National College of Ireland

PsycoPipe-Multimodal Emotion Analysis Framework Using Fine-tuned SER and ASR models with LLM For Psychiatric Conversations

MSc Research Project
MSCAI

Soumya Madhav
Student ID: X23332018

School of Computing
National College of Ireland

Supervisor: Sheresh Zahoor

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Soumya Madhav
Student ID:	X23332018
Programme:	MSCAI **Year:** 2024 - 2025
Module:	MSc Research Project
Lecturer:	Sheresh Zahoor
Submission Due Date:	01-09-2025
Project Title:	PsycoPipe-Multimodal Emotion Analysis Framework Using Fine-tuned SER and ASR models with LLM For Psychiatric Conversations
Word Count:	7915 **Page Count:** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Soumya Madhav

Date: 31-08-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

PsycoPipe-Multimodal Emotion Analysis Framework Using Fine-tuned SER and ASR models with LLM For Psychiatric Conversations.

Abstract -- Accurate emotion recognition in psychiatric conversations is crucial for developing effective clinical decision support systems. Recognising emotions from speech and text is important for creating an effective emotion analysis system . Common challenges in this domain are the lack of domain relevant psychiatric conversational data and lack of domain adaptive models. This thesis introduces PsycoPipe, a multimodal framework that integrates fine-tuned Speech Emotion Recognition (SER) and lightweight tiny Automatic Speech Recognition (ASR) with Large Language Model (LLM) analysis to address these challenges. The study applies a two-phase approach Phase 1 develops a synthetic data generation pipeline using Gemma-3 for conversation generation and IndexTTS with EARS voices for realistic emotional speech synthesis. This dataset is utilised in the fine-tuning of Wav2Vec2-base for SER and fine-tuning of Whisper-tiny for ASR applying selective layer unfreezing and hyperparameter tuning to capture psychiatric terminology and emotional expressions. Phase 2 evaluates inference on a separate synthetic test set combining ASR+LLM semantic analysis and SER acoustic features through a late fusion LLM based ensemble with confidence weighting and label mapping. Results show that fine-tuned Whisper reduces Word Error Rate (WER) by 62%, Fine-tuned Wav2Vec2 achieves 92.3% accuracy with 92.9% F1-score and the ensemble outperforms standalone models, lowering mean emotion distance to 17.38% from 32.4% for SER alone. These findings validate the use of synthetic data and domain specific fine tuning, proposing a scalable and privacy-conscious framework for clinical conversation analysis leading to real-world applications in mental health support.

Keywords : SER , ASR , Fine-tuning , Whisper , Wave2Vec2,WER,Ensemble,LLM

1. Introduction

1.1 Background and motivation

Mental health disorders affect close to a billion people worldwide, proving to be one of the major health concerns. Depression, anxiety and stress being the most widespread conditions throughout the world. Traditional approaches in psychiatric assessment or human emotion analysis rely heavily on clinical observations, which is valuable but can be subjective due to induced human bias and fatigue. Human emotions are complex as they are expressed not just through what an individual says, but also through tone of voice, speaking style, language barriers and surrounding environment. These subtle emotion cues can be easily overlooked or misinterpreted during busy clinical settings or long sessions due limited cognitive ability. This creates a need for systems that can consistently capture , analyse and predict emotions adding to human expertise in psychiatric care.

However rapid progress in artificial intelligence has created promising ways to address this need. Speech emotion recognition has advanced to deep learning models like wave2vec2 where these models automatically extract patterns from human speech than the methods which needed

manual selection of features. At the same time advancements in automatic speech recognition with models like Whisper has shown strong results across various languages and speech patterns. In addition to this large language models can now analyse texts and capture emotions. Together these advancements pave the way to build multimodal systems to analyse both text and audio data for enhanced emotion analysis. Despite the progress emotion recognition systems face major challenges such as lack of data representing real world scenarios like real multi speaker psychiatric conversation data in clinical surroundings due to privacy and ethical concerns. The second challenge seen is that these models are majorly trained on general conversations and may not perform well in analysing psychiatric conversations where emotion patterns, medical terms may be different from everyday exchanges. This research work tackles these challenges through the design of PsycoPipe a multimodal system developed for psychiatric conversation analysis using synthetic data generation, fine-tuning strategies and ensemble learning adding to clinical value by supporting mental health professionals in the decisions.

