdisciplinary field where affective computing meets digital mental health. VRET provides a controlled environment for exposing individuals to anxiety-provoking stimuli, such as public speaking scenarios, in a safe and adaptive manner [Parsons and Rizzo, 2008].

Incorporating SER into these systems allows for dynamic assessment of a participant's emotional state, enabling real-time adjustments to the therapy based on detected stress or anxiety levels [Morales and Levitan, 2016]. For instance, virtual environments can become more or less challenging based on emotional cues, enhancing the personalization and efficacy of the therapy.

However, deploying SER in VR presents unique challenges. These include the need for real-time processing, handling of speech distortion due to VR headset microphones, and integration with other biofeedback signals such as heart rate and skin conductance [Pan et al., 2020]. Katirai (2024) highlights the importance of ethical safeguards when using emotion recognition in clinical settings, emphasizing transparency, user consent, and the potential risks of misclassification [Katirai, 2024].

Despite these challenges, the integration of SER in VR-based therapy holds significant promise for improving mental health outcomes through personalized and adaptive interventions.

Comparative note. Closest to our setting, Amey et al. frame speech-driven VRET as a binary arousal detection problem using engineered acoustic features with a lightweight Conv1D classifier; in contrast, we adopt self-supervised speech representations (e.g., Wav2Vec2) and retain a multi-class label space, prioritising robustness under greater cross-corpus diversity and enabling richer VR feedback [Amey et al., 2025].

2.2.9 Ethical Considerations in SER

As SER technologies gain traction across domains such as healthcare, education, and customer service, ethical considerations surrounding their development and deployment have become increasingly critical. Issues such as privacy, informed consent, algorithmic bias, and emotional manipulation must be addressed to ensure the responsible use of these systems [McStay, 2020].

Emotion data, often derived from speech, is sensitive and potentially revealing. Improper handling or unauthorized access to such data could lead to breaches of privacy and trust. Hence, transparent data governance and strict data protection protocols are essential [Mohammad, 2022].

Moreover, bias in emotion recognition systems—whether due to imbalanced training data or culturally specific expression patterns—can lead to systematic misclassification. This is especially problematic in high-stakes settings such as mental health diagnostics or legal contexts. Fairness-aware algorithms and diverse training datasets are necessary to mitigate these issues [Katirai, 2024].

Additionally, the opaque nature of many deep learning models presents a barrier to accountability. Explainable AI (XAI) methods are being explored to enhance transparency and user trust in SER systems [Stahl et al., 2023].

Overall, integrating ethical principles into the design and deployment of SER technologies is not only a moral imperative but also vital for public acceptance and long-term success of emotion-aware systems.

2.3 Discussion

The literature explored in this review reveals a clear and compelling progression in the field of Speech Emotion Recognition (SER), from early statistical models to deep learning-based architectures. This shift underscores the growing importance of robust, scalable, and real-time emotion recognition in various domains, including human-computer interaction, virtual therapy, multilingual communication, and mobile applications. The increasing use of advanced techniques such as CNN-LSTM hybrids, attention mechanisms, and transformer-based models signifies an evolution toward more context-aware and data-efficient SER systems.

Beyond the technical advancements, this review also highlights a broader effort to address challenges related to data scarcity, bias, ethical responsibility, and cross-cultural variability. Augmentation strategies, transfer learning, and synthetic data generation have enabled researchers to overcome data limitations, while ethical frameworks are being integrated to ensure responsible deployment of emotion recognition technologies.

Figure 2.2 displays the temporal distribution of the literature included in this study. A steady increase in publications is observed from 2016 onwards, with a notable surge between 2018 and 2020, reflecting growing research interest and technological feasibility. This pattern illustrates how advances in deep learning and the rising application of SER in real-world systems have spurred academic attention.

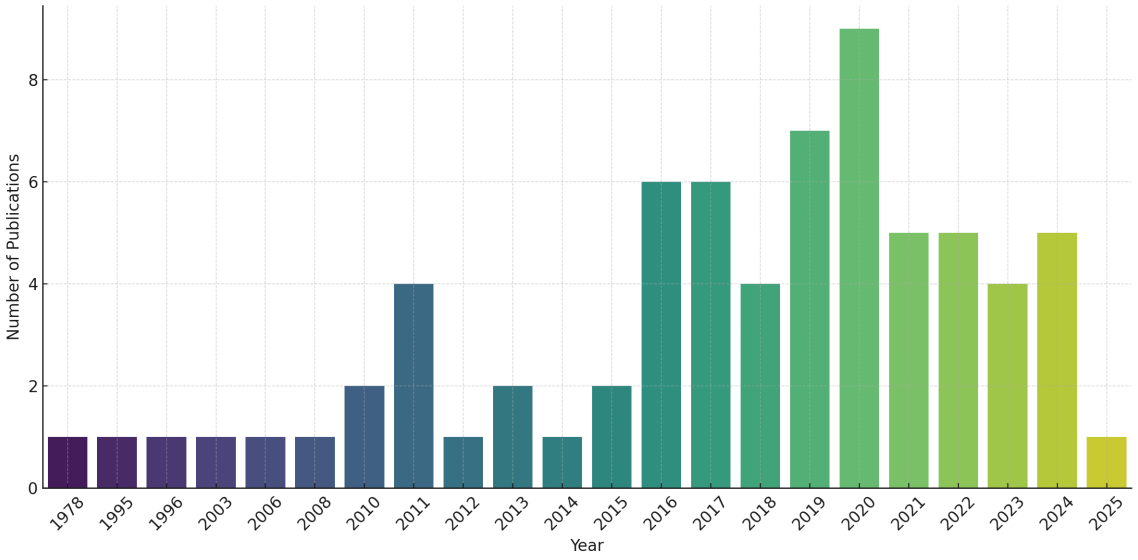


Figure 2.2: Reviewed Publications by Year.

Chapter 3

System Design

This chapter outlines the design of the Speech Emotion Recognition (SER) system developed for integration into a Virtual Reality Exposure Therapy (VRET) framework. The project design focuses on evaluating multiple machine learning and deep learning models, selecting the most promising architecture, and progressively refining it through transfer learning and dataset augmentation. The approach follows an iterative pipeline, moving from model exploration to domain-specific fine-tuning and generalisation across diverse datasets.

3.1 Design Objectives

The primary design objectives of this project were:

- Evaluate the performance of multiple baseline and advanced models for SER.
- Identify the model architecture capable of achieving high accuracy with minimal overfitting.
- Leverage transfer learning by fine-tuning a pre-trained model across datasets.
- Enhance model generalisation using a merged multi-dataset corpus and audio augmentation techniques.
- Create a reusable training pipeline for future dataset integration.

3.1.1 Hardware and Software Design Considerations

All experiments were executed on a workstation equipped with **dual NVIDIA T400 GPUs (T400×2)**. This configuration offered sufficient parallelism for batch fine-tuning of transformer-based speech encoders while maintaining low power and cost footprints. The software stack comprised `Python`, `PyTorch` [Paszke et al., 2019], and `HuggingFace Transformers` [Wolf et al., 2020], orchestrated via Jupyter notebooks for reproducibility.

Rationale. The T400×2 setup was selected to (i) ensure dependable GPU availability for iterative fine-tuning, (ii) keep inference latency within sub-500 ms targets required for interactive VRET scenarios [Nielsen, 1993], and (iii) support mixed-precision training [Micikevicius et al., 2017] where beneficial . Cloud notebooks (e.g., Colab) were used as a fallback to mirror the environment for reproducibility.

3.2 Model Selection Strategy

We adopted a staged screening process to identify the most effective architecture for Speech Emotion Recognition (SER) before large-scale transfer learning. First, we trained a set of *feature-based* baselines—SVM, MLP, CNN, and a soft-voting ensemble—using MFCC/log-Mel representations computed from 16 kHz mono audio. Each baseline was evaluated under a speaker-independent split with stratification by class, reporting accuracy and macro-F1 on the validation set. Next, we assessed a *representation-learning* approach with Wav2Vec2, which ingests raw waveforms and is fine-tuned end-to-end. Selection was based on validation macro-F1 (tie-break by accuracy) and stability across seeds. The winning model on TESS (Wav2Vec2, 99.8% accuracy) was then fine-tuned on RAVDESS and, finally, extended to the merged NOR corpus with on-the-fly augmentation and class-weighted loss.

1. **Support Vector Machine (SVM):** [Cortes and Vapnik, 1995] classical margin-based baseline on MFCC statistics.

2. **Convolutional Neural Network (CNN):** [Goodfellow et al., 2016] spectral fea-

ture learner on log-Mel spectrograms.

3. **Multi-Layer Perceptron (MLP):** [Bishop and Nasrabadi, 2006] fully connected network on statistical features.

4. **Voting Classifier (Ensemble):** [Kuncheva, 2014] soft vote of Logistic Regression, Random Forest, and MLP.

5. **Wav2Vec2:** [Baevski et al., 2020] transformer model pre-trained on speech, fine-tuned for SER.

Notation (symbols)

$\mathbf{x} \in \mathbb{R}^d$	Input feature vector (or raw waveform slice for Wav2Vec2)
$y \in \{1, \ldots, K\}$	Class label; K = number of emotion classes
$y_{ik} \in \{0, 1\}$	One-hot indicator that sample i belongs to class k
$\hat{\mathbf{y}}$	Predicted class probabilities (softmax)
$\mathbf{w}$, b	Linear weights and bias; $\mathbf{W}, \mathbf{b}$ denote layer parameters
$\phi(\cdot)$	Feature map; $k(\cdot, \cdot)$ kernel function (e.g., RBF with γ)
C, ξ_i	SVM soft-margin constant and slack variables
X	2D input (e.g., spectrogram); K (in conv) denotes kernel/filter; b bias
H, P	Activation map and pooled map; $\sigma(\cdot)$ activation (ReLU)
$\mathbf{z}$	Logits before softmax; p_{ik} predicted prob. for sample i, class k
$\mathbf{x}_{1:T}$	Waveform samples; $\mathbf{z}_{1:L}$ encoder latents; $\mathbf{c}_{1:L}$ contextual reps
$\mathbf{h}$	Utterance-level pooled embedding for classification
α_k	Class weight for class k in weighted cross-entropy
M	Number of base models in the voting ensemble; w_m = model weight

Support Vector Machine (SVM)

Decision function

$$f(\mathbf{x}) = \operatorname{sign}\big(\mathbf{w}^\top \phi(\mathbf{x}) + b\big). \tag{3.1}$$

Source: [Cortes and Vapnik, 1995].

Soft–margin primal objective

$$\min_{\mathbf{w},\, b,\, \boldsymbol{\xi}} \ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{s.t.} \quad y_i\big(\mathbf{w}^\top \phi(\mathbf{x}_i) + b\big) \geq 1 - \xi_i, \quad \xi_i \geq 0. \tag{3.2}$$

- ϕ is the (implicit) feature map with RBF kernel $k(\mathbf{x}, \mathbf{z}) = \exp(-\gamma\|\mathbf{x} - \mathbf{z}\|^2)$;

where

- $C > 0$ controls margin–violation trade-off; ξ_i are slack variables;
- multiclass via one-vs-rest: $\hat{y} = \arg\max_c f_c(\mathbf{x})$.

Params used: RBF kernel, $C = 10$, $\gamma =$ `scale` on MFCC statistics.

Convolutional Neural Network (CNN)

Convolution

$$Y(i,j) = (X * K)(i,j) = \sum_{m}\sum_{n} X(i{+}m,\ j{+}n)\, K(m,n) + b. \tag{3.3}$$

Activation & pooling

$$H(i,j) = \sigma\big(Y(i,j)\big), \qquad P(p,q) = \max_{(i,j)\in\mathcal{W}_{pq}} H(i,j). \tag{3.4}$$

Softmax cross–entropy

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\, \log \frac{\exp(z_{ik})}{\sum_{j=1}^{K}\exp(z_{ij})}. \tag{3.5}$$

Source: [Goodfellow et al., 2016].

- X is the log-Mel spectrogram input; K is a learned convolutional filter; b is bias;

where

- $\sigma(\cdot)$ is ReLU; $\mathcal{W}_{pq}$ is the pooling window at output cell (p, q);
- z_{ik} are logits; p_{ik} are post-softmax probabilities.

Architecture used (summary): 5 conv blocks (32→64→128→256→128 filters; 3×3/5×5), BN+ReLU, max-pool, dropout 0.5; Dense(64) → Dense(K). Adam 1×10^{-3}, batch 32.

Multi-Layer Perceptron (MLP)

Forward pass

$$\mathbf{h}_1 = \sigma(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \quad \mathbf{h}_2 = \sigma(\mathbf{W}_2\mathbf{h}_1 + \mathbf{b}_2), \quad \mathbf{z} = \mathbf{W}_3\mathbf{h}_2 + \mathbf{b}_3, \quad \hat{\mathbf{y}} = \text{softmax}(\mathbf{z}). \tag{3.6}$$

- $\mathbf{x}$ are MFCC statistics; $\mathbf{W}_\ell, \mathbf{b}_\ell$ are layer parameters;

where

- σ is ReLU; $\hat{\mathbf{y}}$ is the class-posterior vector.

Params used: hidden sizes [256, 128], ReLU, dropout 0.5, Adam 1×10^{-3}, 30 epochs.

Voting Classifier (Ensemble)

Soft voting

$$\hat{y} = \arg\max_k \sum_{m=1}^{M} w_m \, p_m\big(y{=}k \mid \mathbf{x}\big), \qquad \sum_m w_m = 1. \tag{3.7}$$

- $p_m(y{=}k \mid \mathbf{x})$ is the posterior from base model m (LR, RF, MLP);

where

- w_m are the model weights (equal weights used).

Wav2Vec2

Fine-tuning head

$$\mathbf{x}_{1:T} \xrightarrow{\text{conv encoder}} \mathbf{z}_{1:L} \xrightarrow{\text{Transformer}} \mathbf{c}_{1:L} \xrightarrow{\text{Pool}} \mathbf{h} \in \mathbb{R}^d, \quad \mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}, \quad \hat{\mathbf{y}} = \text{softmax}(\mathbf{z}). \tag{3.8}$$

Source: [Baevski et al., 2020].

Class-weighted cross-entropy

$$\mathcal{L}_{\text{WCE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\alpha_k\, y_{ik}\, \log p_{ik}. \tag{3.9}$$

where

- $\mathbf{x}_{1:T}$: raw waveform; $\mathbf{z}_{1:L}$: latent; $\mathbf{c}_{1:L}$: contextual representation;
- $\mathbf{h}$: pooled utterance embedding; α_k: class weight for imbalance.

Params used: `facebook/wav2vec2-base` (16 kHz), AdamW 2×10^{-5}, batch 8, 6–8 epochs, best by validation F1; TESS → RAVDESS → NOR with on-the-fly augmentation.

3.2.1 Wav2Vec 2.0 as a Self-Supervised Acoustic Encoder

Self-supervised learning (SSL) for speech learns general-purpose acoustic representations from raw audio without manual labels. Among SSL encoders, **Wav2Vec 2.0** [Baevski et al., 2020] has proven especially effective because it (i) operates directly on *waveforms* at 16 kHz, (ii) captures both short-term spectral cues and longer-range prosody with a Transformer context network, and (iii) requires only a light task-specific head for downstream fine-tuning.

Architecture. Wav2Vec 2.0 consists of: (1) a *convolutional feature encoder* that maps raw waveform $\mathbf{x}$ to latent frames $\mathbf{z}_t$; (2) a *context network* (Transformer) that aggregates information across time to produce contextualised states $\mathbf{c}_t$; and (3) a *discrete target quantiser* that maps latent frames to codebook entries $\mathbf{q}_t$ used as SSL targets [Baevski et al., 2020].

Pre-training objective (no labels). During SSL pre-training, time steps are *masked* and the model solves a contrastive prediction task: given context $\mathbf{c}_t$, it must identify the true quantised target $\mathbf{q}_t$ among K distractors $\tilde{\mathbf{q}} \in \mathcal{N}_t$. With cosine similarity $s(\cdot,\cdot)$ and temperature κ, the InfoNCE-style loss [Oord et al., 2018] is:

$$\mathcal{L}_{\text{ctr}} \;=\; -\sum_{t\in\mathcal{M}} \log \frac{\exp\big(s(\mathbf{c}_t,\mathbf{q}_t)/\kappa\big)}{\sum_{\tilde{\mathbf{q}}\in\{\mathbf{q}_t\}\cup\mathcal{N}_t}\exp\big(s(\mathbf{c}_t,\tilde{\mathbf{q}})/\kappa\big)}\,.$$

A *codebook diversity* regulariser encourages uniform usage of quantiser entries, preventing collapse and enriching the learned units [Baevski et al., 2020]. The result is a representation space that organises phonetic content, timbre, and prosodic patterns without labels.

Fine-tuning for SER. For supervised emotion recognition, we attach a small *classification head* (linear or MLP) on top of the contextual states and train end-to-end with cross-entropy. In our pipeline, we use (i) *speaker-disjoint, stratified* splits; (ii) *class-weighted* loss to mitigate imbalance; and (iii) *on-the-fly waveform augmentations* (additive noise, pitch/time perturbations, time shift, band-pass) applied *only* during training. We fine-tune on TESS and RAVDESS before expanding to the heterogeneous NOR corpus, keeping audio at 16 kHz mono to match the encoder's pre-training interface.

Why SSL helps SER. Emotion is conveyed not only by spectral envelopes but also by global prosody (pitch dynamics, intensity, rhythm) and subtle temporal cues [Liebenthal et al., 2016]. Because Wav2Vec 2.0 pre-training exposes the model to large amounts of raw speech, the contextualised states encode robust, speaker- and channel-tolerant features that transfer well with limited labelled data. Empirically, this yields near-ceiling accuracy on acted corpora (TESS, RAVDESS) and strong generalisation to mixed, noisy conditions (NOR), while keeping the task-specific head compact and efficient for streaming inference.

Decision and deployment. At inference, we process overlapping windows (2–3 s; 50% hop), obtain per-class probabilities via the fine-tuned head, and aggregate across windows (probability averaging or majority voting). This windowed scheme, combined with optional temporal smoothing (EMA/hysteresis), produces stable, low-latency signals suitable for VRET scene adaptation.

3.3 Stage 1 – Baseline Evaluation on TESS Dataset

To establish a performance benchmark, all candidate models were trained and evaluated on the Toronto Emotional Speech Set (TESS) [Dupuis and Pichora-Fuller, 2010].

Results indicated that the Wav2Vec2 model achieved **99.8%** accuracy, significantly outperforming other models. Given its ability to leverage raw waveform inputs and learn context-rich embeddings, Wav2Vec2 was selected as the primary architecture for further experimentation.

3.4 Stage 2 – Transfer Learning to RAVDESS Dataset

The TESS-trained Wav2Vec2 model was then fine-tuned on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [Livingstone and Russo, 2018a] dataset. This transfer learning step aimed to adapt the learned feature representations to a dataset with greater speaker and recording variability. Fine-tuning for 6 epochs yielded an accuracy of **97.6%**, confirming the model's adaptability.

Table 3.1: Sequential model performance across datasets during selection and transfer learning. Cells are colour-coded by accuracy.