1.2 Research Questions and Objectives

- Primary Question
    - How effectively can multimodal ensemble framework combining SER, ASR and LLM analysis improve emotion recognition accuracy in psychiatric consultations?
- Secondary questions
    - Can synthetic psychiatric conversations generated provide real-world settings scenarios?
    - How does domain specific fine tuning affect the performance of models in psychiatric contexts?
    - What are the best effective ensemble fusion strategies for combining both sound emotion and semantic predictions while considering model confidence scores?
- Research Objectives
    - Generate synthetic psychiatric conversations using LLMs and TTS models.
    - Fine-tune whisper tiny and wave2vec2 models on the synthetically generated data to improve domain adaptation.
    - Build a SER pipeline using wave2vec2 model
    - Build an ASR with LLM analysis pipeline using Whisper and gemma-3
    - Develop a late fusion multimodal ensemble framework to combine SER, ASR and LLM analysis.

2.Related Work

The related work section will review advancements in ASR, focusing on the Whisper architecture and enhancement strategies, examine synthetic data generation techniques, explore Wav2Vec2 architecture with SER research, fine-tuning approaches and look into evaluation metrics with multimodal fusion methods combining textual and audio data features for enhanced emotion analysis.

2.1 Automation Speech Recognition (ASR): Whisper

(Radford et al., 2022) presents an automatic speech recognition model named whisper to address the lack of an quality pre-trained decoder a weakness seen in speech recognition models limiting their ability to new environments .This encoder - decoder transformer model is trained on 600 k hours of weak supervised audio and text transcripts .Where the audio data is resampled to 16KHz with Mel spectrogram representation ,with audio sliced into 30secs clips after filtering low quality, mismatched or duplicate transcripts. They also used an audio language detector developed by fine tuning prototype models to evaluate the data quality. The spectrograms pass through two convolutional layers with GELU activation. Sinusoidal positional embeddings are added to audio before they pass through transformer layers ,final layer normalisation standardises the output, ensuring quality data for the decoder stage. The decoder uses learned positional embeddings and shares the same token for input and output. All the tasks are combined into one process by using special tokens to start the transcript, no speech, language tag and translation for the decoder to understand. Training was done with AdamW optimisation for 3 epochs. The model performance was measured on word error rate metric obtaining good results in English ASR with error rates reduced by half compared to baseline models. Limitations include weak low resource language results and long-form hallucinations. To tackle these concerns seen in whisper small models in low resource language settings (Ferraz *et al.*, 2024) introduces DistilWhisper, to enhance the ASR performance in *whisper small models* by language specific gated modules plus applying knowledge distillation from *large whisper models*. The model was trained on the common voice 13 data with each language represented by ten thousand audio inputs and evaluated on both CV and FLEURS across eight different languages. Distil Whisper reduced the word error rate gap to the large model by 35% and 75% in data rich language settings. It consistently performed better than standard fine-tuned LoRA models specifically in low resource settings. Drawbacks seen were when the KD from teacher model was flawed the performance was hindered.

Similarly, the paper (Gockcimen, Das and Das, 2024) evaluates the performance of whisper models on Turkish language which is a challenge for ASR models due to its distinct sounds and long forming words attached to root words that add to the confusion. For the data processing the audio paths and transcripts were extracted from CV 11 dataset with punctuations dropped and metadata ignored. The tokenizer was designed according to Turkish linguistics. Whisper tiny, small, medium and large models performance was measured on metrics WER, training loss, validation loss. Results showed that medium shows improved accuracy but increased training time while whisper small model achieved balanced results with improving accuracy and lesser training time. Other models showed better accuracy as the computation cost went up.