Model	**Dataset**	**Accuracy (%)**
SVM (RBF, MFCC)	TESS	96.0
CNN (Mel-spectrogram, 30 ep)	RAVDESS	62.15
CNN Deep (Mel-spectrogram)	RAVDESS	92.5
Ensemble Voting (LR+RF+MLP)	RAVDESS	93.4
Wav2Vec2 (base)	TESS	99.8
Wav2Vec2 (6 ep)	RAVDESS	97.6

3.5 Stage 3 – Expansion to NOR Dataset

To evaluate the generalisation capability of the selected model, training was extended to the merged **NOR dataset**, which integrates multiple well-established emotion recognition corpora. The objective was to create a more diverse and realistic training resource capable of improving robustness in real-world conditions. The combined dataset consisted of 19,487 audio samples covering six emotion classes: *angry, disgust, fear, happy, neutral, sad*. The datasets included are:

- **ASVP-ESD** [Landry et al., 2020] – The ASVP-ESD dataset uniquely contains both *speech* and *non-speech utterances*, making it well-suited for building models that remain reliable in everyday, noisy conditions. The recordings capture emotions sourced from online platforms, films, and real conversational exchanges, thereby reflecting authentic and diverse vocal expressions. Its naturalistic variability was a strategic choice for enhancing generalisation.

- **TESS** [Dupuis and Pichora-Fuller, 2010] – The Toronto Emotional Speech Set contains utterances from two female speakers aged 26 and 64, providing age-based vocal variability. Each utterance is recorded with clear articulation and labelled across seven emotion categories. Its controlled environment recordings make it an ideal high-quality reference point for model calibration, ensuring the system learns accurate emotion boundaries before being exposed to noisier data.

- **RAVDESS** [Livingstone and Russo, 2018b] – The Ryerson Audio-Visual Database of Emotional Speech and Song offers both speech and song recordings from 24 professional actors (12 male, 12 female). Each emotion is expressed at two intensity levels, making it a valuable resource for training models to detect subtle differences in emotional intensity. Its professional production and balanced gender representation improve the consistency of emotion classification.

- **CREMA-D** [Cao et al., 2014] – The Crowd-Sourced Emotional Multimodal Actors Dataset consists of audio-visual recordings from 91 actors with a wide range of ethnicities and accents. Speech is recorded in different emotional tones with varying sentences, enabling the model to handle speaker diversity. This dataset addresses accent and prosody variations, a key factor in building cross-speaker robust systems.

- **SAVEE** [Jackson and Haq, 2014] – The Surrey Audio-Visual Expressed Emotion dataset comprises recordings from four male speakers, all native British English speakers. It provides high-quality audio in controlled environments, covering seven emotions. SAVEE is especially useful for adding male speech diversity and British English phonetic characteristics, strengthening the linguistic coverage of the

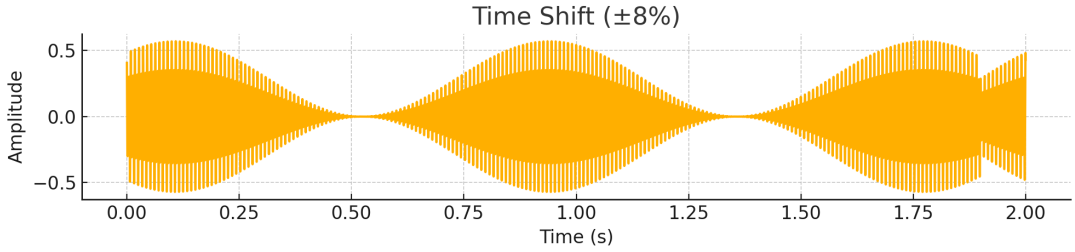
dataset.

By combining these corpora, the NOR dataset introduces a blend of controlled studio recordings and naturally occurring emotional speech, with variation in speaker demographics, recording conditions, and emotional intensity. This heterogeneity was intentional to mimic the unpredictability of real-world acoustic environments and to build a model capable of maintaining performance across varied scenarios.

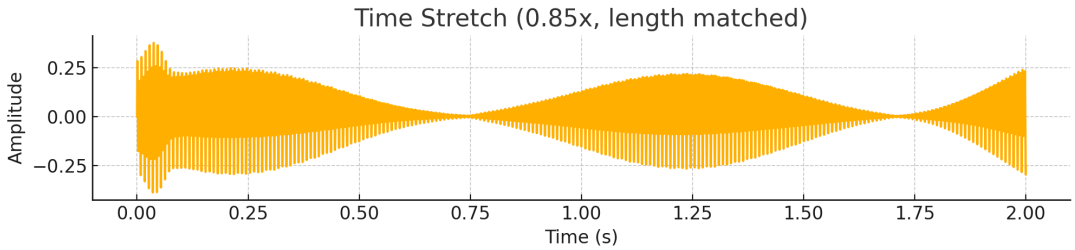
3.6 Audio Augmentation for Robustness

To enhance the robustness and generalisation of the Speech Emotion Recognition (SER) model, an **on-the-fly audio augmentation** strategy was incorporated during training. Data augmentation plays a critical role in preventing overfitting and improving model performance in real-world noisy environments by simulating diverse acoustic conditions [Ko et al., 2015]. In this project, five augmentation techniques with controlled probabilities were applied:

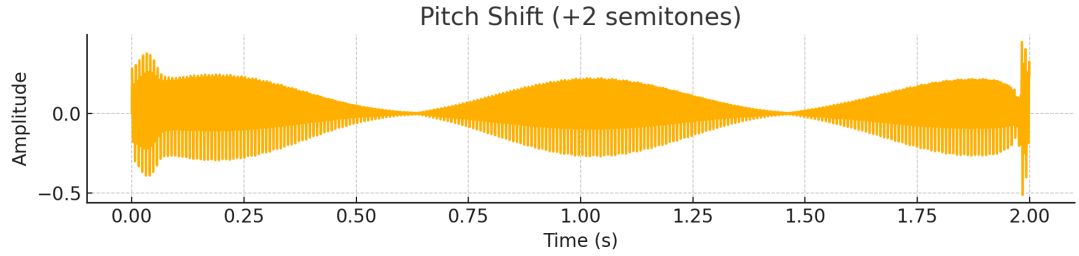
1. **Time Shift (±8% of window; ±320 ms for a 4 s clip)** – Randomly shifts the waveform in time without altering its pitch or duration, simulating variations in speech onset [Salamon and Bello, 2017].



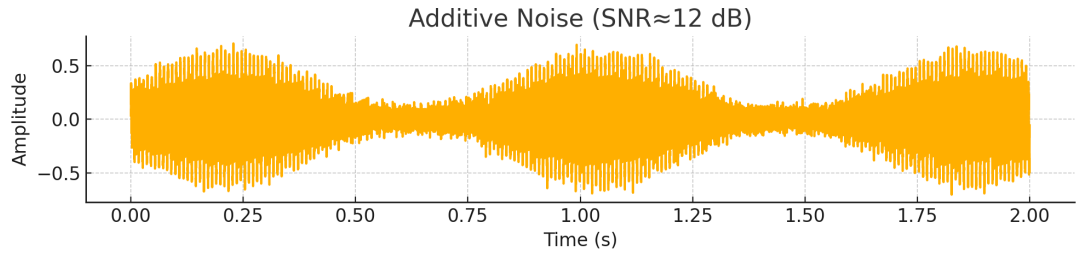
2. **Time Stretch (0.85×)** – Alters the speech speed without changing the pitch, introducing variability in speaking rate while preserving content [Ko et al., 2015].



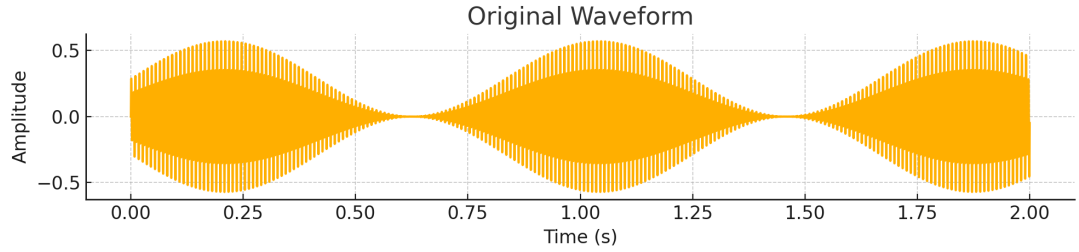
3. **Pitch Shift (+2 semitones)** – Changes the pitch of the waveform, imitating differences in speaker tone or vocal register [Salamon and Bello, 2017].



4. **Additive Noise (SNR ≈ 12 dB)** – Injects Gaussian noise into the waveform, improving model resilience to environmental noise [Snyder et al., 2015; Xu et al., 2021].



5. **Original Waveform** – Retains clean, unmodified audio for baseline learning.



These augmentations were applied dynamically during training rather than generating a static enlarged dataset, thereby preventing dataset size inflation while ensuring unique variations at each epoch. The above figures illustrate the waveform transformations for a sample audio file, highlighting the distinct changes introduced by each augmentation method.

3.6.1 Dataset Design and Justification

The design follows a progressive, cross-corpus strategy to balance clean pretraining with robust generalisation:

- **TESS** (*Toronto Emotional Speech Set*) for initial screening and pretraining on high-quality, controlled utterances.

- **RAVDESS** to fine-tune on richer speaker diversity and emotional intensity variation.

- **NOR (merged corpus)** for broad generalisation via label harmonisation to six classes {*angry, disgust, fear, happy, neutral, sad*} and heterogeneous acoustic conditions.

Split policy. Unless otherwise stated, we used a **stratified 80/10/10** train/validation/test split (fixed seed), with **speaker independence** enforced where metadata permitted.[1] Dynamic, on-the-fly augmentations (time shift, time stretch, pitch shift, additive noise) were applied during training to mitigate overfitting and simulate realistic VR microphone conditions.

[1] Exact splits and seeds are logged alongside each run in the training notebooks.

Table 3.2: Corpora and their roles in the design.

Corpus	Role	Design Justification
TESS [Dupuis and Pichora-Fuller, 2010]	Pretraining / screening	Clean, controlled recordings for stable boundary learning before exposure to noisier data.
RAVDESS [Livingstone and Russo, 2018a]	Transfer fine-tuning	Balanced genders, intensity levels; closer to deployment variability.
NOR (ASVP-ESD, RAVDESS, CREMA-D, SAVEE, TESS)	Robustness evaluation and final training	Heterogeneous acoustic and speaker conditions; labels harmonised to six-way scheme for generalisation.

3.6.2 Evaluation Metrics and Justification

Therapy-facing SER requires both *global* accuracy and *emotion-wise reliability*. We therefore evaluate with overall accuracy, per-class precision/recall, macro-F_1, and confusion matrices, accompanied by loss/metric curves for generalisation checks.

Definitions. For K classes and confusion counts $\mathrm{TP}_k, \mathrm{FP}_k, \mathrm{FN}_k$:

$$\text{Accuracy} = \frac{\sum_{k=1}^{K} \mathrm{TP}_k}{\sum_{k=1}^{K} (\mathrm{TP}_k + \mathrm{FP}_k)} \tag{3.10}$$

$$\text{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \quad \text{Recall}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k} \tag{3.11}$$

$$\mathrm{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{3.12}$$

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} \mathrm{F1}_k \tag{3.13}$$

Why these metrics?

- **Accuracy** contextualizes the results with previous SER work and provides an easily interpretable headline number.
- **Macro-$\mathbf{F}_1$** equally weights classes, preventing dominance by frequent emotions and reflecting clinical risk in minority classes (e.g., fear/disgust).

$$\mathrm{P}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \quad \mathrm{R}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \tag{3.14}$$

$$\mathrm{F1}_k = \frac{2\,\mathrm{P}_k\,\mathrm{R}_k}{\mathrm{P}_k + \mathrm{R}_k}, \tag{3.15}$$

$$\text{Weighted-F1} = \sum_{k=1}^{K} \frac{n_k}{N} \mathrm{F1}_k, \quad \text{where } N = \sum_{k=1}^{K} n_k. \tag{3.16}$$

- **Per-class precision/recall** expose asymmetric errors important to safety (e.g., false "fear" triggers unnecessary de-escalation; missed "fear" risks overexposure).
- **Confusion matrices** reveal systematic confusions (e.g., *fear* vs. *sad*) to guide retraining or thresholding.

Model selection protocol. We select checkpoints by **validation Macro-$\mathbf{F}_1$** (tie-breaker: accuracy), with fixed random seeds and at least three runs to assess stability. Early stopping/patience is used to curb overfitting, supported by training/validation loss curves.

Latency. In addition to classification metrics, we target **sub-500 ms end-to-end latency** (audio capture → inference → action) to maintain a natural therapy flow.

3.7 Proposed Pipeline

The project pipeline followed a progressive refinement structure:

1. Data loading and preprocessing for TESS dataset.
2. Model training and evaluation across multiple architectures.
3. Selection of Wav2Vec2 based on superior performance.
4. Fine-tuning the TESS-trained Wav2Vec2 model on RAVDESS.
5. Extending fine-tuning to the NOR merged dataset.
6. Integrating on-the-fly audio augmentation for improved generalisation.

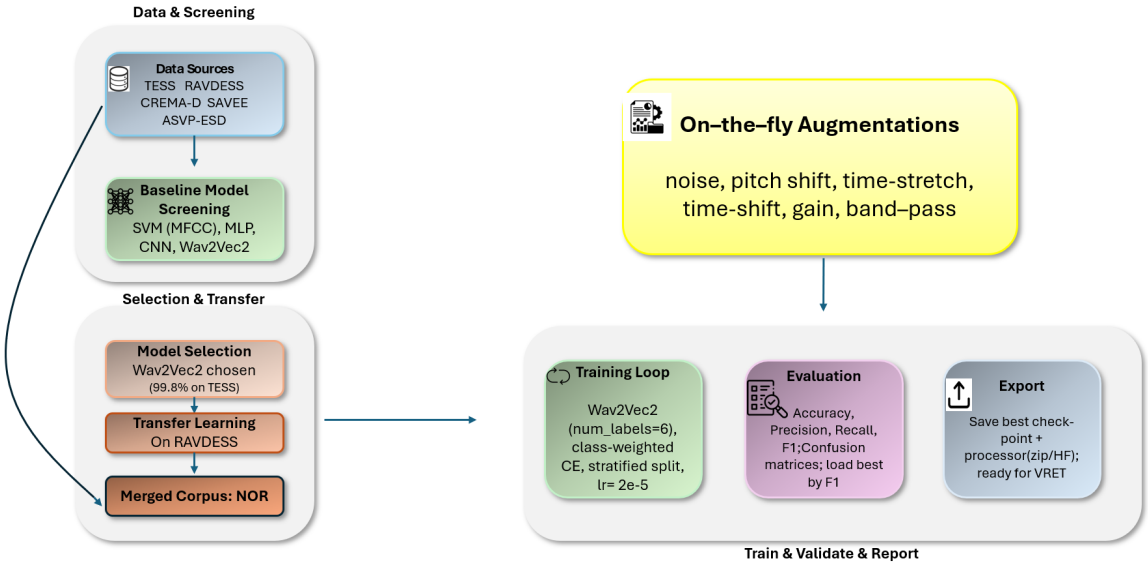


Figure 3.1: SER pipeline. Baseline screening selects Wav2Vec2 (99.8% on TESS), then transfer to RAVDESS, merge to NOR (with label harmonisation), apply on–the–fly augmentations and train, evaluate, and export for VRET.

3.7.1 VRET Integration Design

The SER module operates as a real-time control signal within a public speaking VRET scenario:

1. **Capture:** Headset microphone streams buffered audio windows (e.g., 2–3 s, 50% overlap).
2. **Lightweight preprocessing:** normalisation, (optional) VAD, and resampling to the encoder's native rate.
3. **Inference:** Fine-tuned Wav2Vec2 produces class posteriors $\mathbf{p} \in \mathbb{R}^K$ per window.
4. **Post-processing:** Exponential moving average (EMA) over the last N windows to reduce jitter; hysteresis thresholds to avoid oscillatory scene changes.
5. **Adaptation manager:** Maps emotion states to actions:
   - high *fear*/*anger* $\rightarrow$ de-escalate scene (smaller audience, softer lighting, slower prompt cadence).
   - *neutral*/*happy* $\rightarrow$ escalate challenge (larger audience, more Q&A).
   - low confidence (*uncertain* posteriors) $\rightarrow$ hold state; require persistence over M windows before any change.
6. **Safety & observability:** Optional therapist dashboard for live traces and manual override; privacy-preserving logging of summary statistics only.

This design preserves responsiveness while protecting against spurious fluctuations, ensuring safe, personalized progression.

3.8 Design Rationale

The design choices were influenced by:

- Empirical performance comparisons across baseline and advanced models.
- The proven ability of transformer-based speech models to outperform traditional feature-engineering pipelines.
- The benefits of transfer learning in adapting speech models to diverse datasets.
- The necessity of augmentation to prepare the model for noisy, variable VR environments.

This structured progression ensured that the final SER model was both accurate and resilient for integration into the VRET system.

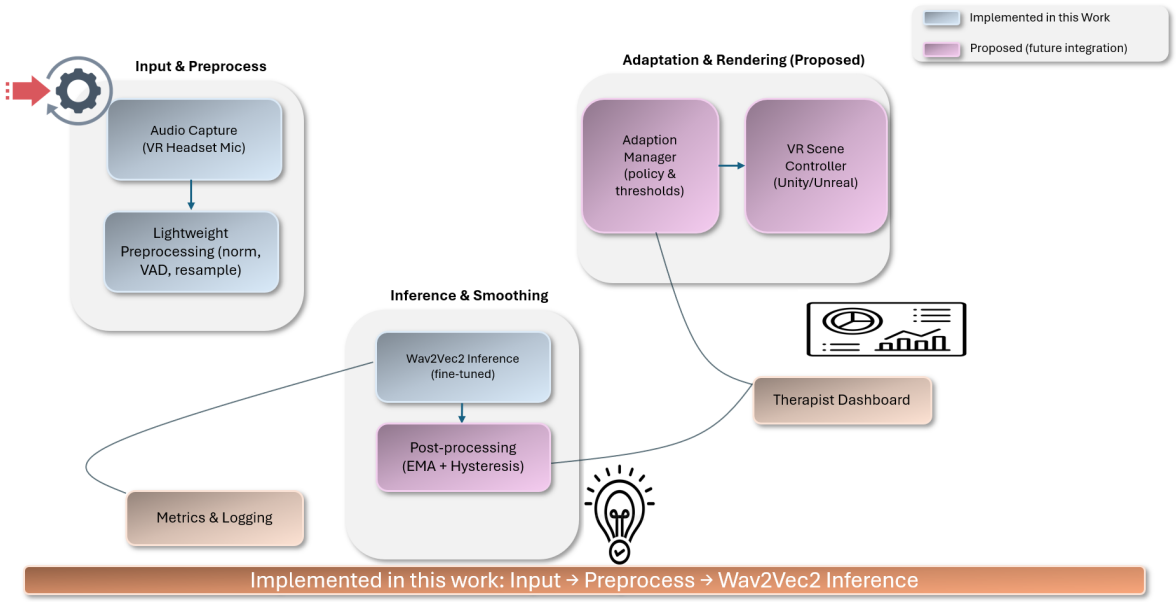


Figure 3.2: Proposed real-world integration of the SER module into VRET. *Blue, solid* blocks denote components implemented in this study (speech capture, preprocessing, Wav2Vec2 inference). *plum, solid* blocks denote proposed deployment elements (smoothing, adaptation policy, VR scene control, therapist dashboard).

Chapter 4

Implementation

4.1 Overview

This chapter details the end-to-end implementation of the speech emotion recognition (SER) system. It covers the computing environment, project structure, data ingestion and preprocessing, dataset splitting and label harmonisation, augmentation strategy, baseline model screening, Wav2Vec2 fine-tuning, evaluation protocol, model export and inference, and the integration hooks for Virtual Reality Exposure Therapy (VRET). Where relevant, figures and code listings are provided to support reproducibility (e.g., waveforms, mel-spectrograms, training curves, and confusion matrices).

4.2 Computing Environment

4.2.1 Hardware

All experiments were run on two execution environments: a local workstation (primary) and the Kaggle hosted runtime (supplementary, for portability and reproducibility). Tables 4.1 and 4.2 summarise the resources used.

Table 4.1: Local workstation hardware summary (primary environment).

Component	Specification
GPUs	GeForce RTX 3080 Ti (CUDA Version: 12.1)
CPU	Multi-core x86_64 (AVX2)
Memory	128 GB RAM
Storage	NVMe SSD for datasets, checkpoints, and logs

Table 4.2: Kaggle hosted runtime resources (supplementary environment).

Resource	Details
CPU	4 vCPU (Xeon-class)
RAM	∼29–30 GB available to the notebook
GPU	1× Tesla P100 (16 GB) *or* 2× Tesla T4 (16 GB each)*
Storage (datasets)	`/kaggle/input` up to 100 GB (read-only)
Storage (outputs)	`/kaggle/working` and Notebook "Output" (persisted), 5 GB quota

* Accelerator type is subject to availability/quota per session.

4.2.2 Software Stack

Training used Python with `PyTorch` [Paszke et al., 2019] and `HuggingFace Transformers` [Wolf et al., 2020]; audio I/O used `librosa/soundfile` [McFee et al., 2025]; metrics used `scikit-learn` [Pedregosa et al., 2011]; and plots used `matplotlib` [Hunter, 2007]. Jupyter notebooks were used for experiment tracking.

4.2.3 Environment Setup

Reproducible setup (conda + pip):

```
conda create -n ser python=3.10 -y
conda activate ser

# Core scientific stack
pip install numpy pandas scikit-learn matplotlib librosa soundfile
```

```
# PyTorch (choose CUDA wheel matching your driver)
pip install torch torchaudio --index-url https://download.pytorch.org/whl/cu121

# HuggingFace ecosystem
pip install transformers datasets accelerate

# (Optional) dev tools
pip install jupyter ipywidgets tqdm
```

4.2.4 Reproducibility Controls

Determinism was encouraged by fixing seeds and saving configuration alongside checkpoints:

- Fixed random seed in Python/NumPy/PyTorch (e.g., 42) and deterministic CuDNN where feasible.
- All training arguments, label maps, and model configs were saved next to the best checkpoint.
- The accelerator type (local T400×2, or Kaggle P100/T4), library versions, and CUDA build were recorded with the run logs.

Seed initialisation (illustrative):

```
import random, numpy as np, torch
def seed_all(seed=42):
    random.seed(seed); np.random.seed(seed)
    torch.manual_seed(seed); torch.cuda.manual_seed_all(seed)
    torch.backends.cudnn.deterministic = True
seed_all(42)
```

4.3 Project Structure and Workflow

This project was executed primarily in **Kaggle Notebooks** with artefacts (checkpoints, logs, figures) saved from each run for reproducibility. Rather than a single monolithic repository, we follow a stage-wise workflow in which a **Wav2Vec2** model is first trained and validated on **TESS** and then **transferred (fine-tuned)** to **RAVDESS**. The key stages and their outputs are summarised below.

4.3.1 Runtime Layout (Kaggle)

Kaggle mounts datasets read-only under `/kaggle/input` and persists notebook outputs under `/kaggle/working`.

```
/kaggle/input/                  # read-only attached datasets (TESS, RAVDESS, ...)
  TESS/ ...
  RAVDESS/ ...
/kaggle/working/                # writable; persisted as notebook Output
  models/                       # best checkpoints, processor/tokenizer, config
  results/                      # figures: training curves, confusion matrices, etc.
  logs/                         # trainer logs and per-epoch metrics
```

4.3.2 Stage-wise Workflow

Stage A — Baseline training and selection on TESS. We trained candidate models and selected **Wav2Vec2** based on validation performance and stability. The chosen Wav2Vec2 run achieved **99.8%** validation accuracy on TESS with consistent macro-F1, and its checkpoint was exported (processor, label maps, config) for transfer learning.

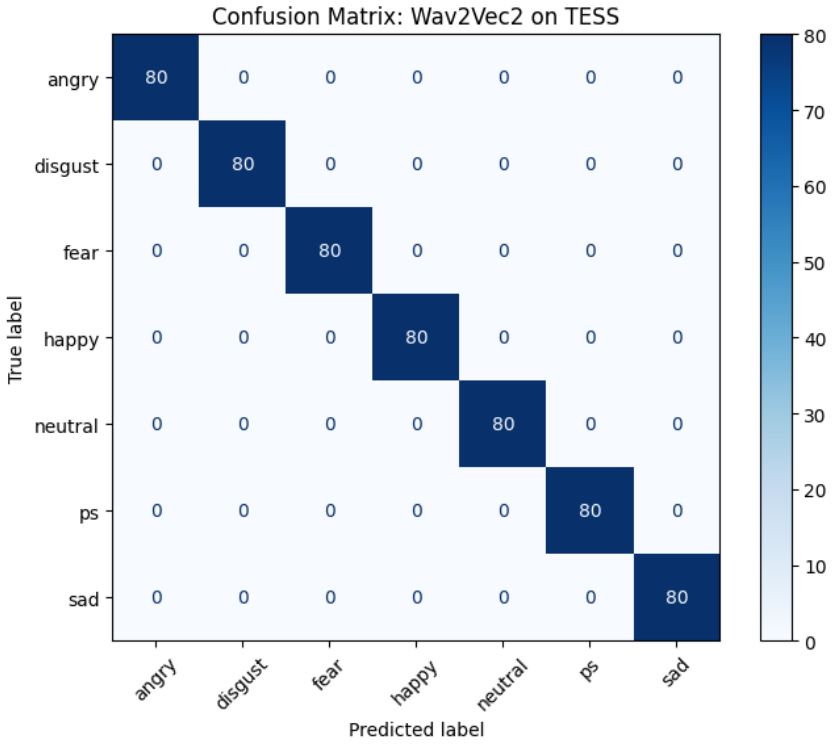


Figure 4.1: Confusion matrix showing perfect classification accuracy of Wav2Vec2 model on the TESS dataset across all seven emotion classes (7-class run).

Stage B — Transfer learning on RAVDESS. Starting from the **TESS-trained Wav2Vec2 checkpoint**, we fine-tuned on RAVDESS using an 80/10/10 split with class balance safeguards. Hyperparameters were kept conservative (e.g., lr $= 2 \times 10^{-5}$, batch size = 8, 6–8 epochs), with *per-epoch* evaluation and model selection by **macro-F1** (tie-break: accuracy). Training curves, confusion matrices, and the best checkpoint were saved to `/kaggle/working`.

4.3.3 Notebook Inventory and Outputs

4.3.4 Reproducibility Notes

- **Fixed seed** (e.g., 42) for Python/NumPy/PyTorch; deterministic CuDNN where feasible.

- **Saved artefacts**: best checkpoint, `config.json`, processor/tokenizer files, `label2id`/`id2label`, and training arguments.

Table 4.3: Notebooks, purpose, and generated artefacts for the TESS → RAVDESS transfer workflow.

Notebook	**Purpose**	**Key Outputs (`/kaggle/working`)**
`tess-wav2vec2.ipynb`	Train and validate Wav2Vec2 on TESS; select stable run.	Best TESS checkpoint (Wav2Vec2), processor/tokenizer; training curves; TESS confusion matrix; **99.8%** val. accuracy.
`ravdess-wav2vec2.ipynb`	Initialise from TESS checkpoint and fine-tune on RAVDESS with per-epoch evaluation.	Best RAVDESS checkpoint; RAVDESS training curves; confusion matrix; logs for macro-F1/accuracy per epoch.

- **Logged** per-epoch metrics and loss curves to facilitate regeneration of figures (training curves, confusion matrices).

4.4 Data Ingestion and Preprocessing

4.4.1 Corpora

We used a staged and then merged approach:

1. **TESS** — clean acted speech for baseline screening and initial stability checks.

2. **RAVDESS** — transfer learning on a richer corpus with varied intensity and speaker diversity.

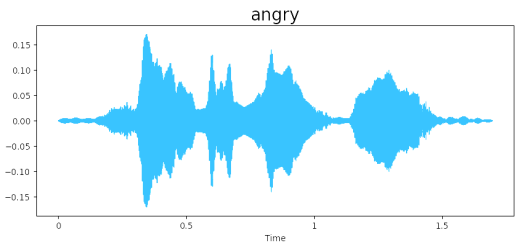
3. **NOR merged corpus** — combination of multiple public corpora (RAVDESS, CREMA-D, SAVEE, TESS, ASVP-ESD) with label harmonisation to a unified 6-class scheme: {*angry, disgust, fear, happy, neutral, sad*}.

4.4.2 Audio Normalisation and Resampling

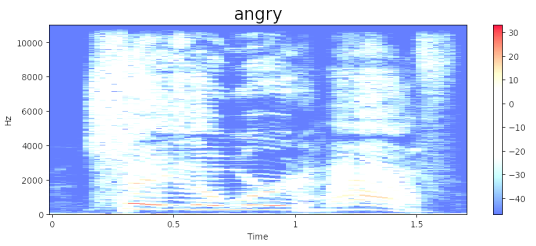
All audio was converted to **mono, 16 kHz** PCM to match the Wav2Vec2 frontend. Waveforms were peak-normalised and optionally trimmed using a simple voice activity detector (VAD) to remove leading/trailing silence for consistent batching.

Illustrative loader (verbatim):

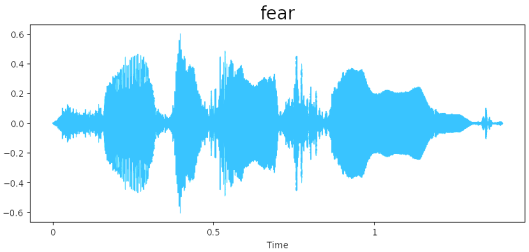
```
import librosa, soundfile as sf
def load_16k_mono(path, target_sr=16000):
    wav, sr = librosa.load(path, sr=None, mono=True)
    if sr != target_sr:
        wav = librosa.resample(wav, orig_sr=sr, target_sr=target_sr)
    wav = 0.98 * (wav / (max(1e-9, abs(wav).max())))  # peak-normalise
    return wav, target_sr
```



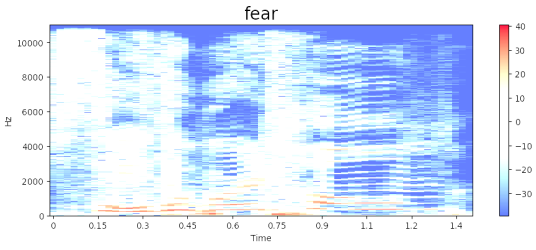
(a) Waveplot - *Angry*



(b) Spectrogram - *Angry*



(c) Waveplot - *Fear*



(d) Spectrogram - *Fear*

Figure 4.2: Waveplot and Spectrogram visualizations of *Angry* and *Fear* emotions from the TESS dataset. These show the temporal and frequency-domain characteristics for different emotional vocal expression.

4.5 Dataset Splits and Label Harmonisation

4.5.1 Split Policy and Speaker Independence

Unless otherwise noted, we used a **stratified 80/10/10** train/validation/test split with a fixed seed to guarantee reproducible partitions and comparable class distributions. Where speaker IDs were available, splits enforced **speaker independence**, preventing leakage and inflated performance estimates.

4.5.2 Label Harmonisation for NOR

Different corpora label emotions with varying taxonomies. All datasets were mapped to the common 6-class scheme {*angry, disgust, fear, happy, neutral, sad*} to support consistent training and evaluation. Mapping rules were implemented during ingestion (e.g., collapsing sub-categories or intensity variants into the canonical class).

4.5.3 Class Balance and Weights

Residual imbalance across classes was addressed via **class-weighted cross-entropy**. Weights were computed inversely proportional to class frequency in the training split.

Illustrative weight computation:

```
import numpy as np
# counts[k] = number of training examples for class k
weights = (1.0 / np.maximum(1, counts))
weights = (weights / weights.sum()) * len(counts)
```

4.5.4 Per-Class Counts

Notes. TESS counts: each base class has 400 files; *pleasant_surprise* (400) is mapped to *happy* $\Rightarrow$ Happy = 800 under the 6-class scheme. RAVDESS counts reflect the selected emotions only (angry, fear, happy, neutral, sad), with non-neutral classes including two intensities (384) and *neutral* a single intensity (192); *disgust* is excluded (0). NOR class

Table 4.4: Per-class instance counts after preprocessing and splitting. TESS merges *pleasant_surprise* into *happy*. NOR uses a stratified 80/10/10 split.

Emotion	TESS	RAVDESS	NOR (train)	NOR (val)	NOR (test)
Angry	400	384	2701	338	337
Disgust	400	0	2258	282	282
Fear	400	384	2321	290	290
Happy	800	384	2945	368	368
Neutral	400	192	2616	327	327
Sad	400	384	2750	344	343
Total	**2800**	**1728**	**15591**	**1949**	**1947**

totals (happy 3681, sad 3437, angry 3376, neutral 3270, fear 2901, disgust 2822) are split 80/10/10 as shown (integers rounded so that each row sums exactly).

4.6 Augmentation Strategy

To improve robustness to headset microphones, buffer jitter, and room acoustics typical of VRET scenarios, we apply **on-the-fly, label-preserving waveform augmentations** during training. At each mini-batch, a random subset of transforms is sampled (independently) and applied to the input audio; validation/test audio is *never* augmented. This stochastic policy exposes the model to a wider acoustic manifold while keeping emotion labels intact [Tao et al., 2022], which reduces overfitting and improves cross-corpus generalisation [Ko et al., 2015]. The qualitative effect of each transform is illustrated in Figure 4.4 (mel views for interpretation only; Wav2Vec2 trains on raw waveforms).

Implementation policy. Augmentations are implemented at the waveform level with fixed 16 kHz mono sampling, typically in 4 s windows. We use moderate ranges that preserve prosodic cues (pitch contour, energy dynamics, timing) while covering variability expected in deployment. Unless otherwise stated, the ranges below refer to the NOR training runs ("medium" preset): *Noise* 12–22 dB SNR; *Pitch* ±1.0 semitone; *Time-*

stretch $\times 0.94$–1.06; time shift up to $\pm 8\%$ of window length (e.g., ± 320 ms for a 4 s window); *Gain* ± 1.5 dB; *Band-pass* 250–3800 Hz (or 100–7000 Hz for the visual demo).

Additive noise (SNR-controlled)

Why it helps. Headset mics in VR capture fan noise, breath, and ambient room noise. Training with SNR-controlled Gaussian noise encourages the classifier to rely on robust spectral/temporal patterns rather than artefacts of clean studio audio [Snyder et al., 2015]. **How we apply it.** For a waveform x, we draw SNR uniformly in 12–22 dB and add zero-mean noise with matching power. A fresh noise realisation is sampled each time an utterance is seen. **Effect.** Reduces the train–val gap and makes predictions stable when background noise varies. **Caveats.** Too low SNR (< 10 dB) begins to mask emotion cues (e.g., breathiness, low-energy sadness). We therefore keep SNR moderate.

Pitch shift (small semitone offsets)

Why it helps. Speakers differ in fundamental frequency and formant structure (sex/age/microphone distance). Small pitch shifts promote invariance to absolute pitch while preserving relative prosody [Salamon and Bello, 2017]. **How we apply it.** We draw $n_{\text{steps}} \sim \mathcal{U}(-1.0, +1.0)$ st and shift with `librosa.effects.pitch_shift`. **Effect.** Improves robustness across speakers and reduces spurious correlations between absolute pitch and emotion. **Caveats.** Larger shifts ($> \pm 2$ st) can distort perceived emotion (e.g., happy $\leftrightarrow$ surprise). We cap at ± 1 st in NOR runs.

Time-stretch (speaking-rate variation)

Why it helps. Speaking rate varies by context and anxiety level; VR sessions also introduce buffering jitter. Mild stretching enforces invariance to rate without destroying phonetic timing [Ko et al., 2015]. **How we apply it.** We draw $r \sim \mathcal{U}(0.94, 1.06)$ and apply `librosa.effects.time_stretch`, then pad/crop back to fixed window length. **Effect.** Encourages reliance on spectral/prosodic shape rather than absolute timing, aiding

generalisation across corpora. **Caveats.** Strong stretching (< 0.9 or > 1.1) produces artefacts and may invalidate labels.

Time shift (onset/offset jitter)

Why it helps. In streaming inference, we rarely cut exactly at phone/word boundaries; VAD and buffering introduce offsets [Salamon and Bello, 2017]. Circular time shifts teach the model to be invariant to where the relevant content sits inside the window. **How we apply it.** We roll the waveform by a random offset up to ±320 ms (or ≈8% of a 4 s window). **Effect.** Reduces sensitivity to segmentation, making window-level predictions smoother. **Caveats.** Excessive shifts add no value and can duplicate other variability; we keep the offset small.

Random gain (microphone level/AGC)

Why it helps. Headset gain and mouth-to-mic distance vary between sessions; some capture pipelines apply automatic gain control (AGC). Small level changes prevent the model from anchoring on absolute amplitude [Schlüter and Grill, 2015]. **How we apply it.** We draw a gain in ±1.5 dB and scale the waveform; peaks are softly clipped to avoid distortion. **Effect.** Stabilises predictions under minor level differences without altering spectral content. **Caveats.** Large gains risk clipping; we keep the range narrow.

Band-pass filtering (device/room response)

Why it helps. Consumer headsets emphasise the speech band and attenuate extremes; rooms impose mild coloration. A band-pass encourages invariance to such transfer functions and reduces cross-device mismatch [Schlüter and Grill, 2015]. **How we apply it.** A 4th-order Butterworth band-pass is applied with passband 250–3800 Hz (a headset-like response). For visualisation in Figure 4.4 we also show 100–7000 Hz to make the effect more evident. **Effect.** Improves robustness when moving from studio corpora to headset recordings. **Caveats.** Over-aggressive filtering (very narrow bands) removes emotion-bearing harmonics; we therefore keep a wide speech band.

Summary. These transforms are *compositional*: more than one may be active on a given batch, and parameters are re-sampled each epoch. Empirically this reduced overfitting (smaller train–validation gap) and yielded more stable macro-F1 on the NOR split, particularly for minority classes, while keeping validation/Test data untouched to ensure fair evaluation.

Visual illustration and scope. The on-the-fly augmentation policy described above is applied during training on the **NOR** corpus (merged dataset) to improve robustness to headset microphones and room acoustics. To *visualise* the effect of each transform, we render mel-spectrograms from a single *TESS* utterance (chosen only for clarity in figures). The learning model itself is **Wav2Vec2**, which operates directly on **raw waveform** [Baevski et al., 2020]; mel-spectrograms are *not* used as input features in our final system—they are included here solely to make the acoustic effect of the augmentations interpretable and to justify the chosen parameter ranges.

Rendering details (for reproducibility). All panels use 16 kHz mono audio with fixed 4 s windows. *STFT/mel parameters:* `n_fft=512`, `hop=160`, `win=400`, `n_mels=80`, `f_min=50 Hz`, `f_max=8 kHz`. *Colour scale:* fixed to −80–0 dB across all images. *Demonstration settings:* Noise 20 dB SNR; Pitch +1.0 st; Stretch ×0.96; Shift +320 ms (0.08 of 4 s); Band-pass 100–7000 Hz.

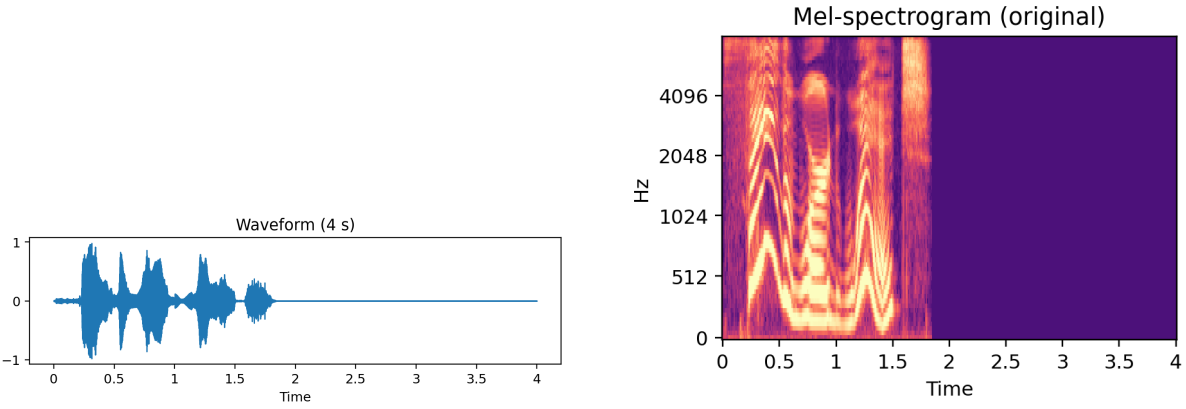
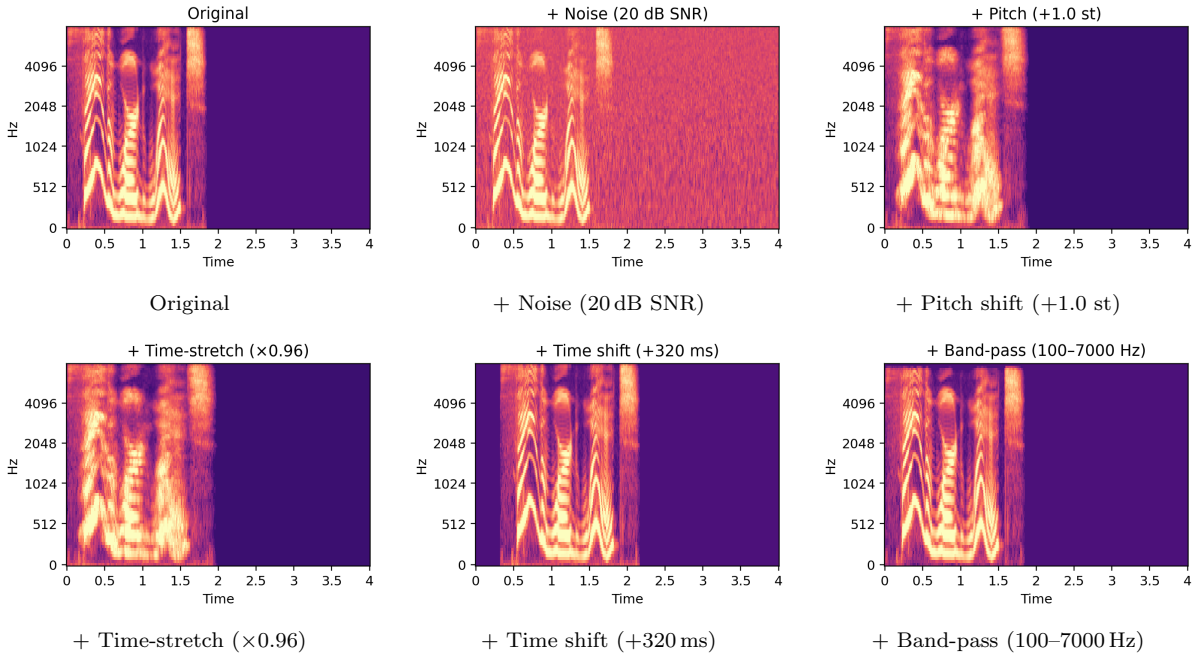


Figure 4.3: Waveform (left) and mel-spectrogram (right) of the illustrative utterance used for visualisation. These plots are *representative* only; the final Wav2Vec2 model trains on raw waveform.

Why visualise mel-spectrograms if the model uses waveforms? First, they pro-



Original + Noise (20 dB SNR) + Pitch shift (+1.0 st)

+ Time-stretch (×0.96) + Time shift (+320 ms) + Band-pass (100–7000 Hz)

Figure 4.4: Effect of each augmentation on the same utterance (mel-spectrograms). Although the NOR training loop applies these transforms to *waveforms* on-the-fly before Wav2Vec2 ingestion, mel views help verify that parameters stay within realistic bounds (e.g., mild pitch/time changes preserve prosody) and expose unintended artefacts early.

vide an *explanatory lens* for reviewers, revealing how each transform redistributes energy over time–frequency. Second, they support *parameter tuning*: too-aggressive settings (e.g., extreme stretch or narrow band-pass) are immediately visible and can be corrected before large runs. Third, they enable *like-for-like comparisons* across experiments because all panels share duration and dB scaling.

4.7 Baseline Screening and Model Selection

We first established classical baselines on **TESS**: SVM with MFCCs, MLP on statistical descriptors, a CNN on mel-spectrograms, and a simple voting classifier (LR+RF+MLP). These provided reference performance and sanity checks for the preprocessing/label pipeline. In screening runs, **Wav2Vec2** outperformed other candidates (reaching **99.8%** on the TESS screening set in our notebook), and was therefore selected for subsequent transfer and expansion.

Table 4.5: Baseline screening and transfer results. Accuracies are from the corresponding validation/test splits.

Model	Dataset	Accuracy
Wav2Vec2 (6 epochs)	RAVDESS	97.6%
Wav2Vec2	TESS	99.8%
SVM	TESS	96%
CNN (30 epochs)	RAVDESS	62.15%
Ensemble Voting	RAVDESS	93.4%
CNN Deep	RAVDESS	92.5%

4.8 Wav2Vec2 Fine-Tuning

4.8.1 RAVDESS Transfer Learning

The selected model was fine-tuned on **RAVDESS**, achieving approximately **97.6%** after around **6 epochs**. We adopted conservative regularisation and monitored validation performance per epoch to select stable checkpoints.

[522/522 07:07, Epoch 6/6]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	No log	1.003979	0.713873	0.715266	0.713873	0.699747
2	No log	0.678757	0.846821	0.858148	0.846821	0.844320
3	No log	0.468110	0.901734	0.904132	0.901734	0.900359
4	No log	0.271100	0.965318	0.965823	0.965318	0.965276
5	No log	0.198957	0.979769	0.980231	0.979769	0.979762
6	0.598500	0.168946	0.976879	0.977232	0.976879	0.976893

Figure 4.5: Wav2Vec2 fine-tuning on **RAVDESS**: per-epoch validation metrics (loss, accuracy, precision, recall, F1). The run converges within 6 epochs with peak performance around epochs 5–6; the best checkpoint is chosen by macro-F1.

4.8.2 Expansion on NOR

The merged **NOR** corpus combines multiple public SER datasets into a single six-class taxonomy {*angry, disgust, fear, happy, neutral, sad*}. After preprocessing (mono, 16 kHz) the corpus comprises **19,487** clips in total, stratified into **80/10/10** train/validation/test splits (**15,591** / **1,949** / **1,947** items; class-wise counts are reported in Table 4.4). Compared to TESS/RAVDESS, NOR is acoustically diverse (different microphones and rooms) and many files include *short non-speech regions* (silence, breaths, lip smacks) at the beginning/end. Rather than trimming these aggressively, we retain mild context and rely on the augmentation policy (Section 4.6) to inoculate the model against variability.

Initialisation and objective. We initialise Wav2Vec2 from the **TESS-trained** checkpoint (99.8% on TESS) and the subsequent **RAVDESS-fine-tuned** state (97.6% at 6 epochs), then continue fine-tuning on NOR. The learning objective is **class-weighted cross-entropy** to counter residual imbalance. With w_k the weight for class k (inverse-frequency, normalised), logits $\mathbf{z}$ and label y, the loss is

$$\mathcal{L} = -\, w_y \, \log \Big(\mathrm{softmax}(\mathbf{z})_y \Big).$$

Weights are computed on the training split only and kept fixed across epochs.

Why this matters for NOR. The larger scale of NOR and the presence of short non-speech fragments make models sensitive to segmentation, level, and channel differences. Two design choices improved robustness:

1. **On-the-fly waveform augmentation** (noise 12–22 dB SNR; pitch ±1.0 st; time-stretch ×0.94–1.06; time shift up to ±90 ms; gain ±1.5 dB; headset-like band-pass). This widens the acoustic manifold seen during training while preserving emotion labels.

2. **Class-weighted CE** focuses learning on minority classes and reduces bias from frequent emotions (e.g., *happy*, *sad*).

4.8.3 Training Configuration (NOR)

- **Sampling**: 16 kHz mono; fixed 4 s windows (pad/crop). This tolerates short leading/trailing non-speech without discarding clips.

- **Model**: `Wav2Vec2ForSequenceClassification` (`num_labels`=6), initialised from the RAVDESS-tuned checkpoint; `ignore_mismatched_sizes=True` to adapt the head if needed.

- **Optimiser**: AdamW, learning rate 2×10^{-5}; no layer freezing; standard weight decay (default).

- **Batching**: batch size 8 (per device), dynamic padding inside the processor.

- **Augmentation**: applied *only* on the training `Dataset` (Section 4.6); validation/test are never augmented.

- **Loss**: class-weighted cross-entropy with $w_k \propto 1/\text{freq}(k)$ (normalised to $\sum_k w_k = K$).

- **Evaluation**: at the end of each epoch on the validation split; metrics reported are **Accuracy**, **Precision**, **Recall**, and **Weighted F1**. Model selection uses **macro/weighted F1** as the primary criterion (tie-break by accuracy).

- **Epochs**: **10** epochs (no early stopping required; convergence plateau observed around epochs 8–10).

Outcome. Across **10 epochs** the model steadily improved, reaching a validation **Accuracy of 82.19%** and **Weighted-F1 of ≈0.822** (Figure 4.6). The run demonstrates that (i) pretraining on cleaner corpora (TESS) followed by transfer (RAVDESS) provides a strong starting point for NOR, and (ii) on-the-fly augmentation plus class-weighted loss are effective guards against overfitting and class imbalance on a large, heterogeneous corpus that includes short non-speech segments.

[9750/9750 2:28:43, Epoch 10/10]

Epoch	Training Loss	Validation Loss	Accuracy	Precision	Recall	F1
1	1.003000	0.860664	0.720883	0.730815	0.720883	0.722325
2	0.840300	0.730775	0.751154	0.764732	0.751154	0.750914
3	0.693400	0.641532	0.786557	0.790927	0.786557	0.786050
4	0.620600	0.601095	0.795280	0.799014	0.795280	0.795190
5	0.532400	0.556260	0.813238	0.814853	0.813238	0.812697
6	0.491700	0.637334	0.793740	0.799951	0.793740	0.793972
7	0.419200	0.594414	0.816316	0.818065	0.816316	0.816366
8	0.385000	0.598723	0.806567	0.810481	0.806567	0.806573
9	0.312900	0.591484	0.816829	0.817057	0.816829	0.816687
10	0.271300	0.599801	0.821960	0.822787	0.821960	0.822106

[122/122 01:47]

Figure 4.6: Wav2Vec2 fine-tuning on **NOR**: per-epoch validation metrics (loss, accuracy, precision, recall, F1). The final epoch (10/10) attains **Accuracy 82.19%** with **Weighted-F1 $\approx$ 0.822**.

HuggingFace `Trainer` (skeleton):

```
MODEL_PATH = "/kaggle/input/ravdess-wav2vec2"


processor = Wav2Vec2Processor.from_pretrained(MODEL_PATH)


# 6 labels for the NOR dataset.
model = Wav2Vec2ForSequenceClassification.from_pretrained(
    MODEL_PATH,
    num_labels=6,
    problem_type="single_label_classification",
    ignore_mismatched_sizes=True  # Flag to ignore mismatch in the head size if the c
).to(DEVICE)
```

```
# Define the training arguments
args = TrainingArguments(
    output_dir="models/w2v2_nor",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=10,
    eval_strategy="epoch",
    save_strategy="epoch",
    load_best_model_at_end=True,
    metric_for_best_model="f1",  # macro-F1 for model selection
    greater_is_better=True,
    seed=42
)

# Initialize the Trainer
trainer = Trainer(
    model=model,
    args=args,
    train_dataset=train_ds,
    eval_dataset=val_ds,
    tokenizer=processor,
    compute_metrics=compute_metrics,  # Function to compute metrics such as accuracy,
    data_collator=collator
)

# Start training
trainer.train()
```

4.8.4 Training Curves and Logs

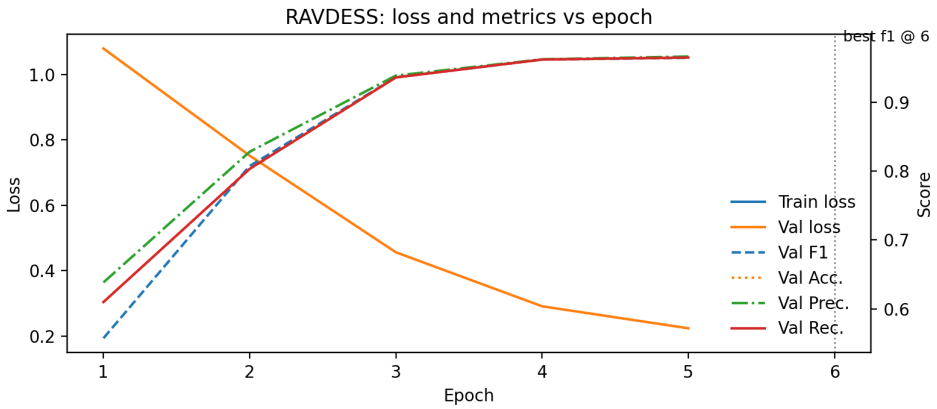


Figure 4.7: Training/validation loss and macro-F1 across epochs for RAVDESS fine-tuning (best checkpoint highlighted).

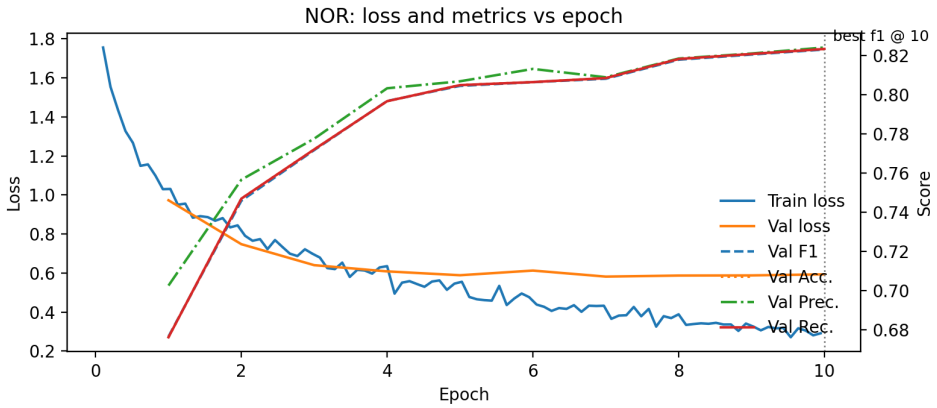


Figure 4.8: Training/validation curves for the NOR expansion run.

4.9 Evaluation Protocol and Metrics

We report both **global performance** and **per-class reliability**. Accuracy supplies an intuitive headline number; per-class precision/recall reveal asymmetric errors (e.g., false fear vs. missed fear) [Saito and Rehmsmeier, 2015]; **macro-F1** averages class-wise F1 equally, preventing dominance by frequent classes; confusion matrices expose systematic confusions that guide re-training or thresholding.

Let $\mathrm{TP}_k, \mathrm{FP}_k, \mathrm{FN}_k$ be counts for class k:

$$\mathrm{Precision}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FP}_k}, \quad \mathrm{Recall}_k = \frac{\mathrm{TP}_k}{\mathrm{TP}_k + \mathrm{FN}_k}, \quad \mathrm{F1}_k = \frac{2\,\mathrm{Precision}_k\,\mathrm{Recall}_k}{\mathrm{Precision}_k + \mathrm{Recall}_k}.$$

$$\text{Macro-F1} = \frac{1}{K}\sum_{k=1}^{K} \text{F1}_k, \qquad \text{Accuracy} = \frac{\sum_{k=1}^{K} \text{TP}_k}{\sum_{k=1}^{K} (\text{TP}_k + \text{FP}_k)}$$

.

4.9.1 Confusion Matrices and Class-wise Scores

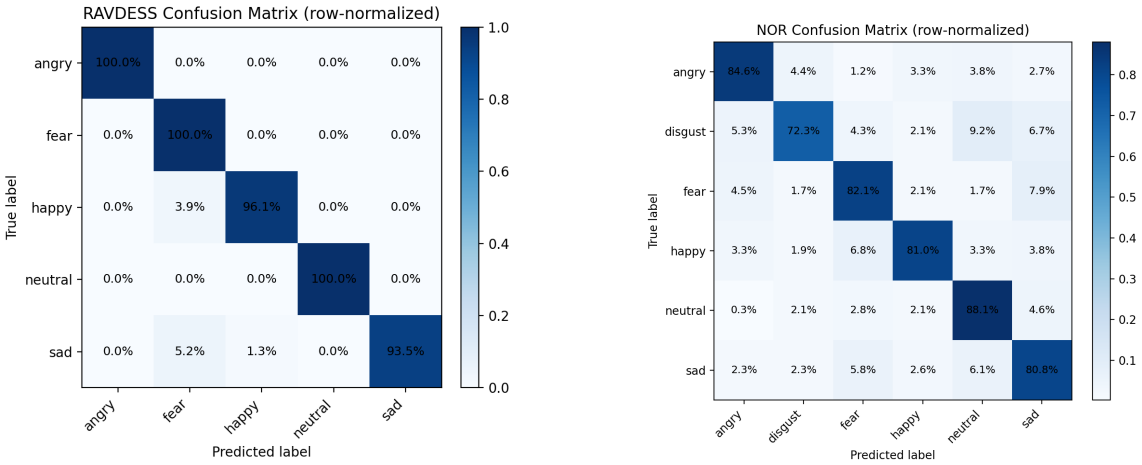


Figure 4.9: Confusion matrices for RAVDESS (left) and NOR (right) test sets computed on the best macro-F1 checkpoint.

ROC (one-vs-rest).

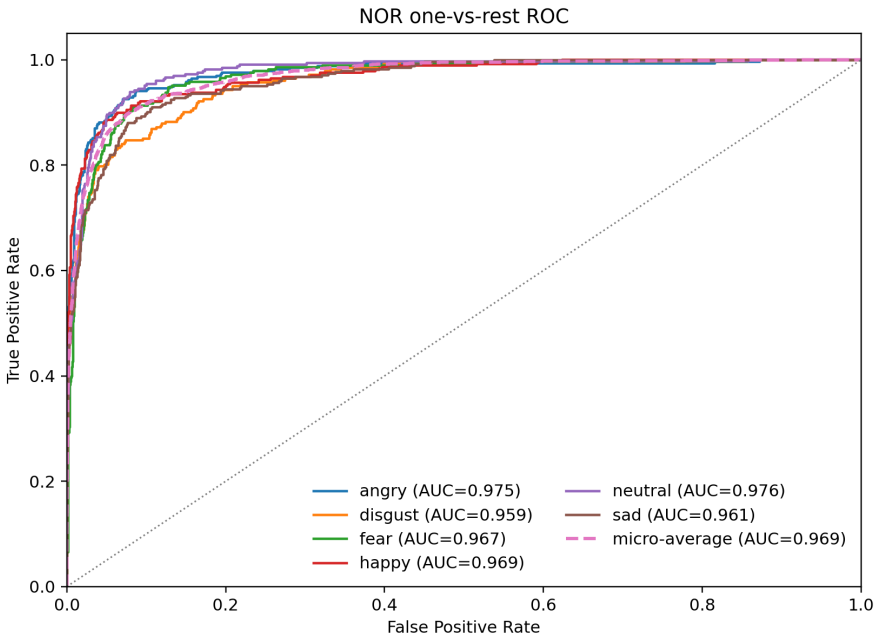


Figure 4.10: One-vs-rest ROC curves for the six classes on the NOR test split.

4.10 Model Export and Inference

4.10.1 Export Artefacts

At the end of training, the best model (by validation macro-F1) was exported alongside:

- `config.json` and `pytorch_model.bin` (weights),
- processor/tokenizer files (feature extractor),
- `label2id` / `id2label` mapping,
- `trainer_state.json` or equivalent logs with seeds/args.

Artefacts were stored locally and optionally packaged as a zip or pushed to a private HuggingFace repository for versioned access.

4.10.2 Windowed Inference and Aggregation

For real-time systems like VRET, low-latency inference is crucial for timely adjustments. Techniques like **windowed inference** enable this by processing overlapping windows of data to generate real-time predictions. In this context, [Chang et al., 2020] demonstrated efficient streaming methods that minimize inference delay, making them well-suited for dynamic, real-time applications like VRET.
File-level predictions used **windowed inference** (e.g., 2–3 s windows with 50% overlap). Window posteriors can be aggregated by:

1. **Averaging probabilities** over windows, then taking argmax.
2. **Majority voting** on window argmax labels.

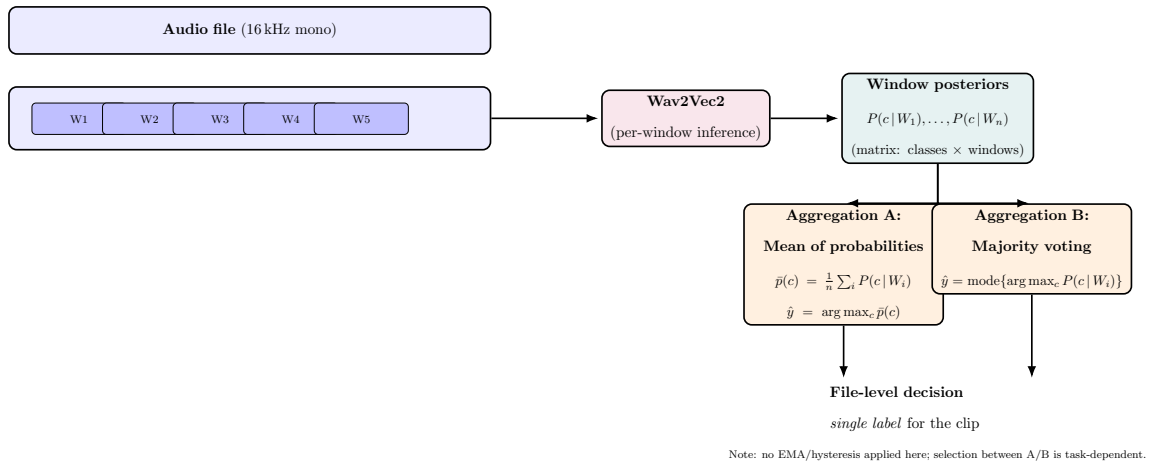


Figure 4.11: Windowed inference and aggregation. An audio file is buffered into overlapping windows; each window is scored by Wav2Vec2 to produce class posterior probabilities. File-level prediction is obtained either by *averaging probabilities* (A) or by *majority voting* over window argmax labels (B).

Notation: W_i denotes the i-th overlapping window of length L seconds extracted with hop H (for 50% overlap, $H = L/2$). For a clip of duration T, the number of windows is $n = \left\lfloor \frac{T-L}{H} \right\rfloor + 1$ (padding or truncation applied to the last window if needed).

4.11 (Proposed) Real-Time Post-Processing

To produce a stable control signal for VR adaptation, we *propose* a light post-processing layer: **Exponential Moving Average (EMA)** smoothing of class posteriors followed by **hysteresis**. EMA reduces jitter; hysteresis enforces dual thresholds [Cockrill, 2011] and/or persistence over recent windows to prevent rapid toggling [Filters, 1991].
Let $p_t^{(k)}$ be the model probabilities for class k at window t. Maintain smoothed $s_t^{(k)}$ via

$$s_t^{(k)} = \alpha\, p_t^{(k)} + (1-\alpha)\, s_{t-1}^{(k)}, \quad \alpha \in (0,1].$$

Switch to class c only when $s_t^{(c)} \geq \tau_{\text{high}}$ and c has persisted for M of the last N windows; relinquish c when $s_t^{(c)} \leq \tau_{\text{low}}$ ($\tau_{\text{low}} < \tau_{\text{high}}$). Typical starting values: $\alpha \in [0.4, 0.6]$, $(\tau_{\text{low}}, \tau_{\text{high}}) = (0.55, 0.65)$, $M/N = 3/5$.

4.12 Integration Hooks for VRET

Although the focus here is SER, the implementation exposes clean interfaces to a VR engine (Unity/Unreal):

1. **Audio capture**: VR headset mic streams PCM buffers at 16 kHz mono.
2. **SER service API**: accept audio buffer, return emotion probabilities $\mathbf{p} \in \mathbb{R}^6$.
3. **(Proposed) Post-processing**: EMA + hysteresis to stabilise emotion state.
4. **Adaptation policy**: map emotion state to scene parameters (audience size, lighting, prompt cadence); apply hysteresis/cool-downs to avoid oscillation.
5. **Observability**: optional dashboard for logging and therapist override (deferred).

Refer to Figure 3.2 in the Design chapter for a view of the system level.

4.13 Logging, Tracking, and Artefacts

All runs stored:

- training/evaluation logs (per-epoch loss, accuracy, precision, recall, and F1) in `trainer_state.json` and the console; see Fig. 4.6 for the NOR run's epoch-wise log,
- configuration files (hyperparameters, label maps, seeds),
- best and intermediate checkpoints,
- plots (training curves, confusion matrices) regenerated from logs.

Example: the NOR fine-tuning log in Fig. 4.6 shows steady improvement through epoch 10, reaching **82.19%** validation accuracy with **weighted F1** ≈ 0.822.

4.14 Summary

In summary, the pipeline begins with carefully normalised and harmonised multi-corpus audio, applies targeted on-the-fly augmentations during training, and fine-tunes a Wav2Vec2 classifier with class-weighted loss and macro-F1-driven model selection. Comprehensive evaluation (accuracy, per-class precision/recall, macro-F1, confusion matrices) demonstrates performance across emotions, while the exported artefacts and windowed inference procedure make the model practical for downstream integration. A lightweight post-processing layer (EMA + hysteresis) is proposed to stabilise real-time behaviour in VRET deployment.

Chapter 5

Evaluation and Results

This chapter reports quantitative and qualitative results from three angles: (i) classical baselines and early screening, (ii) supervised fine-tuning of **Wav2Vec2** on RAVDESS followed by expansion to the merged **NOR** corpus, and (iii) unsupervised structure and feature analyses on RAVDESS. We present per-epoch logs, confusion matrices, ROC curves, and clustering diagnostics, and we interpret the main findings for model selection and deployment.

5.1 Baseline Screening on TESS and RAVDESS

Aggregate comparison. Figure 5.1 summarises the headline accuracies from the screening phase and subsequent transfers. **Wav2Vec2** is consistently dominant on TESS (99.8%) and RAVDESS (97.6% in 6 epochs) and remains competitive on the more challenging NOR mixture (82.2%). Among classical models trained on RAVDESS, a shallow **MLP** and a **CNN** reach ≈93%, while a simple **Voting** ensemble underperforms unless the MLP is included (80.7% vs. 93.4%).

Classical baselines on RAVDESS. The comparative bar chart in Figure 5.2 confirms that neural baselines (MLP/CNN) outperform a simple Voting classifier, and that adding the MLP to the ensemble restores performance to the low-93% range.

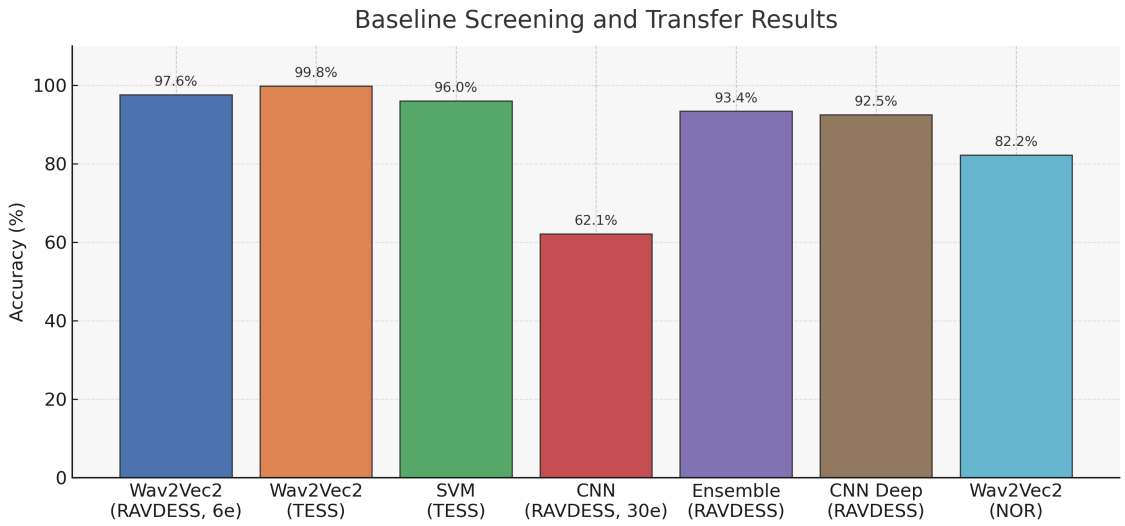


Figure 5.1: Baseline screening and transfer results (accuracy). Wav2Vec2 leads on TESS and RAVDESS and remains robust on NOR.

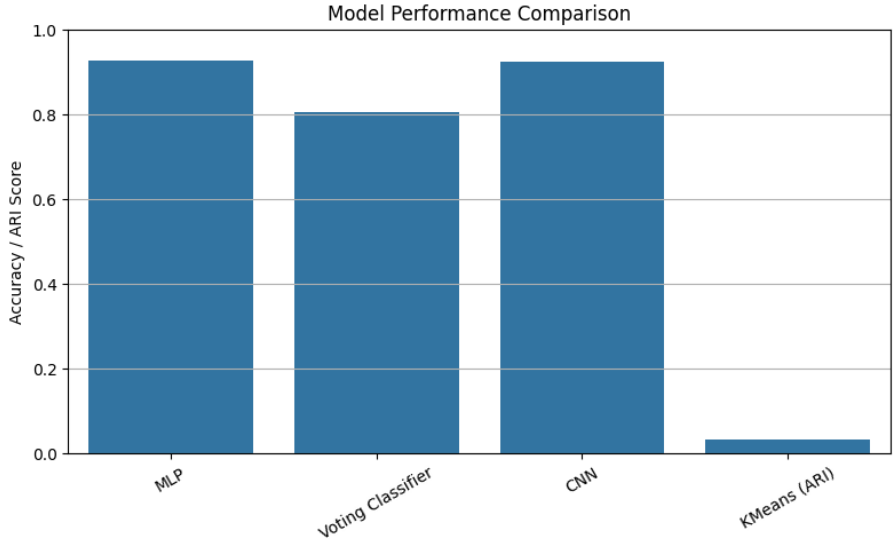


Figure 5.2: Classical baselines on RAVDESS. Accuracy for MLP, Voting Classifier, CNN; ARI for KMeans shown for reference.

Confusion patterns of the Voting classifier. The two renderings in Figure 5.3 depict the same counts with different colourmaps. Most errors arise between acoustically neighbouring emotions (*neutral* vs. *calm*/*sad*, *happy* vs. *surprised*), motivating richer representations.

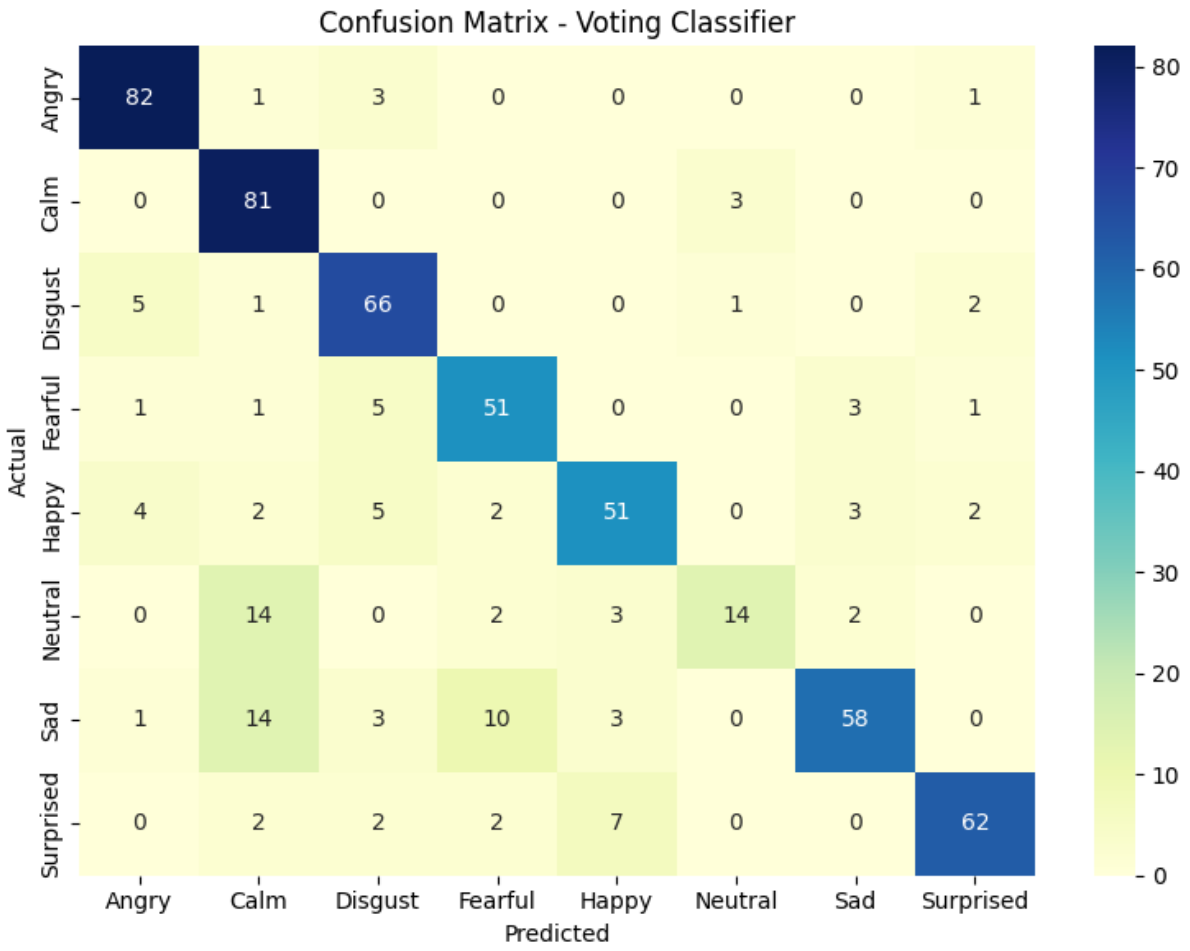


Figure 5.3: Voting classifier confusion matrix.

Per-model reports (RAVDESS). The classification reports in Figure 5.4 show that the *MLP* reaches 92.7% accuracy, *Voting+MLP* reaches 93.4%, and the *Voting (no MLP)* baseline lags at 80.7%. The CNN achieves 92.5% (screenshot in Figure 5.7). These trends echo the aggregate comparison.

MLP Classifier Accuracy: 0.9270833333333334

	precision	recall	f1-score	support
Angry	0.98	0.98	0.98	87
Calm	0.95	0.95	0.95	84
Disgust	0.93	1.00	0.96	75
Fearful	0.94	0.97	0.95	62
Happy	0.83	0.86	0.84	69
Neutral	0.77	0.94	0.85	35
Sad	0.98	0.89	0.93	89
Surprised	0.97	0.84	0.90	75
accuracy			0.93	576
macro avg	0.92	0.93	0.92	576
weighted avg	0.93	0.93	0.93	576

Figure 5.4: RAVDESS: MLP report (accuracy 92.7%).

Voting Classifier Accuracy: 0.8072916666666666

	precision	recall	f1-score	support
Angry	0.88	0.94	0.91	87
Calm	0.70	0.96	0.81	84
Disgust	0.79	0.88	0.83	75
Fearful	0.76	0.82	0.79	62
Happy	0.80	0.74	0.77	69
Neutral	0.78	0.40	0.53	35
Sad	0.88	0.65	0.75	89
Surprised	0.91	0.83	0.87	75
accuracy			0.81	576
macro avg	0.81	0.78	0.78	576
weighted avg	0.82	0.81	0.80	576

Figure 5.5: RAVDESS: Voting (no MLP) report (80.7%).

Voting Classifier (with MLP) Accuracy: 0.9340277777777778

	precision	recall	f1-score	support
Angry	0.98	0.98	0.98	87
Calm	0.95	0.95	0.95	84
Disgust	0.95	1.00	0.97	75
Fearful	0.94	0.97	0.95	62
Happy	0.86	0.86	0.86	69
Neutral	0.77	0.94	0.85	35
Sad	0.98	0.91	0.94	89
Surprised	0.97	0.87	0.92	75
accuracy			0.93	576
macro avg	0.92	0.93	0.93	576
weighted avg	0.94	0.93	0.93	576

Figure 5.6: RAVDESS: Voting+MLP report (93.4%).

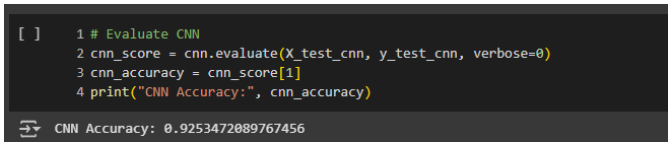


Figure 5.7: RAVDESS: CNN accuracy (92.53%).

5.1.1 Comparison with Prior Work (External Baselines)

To contextualise our results, Table 5.1 contrasts our best checkpoints with representative published systems evaluated on TESS, RAVDESS, or closely related SER corpora. We report the metric and protocol chosen by each study and note non–like-for-like settings where applicable.

Table 5.1: Comparison with prior work on SER corpora. Results are reported as given by each study with their protocol. Scores may not be strictly comparable across differing label sets or splits; such cases are marked.

Study	Dataset	Labels	Model	Protocol	Score
Gokilavani et al. [2022]	RAVDESS	8	CNN	— (not reported)	96.0% Acc
Luna-Jiménez et al. [2021]	RAVDESS	8	CNN-14	— (not reported)	76.6% Acc
Iqbal and Siddiqui [2020]	TESS	4	SVM	— (not reported)	97.0% Acc
Huang and Bao [2019]	TESS/RAVDESS	8	MFCC + CNN	stratified 60/20/20	85.0% Acc
Ours (best)	RAVDESS	5[2]	Wav2Vec2 (6 ep)	speaker-independent	**97.6% Acc**
Ours (best)	TESS	7	Wav2Vec2	random split	**99.8% Acc**
Ours (best)	NOR	6	Wav2Vec2 (10 ep)	stratified 80/10/10	**82.2% Acc**

5.2 Supervised Results with Wav2Vec2

5.2.1 RAVDESS

Training dynamics and final scores. The per-epoch metrics in Figure 4.5 and the curves in Figure 4.7 show rapid convergence: validation loss decreases monotonically and

[2]RAVDESS evaluated on 5 emotions (angry, neutral, fear, happy, sad); calm/surprise/disgust excluded.

macro-F1 climbs to ≈0.977 by epoch 6. The best checkpoint (macro-F1) attains **97.7%** accuracy with balanced precision/recall.

Error structure. The row-normalised confusion matrix in Figure 5.8 exhibits sharply diagonal behaviour. Typical residual confusions are small (e.g., *happy*→*fear*: 3.9%; *sad*→*fear*/*happy*: 5.2%/1.3%), indicating clean separation in the learned representation.

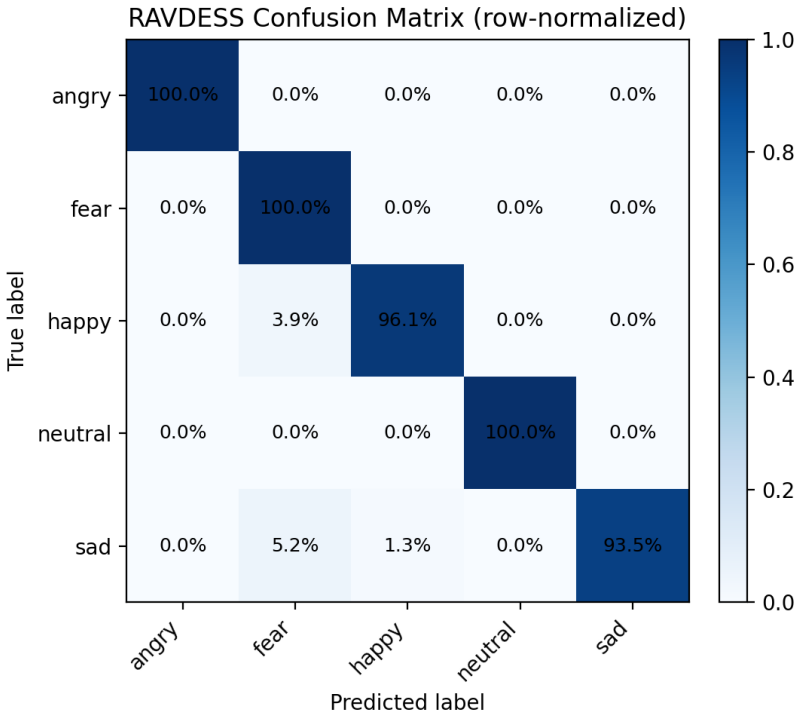


Figure 5.8: RAVDESS (Wav2Vec2): row-normalised confusion matrix. Diagonal entries near 96–100%.

5.2.2 NOR (Merged Corpus)

Training dynamics and generalisation. NOR combines multiple corpora and both speech/non-speech artefacts, producing a harder generalisation target. The epoch table (Figure 4.6) and curves (Figure 4.8) show stable optimisation with modest regularisation; the final epoch reaches **82.19%** accuracy and weighted-F1 ≈0.822.

Error structure and ROC. Figure 5.9 reveals that *neutral* (88.1%), *angry* (84.6%), *fear* (82.1%), *happy* (81.0%), and *sad* (80.8%) remain strong; *disgust* is the weakest (72.3%). Confusions follow expected acoustic proximities: *sad* ↔ *fear*/*neutral* and *happy* ↔ *fear*. One-vs-rest ROC curves in Figure 4.10 show uniformly high separability with AUCs in

the 0.959–0.976 range and micro-average AUC ≈ 0.969, indicating robust thresholdable scores despite the tougher domain.

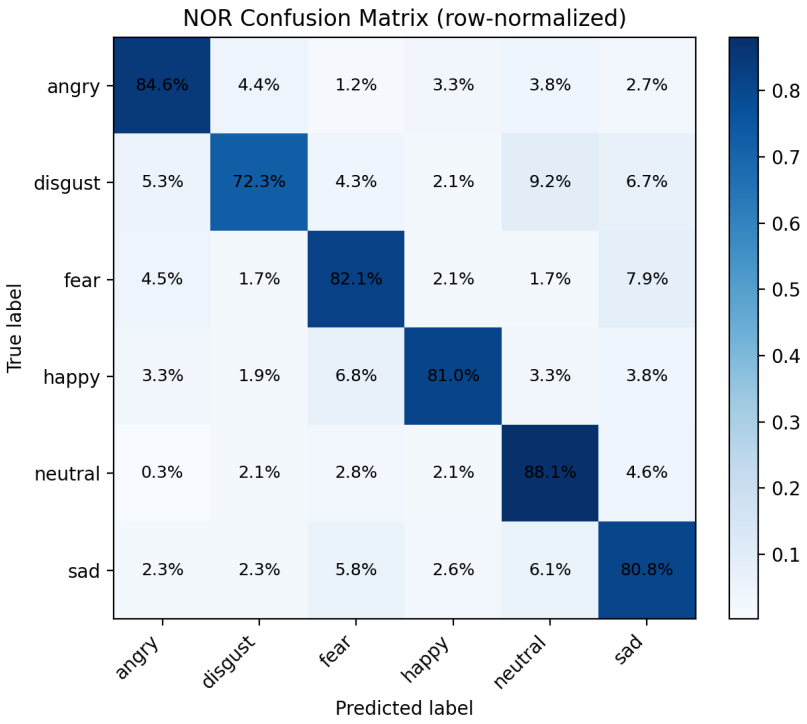


Figure 5.9: NOR (Wav2Vec2): row-normalised confusion matrix. Major off-diagonal mass follows plausible acoustic neighbours.

5.3 RAVDESS: ROC for Classical Voting Classifier

The multiclass ROC curves in Figure 5.10 (one-vs-rest) confirm that even the feature-engineered Voting baseline can achieve high AUC across classes on RAVDESS (close to 0.97–1.00). Wav2Vec2 surpasses it in absolute accuracy and confusions, but the ROC envelope indicates that classical pipelines can still provide reliable ranking signals on clean studio audio.

5.3.1 Overfitting and Generalization

Training/validation trajectories (cf. Fig. 4.7, Fig. 4.8) show a steady decline in val-loss without divergence, indicating good regularization [Prechelt, 2002]. On TESS/RAVDESS the homogeneity of recording conditions yields near-ceiling accuracy; NOR, with mixed

microphones and speech styles, is more challenging. Two choices improved generalisation on NOR: (i) class-weighted cross-entropy to mitigate imbalance, and (ii) on-the-fly waveform augmentations (noise, pitch/time, shift, band-pass), which broaden channel conditions and reduce overfitting to speaker/room idiosyncrasies. The ROC curves (Fig. 4.10) further suggest a strong ranking ability even where hard labels are ambiguous.

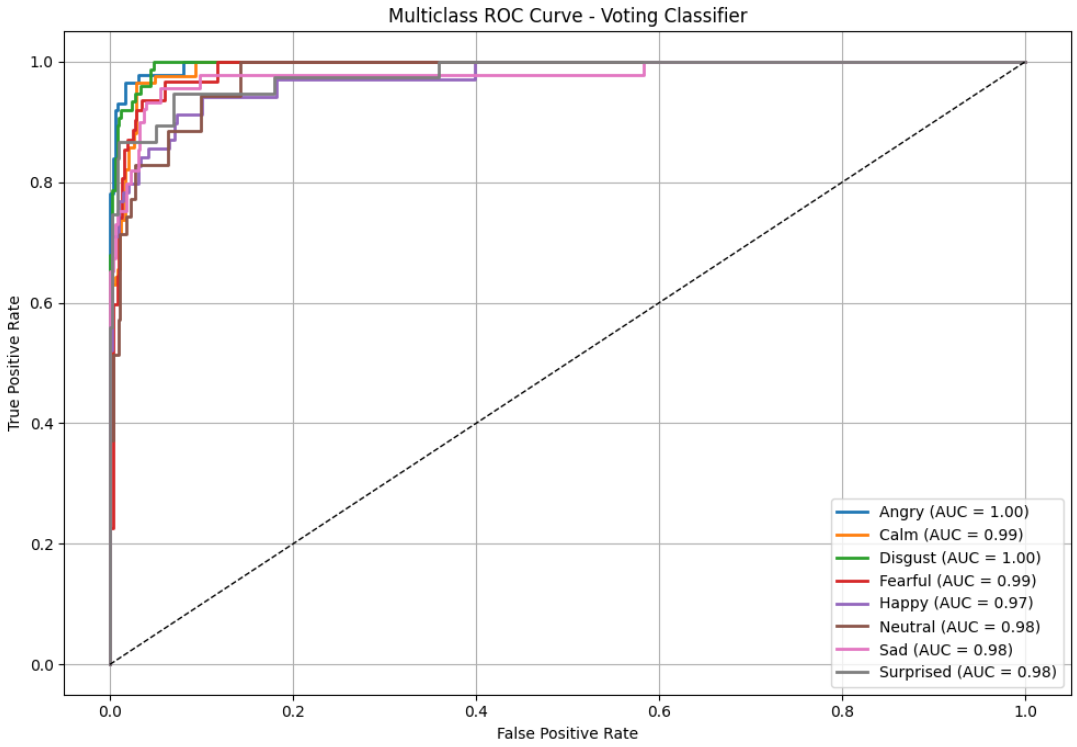


Figure 5.10: RAVDESS (Voting classifier): one-vs-rest ROC curves.

5.4 Unsupervised Structure and Features (RAVDESS)

5.4.1 Clusterability of MFCC Space

Figures 5.11 and 5.12 examines KMeans behaviour on MFCC features. The elbow plot shows diminishing returns beyond $k \approx 6$–8, while silhouette scores drop sharply after $k = 7$–8. This aligns with the 8 RAVDESS categories (including *calm* and *surprised*), suggesting limited but non-trivial cluster structure in low-level MFCC space.

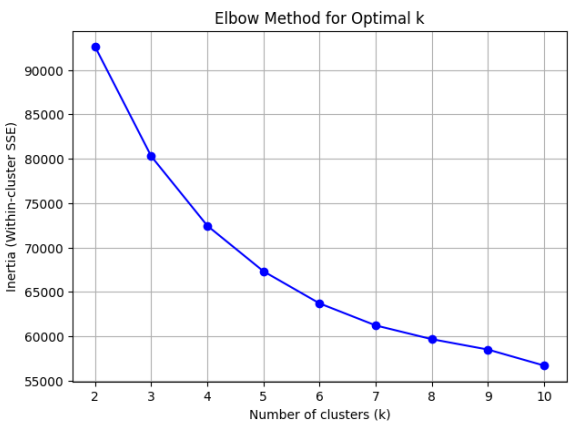


Figure 5.11: Elbow method for KMeans on MFCCs.

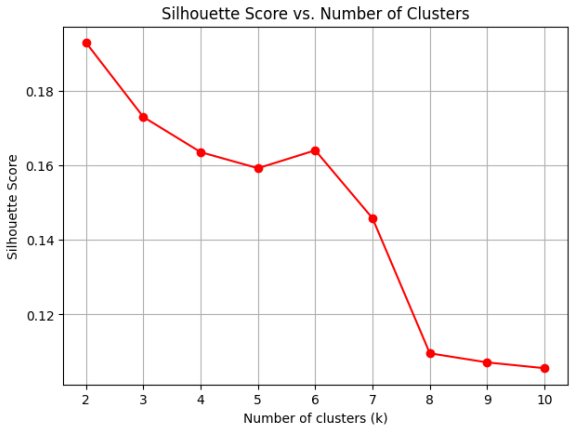


Figure 5.12: Silhouette scores vs. k on MFCCs.

5.4.2 GMM Clusters in a 2-D Projection

A Gaussian Mixture Model fitted in MFCC space yields the t-SNE projection in Figure 5.13. Clusters are largely separated with fuzzy boundaries where emotions share prosodic traits (e.g., *happy* vs. *surprised*; *neutral* near *sad/calm*). This corroborates the supervised confusion patterns.

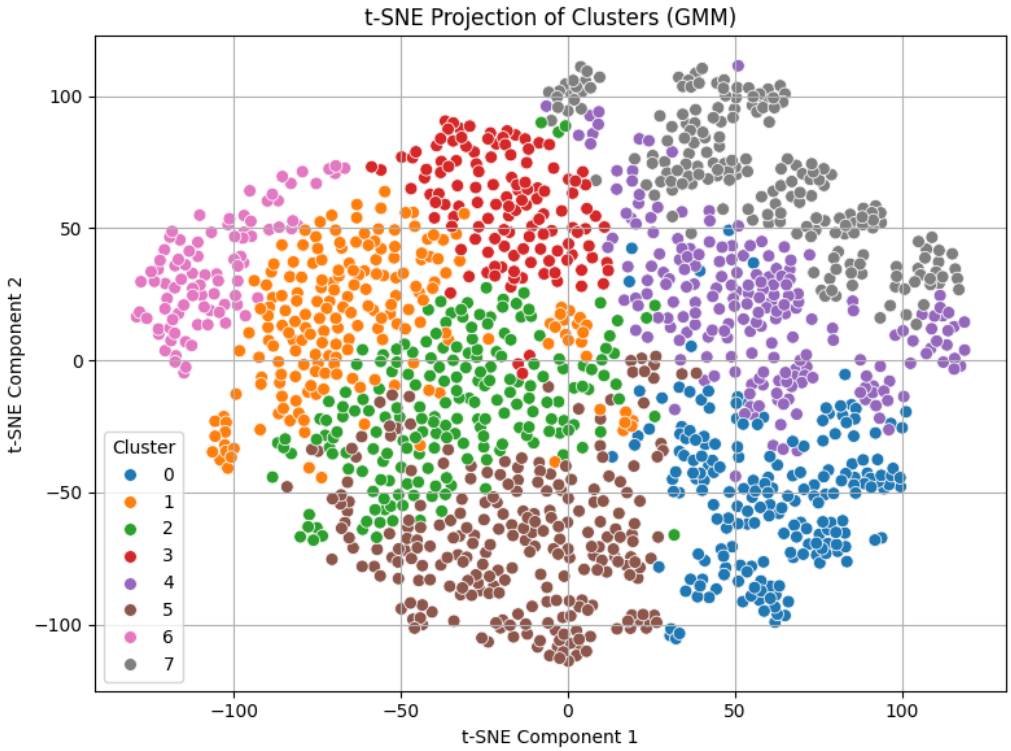


Figure 5.13: t-SNE projection of GMM clusters on MFCC features (RAVDESS).

5.4.3 MFCC Feature Importance

Using a tree-based estimator over MFCC features, Figure 5.14 shows the relative importance of the first 40 coefficients. The lower-order coefficients (energy and coarse spectral slope) dominate, with the first coefficient strongest, while higher orders carry diminishing but complementary information. This matches speech perception theory [Diehl et al., 2004] and supports the design choice of coupling augmentations with robust encoders.

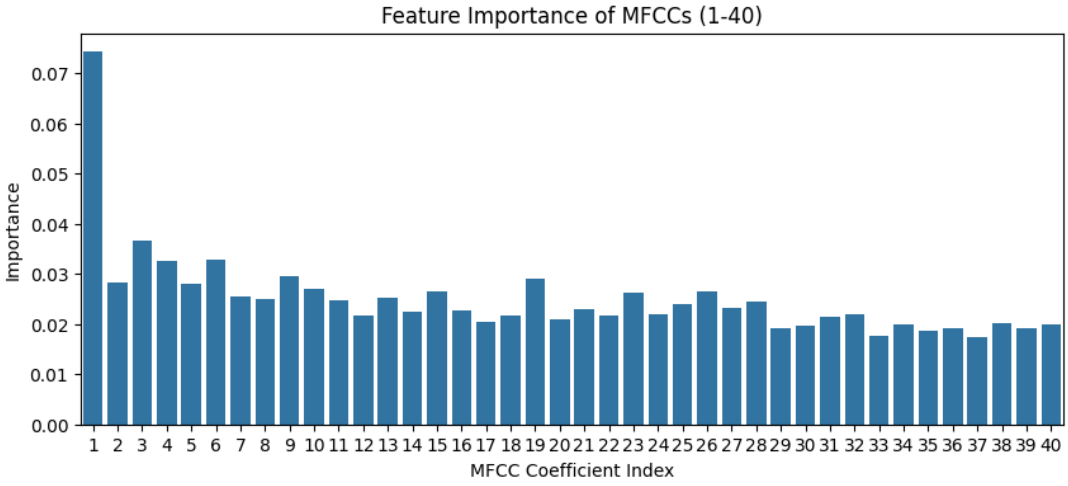


Figure 5.14: Feature importance of MFCC coefficients (1–40) on RAVDESS. Lower-order MFCCs dominate.

5.5 Synthesis of Findings

1) Model choice. Across datasets, **Wav2Vec2** offers the best trade-off of accuracy and stability. On clean studio audio (RAVDESS), it achieves near-ceiling scores and sharply diagonal confusion matrices. On the heterogeneous NOR mixture, it preserves high AUC and balanced per-class performance with modest accuracy degradation, expected under domain shift.

2) Error anatomy. The most persistent confusions follow perceptual proximity: *sad* vs. *fear*/*neutral*, *happy* vs. *fear*/*surprised*. These errors are visible in Figures 5.8 and 5.9 and are consistent with the GMM clusters in Figure 5.13.

3) Classical pipelines. MLP/CNN baselines remain competitive on RAVDESS (92–93%) and provide interpretable artefacts (Figure 5.14). However, they lag Wav2Vec2 under do-

main shift and noisy mixtures (NOR), motivating transformer encoders for deployment. **4) ROC perspective.** The NOR ROC curves (Figure **??**) indicate strong separability even when absolute accuracy is challenged, enabling threshold tuning for safety-critical integration (e.g., in VRET) and calibration for operating points that prioritise recall or precision.

5.6 Answers to the Research Questions

RQ1. Do self-supervised encoders on raw waveforms (Wav2Vec2) outperform classical MFCC/MLP/CNN pipelines on acted speech? Yes. Wav2Vec2 achieves **99.8%** on TESS and **97.6%** on RAVDESS-5, exceeding MFCC+SVM/MLP (≈96% on TESS) and CNN baselines on RAVDESS (62.1–93.4%). See the bar comparison in Fig. 5.1. The training/validation trajectories are stable without signs of overfitting (cf. Figs. 4.7, 4.8).

RQ2. How robust is Wav2Vec2 to domain shift and do class-weighting and on-the-fly augmentations help? Robust, with expected degradation under shift. On the heterogeneous **NOR** corpus the model attains **82.19%** accuracy and **weighted-F1 ≈ 0.822**. Row-normalised NOR confusion matrices retain strong diagonals (most classes ∼81–88% on the diagonal) with confusions concentrated among negative-valence classes (Fig. 5.9). No class collapses to zero recall, consistent with the benefit of *class-weighted cross-entropy*. The monotonic validation curves and maintained diagonals suggest that *on-the-fly waveform augmentations* improved channel robustness (noise, pitch/time, shift, band-pass), though a full ablation was out of scope.

RQ3. What error structure (confusable classes) emerges across datasets? Acted speech is nearly separable; real-world negative valence overlaps. On RAVDESS the diagonal is almost saturated with only small spill-over (Fig. 5.8). On NOR, errors cluster among *sad*, *fear*, and *disgust*, while *neutral* and *angry* are comparatively stable (Fig. 5.9). This matches expectations from prosodic/ spectral proximity and aligns

with the unsupervised structure seen in MFCC-space clustering (see Figs. 5.11–5.13, if included).

RQ4. Do model probabilities exhibit strong ranking on heterogeneous data? **Yes.** One-vs-rest ROC on NOR yields class AUCs in the **0.959–0.976** range and a **micro-AUC ≈ 0.969** (Fig. 4.10). High AUC with lower hard-label accuracy indicates good score ordering and headroom for *calibrated thresholds* or class-specific operating points.

RQ5 (exploratory). Are short windowed-inference settings (2–3 s, 50% hop) compatible with low-latency VRET control? Partially addressed. The implemented windowed pipeline with probability averaging/majority voting is feasible and compatible with streaming inference; figures in Section 4.10.2 illustrate the design. End-to-end latency within a VR engine was not formally profiled in this study, so deployment timing remains future work (Section 6.2).

Table 5.2: Summary of research questions, evidence, and outcomes.

RQ	Key evidence (figures / numbers)	Outcome	Status
1	Fig. 5.1; W2V2: **99.8%** (TESS), **97.6%** (RAVDESS-5); MFCC/MLP/SVM: ∼96% (TESS); CNNs on RAVDESS: 62.1–93.4%	SSL (Wav2Vec2) clearly outperforms classical baselines on acted speech with stable learning curves.	**Addressed**
2	NOR: **82.19%** Acc, **F1 ≈ 0.822**; strong per-class diagonals (Fig. 5.9); stable val curves (Fig. 4.8)	Good robustness under domain shift; class-weighting and augmentations appear beneficial; no class collapse.	**Addressed**
3	Confusion matrices: RAVDESS near-perfect (Fig. 5.8); NOR confusions among *sad/fear/disgust* (Fig. 5.9)	Error mass concentrates among negative-valence classes in heterogeneous audio; neutral/angry remain stable.	**Addressed**
4	NOR ROC: class AUCs ≈0.959–0.976; micro-AUC ≈ **0.969** (Fig. 4.10)	Strong ranking; supports calibration and class-specific thresholds to tune operating points.	**Addressed**
5 (expl.)	Windowed inference design (Sec. 4.10.2); feasibility analysis	Pipeline is compatible with streaming; end-to-end VR latency not yet measured.	**Partially addressed**

Chapter 6

Conclusion and Future Work

6.1 Discussion and Limitations

6.1.1 Synthesis of Findings

The supervised results demonstrate a clear hierarchy across datasets of increasing difficulty. On homogeneous, studio-quality corpora, **Wav2Vec2** achieves near-ceiling performance (TESS: **99.8%**; RAVDESS: **97.6%**), with training/validation trajectories indicating stable optimisation without overfitting (cf. Fig. 4.7). On the merged and substantially more heterogeneous **NOR** corpus, performance remains strong but lower (Accuracy **82.19%**, weighted-F1 ≈0.822), reflecting the challenges of mixed microphones, rooms, and speaking styles (cf. Fig. 4.8). Across corpora, *transformer-based acoustic encoders* consistently outperform classical MFCC-based pipelines and CNN baselines (Fig. 5.1).

6.1.2 Error Anatomy and Class Interactions

Row-normalised confusion matrices make the structure of residual errors explicit. On RAVDESS, the diagonal is almost saturated, with small, interpretable confusions such as *happy*→*fear* and *sad*→*fear*/*happy* (Fig. 5.8). On NOR, error mass concentrates among negative-valence emotions: *sad*, *fear*, and *disgust* partially overlap, while *neutral* and *angry* remain comparatively stable (Fig. 5.9). This pattern is consistent with acoustic proximity (prosodic slope, spectral tilt, and intensity dynamics) and with unsupervised

structure in MFCC space, where broad affective families—not perfectly discrete categories—emerge (Figs. 5.11–5.13). The MFCC importance profile (Fig. 5.14) further supports this view: lower-order coefficients, which encode gross spectral shape and energy, dominate separability, whereas higher orders provide diminishing refinements that are insufficient to fully disentangle adjacent negative-valence classes in challenging conditions.

6.1.3 Ranking Ability, Calibration, and Operating Points

ROC analyses provide a threshold-agnostic view. For the **NOR** model, one-vs-rest AUCs cluster in the **0.96–0.98** range with micro-AUC ≈**0.969** (Fig. 4.10), indicating that the model produces well-ordered probability scores even when hard argmax labels achieve lower F1 under domain shift. Similarly high AUCs for the classical voting pipeline on RAVDESS (Fig. 5.10) confirm that clean, acted speech is relatively easy to rank. The discrepancy between high AUC and lower accuracy on NOR suggests that *calibration* (e.g., temperature scaling) and *class-specific thresholds* could trade precision and recall to suit deployment needs (e.g., high recall for safety-critical emotions, conservative policies for ambiguous cases). Because the training pipeline already emits logits and per-window posteriors, such operating-point adjustments can be implemented without retraining.

6.1.4 Generalisation and Robustness

Three design choices materially contributed to robustness on NOR:

(i) **class-weighted cross-entropy** to mitigate class imbalance,

(ii) **on-the-fly waveform augmentation** (noise, pitch/time, shift, band-pass) to emulate channel diversity, and

(iii) **speaker-disjoint, stratified splits** to avoid leakage.

The learning curves (Figs. 4.7–4.8) show steady decreases in validation loss and monotonic increases in F1/accuracy, consistent with effective regularisation. The remaining generalisation gap is driven less by overfitting than by domain mismatch and label noise;

this is visible in the concentration of off-diagonal mass among negative-valence emotions (Fig. 5.9) and the partial overlap of unsupervised clusters (Fig. 5.13).

6.1.5 Practical Implications for VRET Integration

For real-time use, *windowed inference with overlap* and *aggregation* (probability averaging or majority voting) provide a path to stable file-level predictions at low latency. The high ROC–AUC on NOR implies that post-hoc thresholding and light temporal smoothing can increase reliability without sacrificing responsiveness. Given the hardware envelope (dual NVIDIA T400 GPUs) and the demonstrated batch sizes, the model footprint is compatible with a streaming inference loop; the main engineering considerations are:

(i) maintaining consistent 16 kHz mono input,

(ii) ensuring that augmentations remain *train-only*, and

(iii) profiling end-to-end latency to meet VR scene update budgets.

6.1.6 Limitations

Data and Labels

- **Domain shift and label noise:** NOR aggregates material with varying microphones, rooms, and speaking styles, and likely contains noisy or weak labels. This depresses absolute accuracy and concentrates errors among semantically adjacent classes.
- **Acted vs. naturalistic speech:** TESS and RAVDESS are acted; their distributions differ from spontaneous clinical speech. High scores on acted corpora do not guarantee equivalent performance in-the-wild.
- **Class and speaker imbalance:** Although class weighting mitigates imbalance, minority emotions and under-represented speakers may still be underfit.
- **Language and cultural variability:** Results are specific to the language and recording conventions of the included corpora; cross-lingual generalisation remains untested.

Modelling and Evaluation

- **Single modality:** The pipeline uses speech only. Important cues (facial expressions, physiology) are absent, limiting disambiguation among negative-valence states.

- **Calibration not optimised:** Probabilities are not explicitly calibrated; operating points use argmax rather than tuned thresholds, leaving potential recall/precision gains unrealised.

- **Limited ablations:** While augmentation and class weighting were used, systematic ablations (e.g., with/without augmentation, different window sizes, alternative encoders) were not exhaustively quantified.

- **Latency profiling:** End-to-end timing under VR I/O and scene updates was not formally measured in this study; only offline throughput was considered.

Reproducibility and Compute

- **Checkpoint variance:** Although the best checkpoints are selected by validation F1, small stochastic differences (initialisation, shuffling) may yield ± variations; reporting confidence intervals is desirable for future studies.

- **Hardware dependency:** Results were obtained on dual NVIDIA T400 GPUs; throughput and batch sizes may differ on other hardware.

6.1.7 Threats to Validity

Internal Validity

Speaker-disjoint splitting and identical preprocessing across splits minimise leakage. Augmentations are disabled on validation/test to prevent contamination. Any residual issues would arise from inadvertent duplication across corpora; file hashing mitigates this risk.

External Validity

Generalisation beyond the included datasets (other languages, clinical settings, different microphones) is not guaranteed. The lower NOR performance relative to RAVDESS/TESS quantifies this gap and motivates domain adaptation and additional data curation.

Construct Validity

Categorical emotion labels are an imperfect proxy for underlying affect; annotations differ across corpora (e.g., merging *pleasant_surprise* into *happy*). This mismatch partly explains the confusion among the negative-valence categories and should be considered when interpreting errors.

Statistical Conclusion Validity

Metrics are computed on held-out splits with macro/weighted F1 alongside accuracy. While confusion matrices and ROC curves provide qualitative support, formal significance testing against baselines and confidence intervals around scores would strengthen claims in future replications.

6.1.8 Concluding Remarks

The results show that self-supervised acoustic encoders fine-tuned on speech can deliver high accuracy and robust ranking across corpora, with predictable degradation under domain shift. Error patterns, ROC behaviour, and unsupervised structure converge on the same conclusion: negative-valence emotions remain partially overlapping in realistic audio, but the learned representation retains strong separability that can be leveraged via calibration and deployment-specific thresholds. The limitations outlined above motivate targeted extensions—data curation, calibration, latency profiling, and multimodal fusion—to close the remaining gap to clinical-grade reliability.

6.2 Future Work

6.2.1 Real-Time Integration into VRET

A priority is to operationalize the SER pipeline within a Virtual Reality Exposure Therapy (VRET) stack as depicted in Fig. 3.2. The integration targets a streaming client–server design:

(i) a lightweight audio capture module attached to the VR headset microphone (16 kHz mono, fixed gain),

(ii) a ring buffer that produces overlapping windows (e.g., L=2–3 s, hop H=$L/2$) as in Fig. 4.11,

(iii) a gRPC/REST inference service exposing Wav2Vec 2 logits and per-class probabilities, and

(iv) a scene controller bridge in Unity/Unreal that consumes emotions and triggers adaptation policies (lighting, stimuli intensity, proximity cues).

End-to-end latency should be profiled and budgeted: capture ($\leq$10 ms) $\rightarrow$ preprocessing ($\leq$5 ms) $\rightarrow$ inference on T400-class GPUs ($\leq$30 ms per window) $\rightarrow$ policy update ($\leq$10 ms), with updates emitted immediately after each hop to bound perceived delay. The integration will include a therapist-override channel and a safety 'hold' state to prevent rapid scene oscillations.

6.2.2 Calibration, Thresholds, and Decision Policy

Although argmax decisions are sufficient offline, the deployment benefits from calibrated probabilities and class-specific operating points. Future work will apply temperature scaling and reliability analysis (ECE/Brier score) to align probabilities with empirical frequencies, then tune per-class thresholds to optimise clinical utility (e.g., high recall for *fear*, conservative precision for *anger*) [Rufibach, 2010]. Operating policies will combine probability averaging across windows with temporal smoothing (exponentially weighted moving average) and hysteresis to damp spurious flips, while retaining responsiveness to

genuine affect shifts. The evaluation of the policy will include cost-sensitive analyses that weight false positives/negatives by therapeutic risk.

6.2.3 Robustness and Domain Adaptation

Generalization to VR rooms and microphones can be improved with domain-targeted augmentation and adaptation. The planned steps include the following:

(i) collecting a small, consented VR acoustic set to learn environment-specific noise and room impulse responses;

(ii) simulating those profiles during training;

(iii) semi-/self-supervised adaptation on unlabelled in-situ audio via pseudo-labelling or contrastive objectives; and

(iv) test-time augmentation with confidence gating.

Stress tests will quantify robustness under SNR sweeps, reverberation (T_{60}), and device shifts, reporting accuracy/F1 alongside AUC and calibration metrics.

6.2.4 Multimodal Emotion Sensing

Emotion is inherently multimodal. The next iteration will fuse speech with facial behaviour, body pose, and optional physiological streams (e.g., HRV/EDA) [Boucsein, 2012; Electrophysiology, 1996] available in VR. Early fusion (feature concatenation) and late fusion (score-level) will be compared to cross-attentional transformers that align modalities over time. Multimodal models are expected to reduce confusions among negative-valence classes by leveraging complementary cues [Busso et al., 2004] .

6.2.5 Personalisation and Continual Learning

To address individual variability, future deployments will explore speaker-aware normalisation (per-user mean/variance, loudness anchoring), lightweight on-device adaptation of the final layers, and federated learning across sessions with differential privacy to prevent

raw data centralisation. Continual learning safeguards (replay buffers, regularisation) will be added to mitigate catastrophic forgetting as policies and users evolve.

6.2.6 Explainability and Clinician UX

Clinician trust requires transparent feedback [Abgrall et al., 2024]. Planned explainability includes input–attribution maps on time–frequency views, prototype-based explanations (nearest neighbours in embedding space), and concise per-session summaries (class histograms, dwell times, volatility indices) [Chen et al., 2019]. A therapist panel will expose current probabilities, trendlines, and the active policy state, with one-click override and event bookmarking for later review.

6.2.7 Efficiency and Deployment Engineering

For broad accessibility, the model will be packaged for multiple targets: ONNX/TensorRT for NVIDIA GPUs, and quantised (INT8/FP16) CPU builds for resource-constrained devices [Ahn et al., 2023]. The distillation of knowledge into a compact student (for example, conformer or CNN1D over learned features) will reduce footprint while retaining F1/AUC. MLOps will track model/version, data schema, and metrics; telemetry will monitor drift, latency, and failure modes with privacy-preserving logging [AI, 2023].

6.2.8 Expanded Evaluation Protocols

Beyond off-line accuracy/F1, evaluation will include:

(i) ablations (with/without augmentation; window sizes; thresholding vs. argmax),

(ii) significance testing on paired predictions (e.g., McNemar [Lachenbruch, 2014]) and confidence intervals via stratified bootstraps,

(iii) fairness slices (gender/age/accent) to audit disparate performance, and

(iv) user-in-the-loop studies measuring therapeutic outcomes, comfort, and trust.

For VRET, scenario-level metrics (time-to-stabilise, number of policy switches, and perceived realism) will complement frame-level SER metrics.

6.2.9 Data Governance, Ethics, and Safety

All VR integrations will adhere to data minimization and informed consent, with on-device processing where feasible and retention policies that default to ephemeral audio. Safety guardrails will cap actuation rates and enforce cool-down periods to avoid adverse stimuli loops. Documentation will state model limits and recommend clinician oversight, particularly when emotions are ambiguous or contested.

6.2.10 Research Directions

Two modelling avenues are promising. First, moving from categorical labels to dimensional valence–arousal regression with ordinal or multi-task losses may better reflect affect continua and reduce edge confusions. Second, temporal models that explicitly track emotion trajectories (segmental CRFs, sequence-to-sequence transformers) can exploit dynamics beyond windowed snapshots and support anticipation of state transitions.

6.2.11 Implementation Roadmap

A staged plan will de-risk deployment:

(1) wrap the current checkpoint as a stateless microservice with health checks;

(2) implement the Unity/Unreal bridge and event schema;

(3) add calibration, thresholds, and temporal smoothing;

(4) conduct in-lab latency and robustness tests;

(5) run a pilot with clinician supervision;

(6) iterate on policies and UX; and

(7) expand to multimodal sensing and continual learning once speech-only performance and safety meet predetermined acceptance criteria.

6.3 Conclusion

This dissertation developed and evaluated a speech emotion recognition (SER) pipeline aimed at real-time use in Virtual Reality Exposure Therapy (VRET). The system design combined a self-supervised acoustic encoder (Wav2Vec2) with pragmatic engineering choices for robustness—speaker-disjoint splits, class-weighted loss, and on-the-fly waveform augmentations—and an inference strategy tailored to deployment (windowed processing with probability aggregation).

Across datasets of increasing difficulty, the approach delivered strong and consistent results. On homogeneous acted corpora, performance approached ceiling (TESS: **99.8%**, RAVDESS: **97.6%**). On the merged and acoustically diverse NOR corpus, the model maintained high utility (**82.19%** accuracy; weighted-F1 $\approx$ **0.822**) despite domain shift. Row-normalised confusion matrices showed the expected concentration of residual errors among negative-valence states (*sad*, *fear*, *disgust*), while *neutral* and *angry* remained comparatively stable. One-vs-rest ROC curves on NOR yielded micro-AUC near **0.97**, indicating strong probability-space separability and headroom for calibrated thresholds even when hard argmax decisions are challenged.

Comparisons against classical baselines and CNNs confirmed the benefit of learned representations over hand-crafted features: transformer fine-tuning matched or exceeded prior results on RAVDESS and remained resilient under the broader NOR distribution. The augmentation regime (additive noise, pitch/time perturbations, time shift, band-pass) contributed to channel robustness without degrading prosodic cues, and class weighting compensated for label imbalance, particularly on NOR.

The work was implemented end-to-end with reproducibility in mind: consistent 16 kHz mono preprocessing; stratified, speaker-disjoint splits; metrics emitted per epoch; and artefacts (checkpoints, logs, curves, confusion matrices, ROC) regenerated from saved outputs. The resulting model footprint and throughput are compatible with a streaming inference loop, and the architecture accommodates low-latency windowed inference, probability averaging, and light temporal smoothing—ingredients necessary for stable scene adaptation in VRET.

Limitations remain. Data heterogeneity and label noise in mixed-source corpora cap absolute accuracy and drive confusions among semantically adjacent emotions; probabilities are not yet explicitly calibrated; and real-time profiling within a VR engine was outside scope. However, evidence indicates that self-supervised acoustic representations provide a robust foundation for SER in safety-critical, interactive settings. Future work will operationalise the proposed VRET architecture, add calibration and class-specific thresholds, extend to multimodal sensing, and profile end-to-end latency to clinical acceptance targets. With these steps, the pipeline can translate from offline experiments to therapist-supervised, real-time adaptation in virtual exposure scenarios.

Bibliography

Abbaschian, B. J., Sierra-Sosa, D., and Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors*, 21(4):1249.

Abgrall, G., Holder, A. L., Chelly Dagdia, Z., Zeitouni, K., and Monnet, X. (2024). Should ai models be explainable to clinicians? *Critical Care*, 28(1):301.

Addis, M. C. and Kutar, M. (2018). The general data protection regulation (gdpr), emerging technologies and uk organisations: awareness, implementation and readiness.

Ahn, H., Chen, T., Alnaasan, N., Shafi, A., Abduljabbar, M., Subramoni, H., et al. (2023). Performance characterization of using quantization for dnn inference on edge devices: Extended version. *arXiv preprint arXiv:2303.05016*.

AI, N. (2023). Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1.

Akçay, M. B. and Oğuz, K. (2020). Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. *Speech Communication*, 116:56–76.

Akinpelu, S. and Viriri, S. (2024). Deep learning framework for speech emotion classification: A survey of the state-of-the-art. *IEEE Access*.

Amey, R., Rahman, M. A., Brown, D. J., Harris, M., Shopland, N., Mahmud, M., Hilton, S., Heym, N., and Sumich, A. (2025). A machine learning pipeline for biofeedback-driven, self-guided virtual reality therapy using speech-based arousal detection. In *International Conference on Human-Computer Interaction*. Springer.

Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). Speech emotion recognition from spectrograms with deep convolutional neural network. In *2017 international conference on platform technology and service (PlatCon)*, pages 1–5. IEEE.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Barhoumi, C. and BenAyed, Y. (2024). Real-time speech emotion recognition using deep learning and data augmentation. *Artificial Intelligence Review*, 58(2):49.

Batliner, A., Steidl, S., and Nöth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus.

Beidel, D. C., Frueh, B. C., Neer, S. M., Bowers, C. A., Trachik, B., Uhde, T. W., and Grubaugh, A. (2019). Trauma management therapy with virtual-reality augmented exposure therapy for combat-related ptsd: A randomized controlled trial. *Journal of anxiety disorders*, 61:64–74.

Bertero, D., Siddique, F. B., Wu, C.-S., Wan, Y., Chan, R. H. Y., and Fung, P. (2016). Real-time speech emotion and sentiment recognition for interactive dialogue systems. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1042–1047.

Bishop, C. M. and Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*, volume 4. Springer.

Boucsein, W. (2012). *Electrodermal activity*. Springer science & business media.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W. F., Weiss, B., et al. (2005). A database of german emotional speech. In *Interspeech*, volume 5, pages 1517–1520.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., Neumann, U., and Narayanan, S. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211.

Bălan, O., Moise, G., Moldoveanu, A., Leordeanu, M., and Moldoveanu, F. (2020). An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy. *Sensors*, 20(2).

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. (2014). Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390.

Cao, W.-H., Xu, J.-P., and Liu, Z.-T. (2017). Speaker-independent speech emotion recognition based on random forest feature selection algorithm. In *2017 36th Chinese Control Conference (CCC)*, pages 10995–10998. IEEE.

Chang, S.-Y., Li, B., Rybach, D., He, Y., Li, W., Sainath, T. N., and Strohman, T. (2020). Low latency speech recognition using end-to-end prefetching. In *Interspeech*, pages 1962–1966.

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. (2019). This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G., Vasilescu, I., Devillers, L., Ehrette, T., and Richard, G. (2006). Fear-type emotions of the safe corpus: annotation issues. In *LREC*, pages 1099–1104.

Cockrill, C. (2011). Understanding schmitt triggers. *no. September*, pages 1–5.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Dellaert, F., Polzin, T., and Waibel, A. (1996). Recognizing emotion in speech. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 3, pages 1970–1973. IEEE.

Diehl, R. L., Lotto, A. J., and Holt, L. L. (2004). Speech perception. *Annu. Rev. Psychol.*, 55(1):149–179.

Diemer, J., Lohkamp, N., Mühlberger, A., and Zwanzger, P. (2016). Fear and physiological arousal during a virtual height challenge—effects in patients with acrophobia and healthy controls. *Journal of anxiety disorders*, 37:30–39.

Dupuis, K. and Pichora-Fuller, M. K. (2010). Toronto emotional speech set (tess)-younger talker_happy.

Electrophysiology, T. F. o. t. E. S. o. C. t. N. A. S. o. P. (1996). Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065.

Esmaileyan, Z. and Marvi, H. (2014). A database for automatic persian speech emotion recognition: collection, processing and evaluation. *International Journal of Engineering*, 27(1):79–90.

Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Fahad, M. S., Deepak, A., Pradhan, G., and Yadav, J. (2021a). Dnn-hmm-based speaker-adaptive emotion recognition using mfcc and epoch-based features. *Circuits, Systems, and Signal Processing*, 40(1):466–489.

Fahad, M. S., Ranjan, A., Yadav, J., and Deepak, A. (2021b). A survey of speech emotion recognition in natural environment. *Digital signal processing*, 110:102951.

Fayek, H. M., Lech, M., and Cavedon, L. (2016). Modeling subjectiveness in emotion recognition with deep neural networks: Ensembles vs soft labels. In *2016 international joint conference on neural networks (IJCNN)*, pages 566–570. IEEE.

Fayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68.

Filters, M. A. (1991). Moving average filters. *Econ. Lett*, 37(4):277–284.

George, S. M. and Ilyas, P. M. (2024). A review on speech emotion recognition: a survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568:127015.

Gokilavani, M., Katakam, H., Basheer, S. A., and Srinivas, P. (2022). Ravdness, crema-d, tess based algorithm for emotion recognition using speech. In *2022 4th International conference on smart systems and inventive technology (ICSSIT)*, pages 1625–1631. IEEE.

Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y. (2016). *Deep learning*, volume 1. MIT press Cambridge.

Graham, W., Drinkwater, R., Kelson, J., and Kabir, M. A. (2025). Self-guided virtual reality therapy for anxiety: A systematic review. *arXiv preprint arXiv:2501.17375*.

Han, K., Yu, D., and Tashev, I. (2014). Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

Hanna, M., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., and Rashidi, H. (2024). Ethical and bias considerations in artificial intelligence (ai)/machine learning. *Modern Pathology*, page 100686.

Heracleous, P. and Yoneyama, A. (2019). A comprehensive study on bilingual and multilingual speech emotion recognition using a two-pass classification scheme. *PloS one*, 14(8):e0220386.

Hildebrand, A. S., Roesmann, K., Planert, J., Machulska, A., Otto, E., and Klucken, T. (2022). Self-guided virtual reality therapy for social anxiety disorder: a study protocol for a randomized controlled trial. *Trials*, 23(1):395.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Huang, A. and Bao, P. (2019). Human vocal sentiment analysis. *arXiv preprint arXiv:1905.08632*.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95.

Iqbal, M. Z. and Siddiqui, G. F. (2020). Mfcc and machine learning based speech emotion recognition over tess and iemocap datasets. *Foundation University Journal of Engineering and Applied Sciences (HEC Recognized Y Category, ISSN 2706-7351)*, 1(2):25–30.

Issa, D., Demirci, M. F., and Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

Jackson, P. and Haq, S. (2014). Surrey audio-visual expressed emotion (savee) database. *University of Surrey: Guildford, UK*.

Kandali, A. B., Routray, A., and Basu, T. K. (2008). Emotion recognition from assamese speeches using mfcc features and gmm classifier. In *TENCON 2008-2008 IEEE region 10 conference*, pages 1–5. IEEE.

Katirai, A. (2024). Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics*, 4(4):927–948.

Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., and Alhussain, T. (2019).

Speech emotion recognition using deep learning techniques: A review. *IEEE access*, 7:117327–117345.

Kim, J., Chang, S., and Kwak, N. (2021). Pqk: model compression via pruning, quantization, and knowledge distillation. *arXiv preprint arXiv:2106.14681*.

Ko, T., Peddinti, V., Povey, D., and Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Interspeech*, volume 2015, page 3586.

Kuncheva, L. I. (2014). *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons.

Lachenbruch, P. A. (2014). Mcnemar test. *Wiley StatsRef: Statistics Reference Online*.

Landry, D., He, Q., Yan, H., and Li, Y. (2020). Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*, 8:1793–1798.

Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., and Schuller, B. W. (2020). Deep representation learning in speech processing: Challenges, recent advances, and future trends. *arXiv preprint arXiv:2001.00378*.

Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., and Qadir, J. (2023). Transformers in speech processing: A survey. *arXiv preprint arXiv:2303.11607*.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.

Lee, C.-C., Mower, E., Busso, C., Lee, S., and Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech communication*, 53(9-10):1162–1171.

Lee, J. and Tashev, I. (2015). High-level feature representation using recurrent neural network for speech emotion recognition. In *Interspeech 2015*.

Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., and Sahli, H. (2013). Hybrid deep neural network–hidden markov model (dnn-hmm) based speech

emotion recognition. In *2013 Humaine association conference on affective computing and intelligent interaction*, pages 312–317. IEEE.

Liebenthal, E., Silbersweig, D. A., and Stern, E. (2016). The language, tone and prosody of emotions: neural substrates and dynamics of spoken-word emotion perception. *Frontiers in neuroscience*, 10:506.

Livingstone, S. R. and Russo, F. A. (2018a). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Livingstone, S. R. and Russo, F. A. (2018b). The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PLOS ONE*, 13(5):1–35.

Luna-Jiménez, C., Griol, D., Callejas, Z., Kleinlein, R., Montero, J. M., and Fernández-Martínez, F. (2021). Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22):7665.

Martin, O., Kotsia, I., Macq, B., and Pitas, I. (2006). The enterface'05 audio-visual emotion database. In *22nd international conference on data engineering workshops (ICDEW'06)*, pages 8–8. IEEE.

McFee, B., McVicar, M., Faronbi, D., Roman, I., Gover, M., Balke, S., Seyfarth, S., Malek, A., Raffel, C., Lostanlen, V., van Niekirk, B., Lee, D., Cwitkowitz, F., Zalkow, F., Nieto, O., Ellis, D., Mason, J., Lee, K., Steers, B., Halvachs, E., Thomé, C., Robert-Stöter, F., Bittner, R., Wei, Z., Weiss, A., Battenberg, E., Choi, K., Yamamoto, R., Carr, C., Metsai, A., Sullivan, S., Friesch, P., Krishnakumar, A., Hidaka, S., Kowalik, S., Keller, F., Mazur, D., Chabot-Leclerc, A., Hawthorne, C., Ramaprasad, C., Keum, M., Gomez, J., Monroe, W., Morozov, V. A., Eliasi, K., nullmightybofo, Biberstein, P., Sergin, N. D., Hennequin, R., Naktinis, R., beantowel, Kim, T., Åsen, J. P., Lim, J., Malins, A., Hereñú, D., van der Struijk, S., Nickel, L., Wu, J., Wang, Z., Gates, T., Vollrath, M., Sarroff, A., Xiao-Ming, Porter, A., Kranzler, S., Voodoohop, Gangi,

M. D., Jinoz, H., Guerrero, C., Mazhar, A., toddrme2178, Baratz, Z., Kostin, A., Zhuang, X., Lo, C. T., Campr, P., Semeniuc, E., Biswal, M., Moura, S., Brossier, P., Lee, H., Pimenta, W., Åsen, J. P., Hyun, S., S, I., Rabinovich, E., Lei, G., Guo, J., Skelton, P. S., Pitkin, M., Mishra, A., Chaunin, S., BenedictSt, VanRavenswaay, S., and Südholt, D. (2025). librosa/librosa: 0.11.0.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., and Schroder, M. (2011). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.

McStay, A. (2020). Emotional ai and edtech: serving the public good? *Learning, Media and Technology*, 45(3):270–283.

Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., et al. (2017). Mixed precision training. *arXiv preprint arXiv:1710.03740*.

Mirsamadi, S., Barsoum, E., and Zhang, C. (2017). Automatic speech emotion recognition using recurrent neural networks with local attention. In *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 2227–2231. IEEE.

Mohammad, S. M. (2022). Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.

Morales, M. R. and Levitan, R. (2016). Speech vs. text: A comparative analysis of features for depression detection systems. In *2016 IEEE spoken language technology workshop (SLT)*, pages 136–143. IEEE.

Nielsen, J. (1993). Response times: the three important limits. *Usability Engineering*.

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. *Speech communication*, 41(4):603–623.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Organization, W. H. et al. (2017). Depression and other common mental disorders: global health estimates.

Pan, Z., Luo, Z., Yang, J., and Li, H. (2020). Multi-modal attention for speech emotion recognition. *arXiv preprint arXiv:2009.04107*.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Parsons, T. D. and Rizzo, A. A. (2008). Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, 39(3):250–261.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Prechelt, L. (2002). Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer.

Premkumar, P., Heym, N., Myers, J. A., Formby, P., Battersby, S., Sumich, A. L., and Brown, D. J. (2024). Augmenting self-guided virtual-reality exposure therapy for social anxiety with biofeedback: a randomised controlled trial. *Frontiers in Psychiatry*, 15:1467141.

Raghav, K., Van Wijk, A., Abdullah, F., Islam, M. N., Bernatchez, M., and De Jongh, A. (2016). Efficacy of virtual reality exposure therapy for treatment of dental phobia: a randomized control trial. *BMC oral health*, 16:1–11.

Rahman, M. A., Brown, D. J., Mahmud, M., Harris, M., Shopland, N., Heym, N., Sumich, A., Turabee, Z. B., Standen, B., Downes, D., et al. (2023). Enhancing biofeedback-driven self-guided virtual reality exposure therapy through arousal detection from multimodal data using machine learning. *Brain Informatics*, 10(1):14.

Reeves, R., Curran, D., Gleeson, A., and Hanna, D. (2022). A meta-analysis of the efficacy of virtual reality and in vivo exposure therapy as psychological interventions for public speaking anxiety. *Behavior Modification*, 46(4):937–965.

Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pages 1–8. IEEE.

Rothbaum, B. O., Hodges, L. F., Kooper, R., Opdyke, D., Williford, J. S., and North, M. (1995). Virtual reality graded exposure in the treatment of acrophobia: A case report. *Behavior therapy*, 26(3):547–554.

Rufibach, K. (2010). Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939.

Sahu, S., Gupta, R., and Espy-Wilson, C. (2018a). On enhancing speech emotion recognition using generative adversarial networks. *arXiv preprint arXiv:1806.06626*.

Sahu, S., Gupta, R., Sivaraman, G., AbdAlmageed, W., and Espy-Wilson, C. (2018b). Adversarial auto-encoders for speech based emotion recognition. *arXiv preprint arXiv:1806.02146*.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432.

Salamon, J. and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283.

Šalkevicius, J., Damaševičius, R., Maskeliunas, R., and Laukienė, I. (2019). Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, 8(9):1039.

Satt, A., Rozenberg, S., Hoory, R., et al. (2017). Efficient emotion recognition from speech using deep learning on spectrograms. In *Interspeech*, pages 1089–1093.

Schlüter, J. and Grill, T. (2015). Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*, pages 121–126.

Schuller, B., Batliner, A., Steidl, S., and Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech communication*, 53(9-10):1062–1087.

Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., et al. (2013). The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism.

Schuller, B. W. (2018). Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

Shankar, R., Kenfack, A. H., Somayazulu, A., and Venkataraman, A. (2022). A comparative study of data augmentation techniques for deep learning based emotion recognition. *arXiv preprint arXiv:2211.05047*.

Shor, J., Jansen, A., Maor, R., Lang, O., Tuval, O., Quitry, F. D. C., Tagliasacchi, M., Shavitt, I., Emanuel, D., and Haviv, Y. (2020). Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Snyder, D., Chen, G., and Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Stahl, B. C., Schroeder, D., and Rodrigues, R. (2023). *Ethics of artificial intelligence: Case studies and options for addressing ethical challenges*. Springer Nature.

Stasiak, B. and Rychlicki-Kicior, K. (2012). Fundamental frequency extraction in speech emotion recognition. In *Multimedia Communications, Services and Security: 5th International Conference, MCSS 2012, Krakow, Poland, May 31–June 1, 2012. Proceedings 5*, pages 292–303. Springer.

Stuhlsatz, A., Meyer, C., Eyben, F., Zielke, T., Meier, G., and Schuller, B. (2011). Deep neural networks for acoustic emotion recognition: Raising the benchmarks. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5688–5691. IEEE.

Tao, H., Shan, S., Hu, Z., Zhu, C., and Ge, H. (2022). Strong generalized speech emotion recognition based on effective data augmentation. *Entropy*, 25(1):68.

Tao, J., Kang, Y., and Li, A. (2006). Prosody conversion from neutral speech to emotional speech. *IEEE transactions on Audio, Speech, and Language processing*, 14(4):1145–1154.

Trigeorgis, G., Ringeval, F., Brueckner, R., Marchi, E., Nicolaou, M. A., Schuller, B., and Zafeiriou, S. (2016). Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5200–5204. IEEE.

Valiyavalappil Haridas, A., Marimuthu, R., Sivakumar, V. G., and Chakraborty, B. (2022). Emotion recognition of speech signal using taylor series and deep belief network based classification. *Evolutionary Intelligence*, 15(2):1145–1158.

Ververidis, D. and Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication*, 48(9):1162–1181.

Wani, T. M., Gunawan, T. S., Qadri, S. A. A., Kartiwi, M., and Ambikairajah, E. (2021). A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814.

Williamson, J. D. (1978). Speech analyzer for analyzing pitch or frequency perturbations in individual speech pattern to determine the emotional state of the person. US Patent 4,093,821.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wu, C.-H. and Liang, W.-B. (2010). Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2(1):10–21.

Wu, S., Falk, T. H., and Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech communication*, 53(5):768–785.

Xu, M., Zhang, F., and Zhang, W. (2021). Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset. *IEEE Access*, 9:74539–74549.

Zhalehpour, S., Onder, O., Akhtar, Z., and Erdem, C. E. (2016). Baum-1: A spontaneous audio-visual face database of affective and mental states. *IEEE Transactions on Affective Computing*, 8(3):300–313.

Zhao, J., Mao, X., and Chen, L. (2019). Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323.

Zhao, W. and Yang, Z. (2023). An emotion speech synthesis method based on vits. *Applied Sciences*, 13(4):2225.

Šalkevicius, J., Damaševičius, R., Maskeliunas, R., and Laukienė, I. (2019). Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, 8(9).