Whisper must listen to the whole audio before transcribing the data. (Zhou *et al.*, 2025) demonstrates a method where a two pass decoding method is adapted to generate transcripts for live streaming data. This method has a CTC decoder which runs and gets instant partials

and the original attention decoder whisper is used to rescore the final output. A hybrid tokenizer is adapted where CTC is restricted to a lesser number of tokens and the attention decoder keeps all the tokens to improve generalisation. The U2 model got up to 17% on the earnings test audio functioning on CPU cores without the need for GPU s with very little loss of accuracy.(Qu et al., 2024) utilises whisper for SER in emotional support conversations. Here the whisper encoder layer remains frozen and extra transformer, projection layers added are fine-tuned on RAVDES data. The SER whisper approach achieved a promising result of 94% accuracy compared to baseline models. Implementing a two-stage weighted training method (Chou et al., 2024) develops Whisper-SER a unified Whisper Tiny-based model that combines ASR and multi-label SER. It achieves ASR error reductions 29.75%/63.58% with minimal SER performance loss, proving the first successful shared encoder–decoder approach without removing unclear emotional data.

2.2 Speech Emotion Recognition – Wave2Vec2

The study here (Baevski et al., 2020) introduces wave2Vec2 a self-supervised learning framework ideated to reduce the dependency on large amounts of transcribed audio for speech recognition in previous systems and adapts the human learning way that is from audio representations. First the raw audio is passed through a convolutional encoder converted into latent features and normalised.A small convolution layer applies relative position embeddings to avoid failures during longer sequence.The masked features are then fed to the transformer for contextual representations. GELU activation was applied on both CNN and transformer outputs .At the same time the encoder outputs are passed through a quantisation module that discretises them into speech units using product quantisation with Gumbel softmax. System is trained to identify the correct target among distractors through a contrastive loss along with diversity loss to balance the model learning. In the second stage these pretrained representations are fine-tuned using connectionist temporal classification (CTC)loss function layer on labelled data . CTC is used here to map audio and text of different lengths. The model was tested on three datasets Librispeech,Librivox and TMIT and with ten mins of labelled data the model was showed 4.8% word error rate in clean data and 8.2% WER with noisy data better than earlier models .Drawbacks seen were is that training was computationally expensive.

Due to high computational needs of SER models like Wav2Vec2 ,WavLM which are too large for low-resource environments. The research here (Kounadis-Bastian et al., 2024) aims to develop a model that maintains performance while reducing parameter count and memory. The methodology involves a knowledge distillation approach where a teacher model with ensemble of WavLM and Dawn was created to output soft labels instead of hard labels to learn subtle cues for training smaller model. Avoiding issues like human annotation inaccuracies and allow data augmentation. In the Wav2Small architecture input features are processed through a VGG7 convolutional block with a novel token vectorization method . Quadrant correction loss function was also introduced to penalise student models if the predictions were different from teachers. Results show that Wav2Small with only 72,000 parameters and 9MB RAM usage, achieves comparable concordance correlation coefficient (CCC) score to larger models on both datasets IEMOCAP,MSP Podcast that is CCC of 0.66 for a particular emotion on MSP podcast marking the efficiency of light models.

The study here (Zhao et al., 2024) looks into speech emotion recognition framework that highlights the limitations of depending on single acoustic features. Traditional methods like

using MFCC, spectrograms or Wav2Vec2 embedding on their own misses crucial cues, that is MFCC gives in details on pitch, spectrograms gives us pitch patterns and embeddings gives us the deep context of the scenario. Simple feature combination of these loses deep context information. To address this the paper here proposes a cross fusion method that enables interaction between three feature types MFCC processed with BiLSTM, spectrograms through AlexNet, and Wav2Vec2 embeddings . Each audio from IEMOCAP dataset is segmented into 3 secs MFCCs derived from librosa, spectrograms taken from STFT with Hamming windows and embeddings taken directly from the raw data. Cross fusion is done using 1D convolutions to achieve same features width continued with multi-head attention so the features complement each other .Out of all the ways to combine features MFCC + embeddings merged into spectrograms followed by a self-attention layer achieved the best results that is weighted accuracy of 72% and unweighted accuracy of 73 % higher than the baseline by 2-4 %.

Recent work showcases the usability of Wav2vec2 in mental health analysing tasks. (Huang et al., 2024) proposes a voice-based depression recognition framework using Wav2vec2.0 pre-training on achieving high accuracy 96.5% binary and 94.8% multi-class displaying its efficiency for early depression screening. In addition (Wang and Yang et al., 2025) fine-tuned Wav2vec2.0 with a Neural Controlled Differential Equations (NCDE) classifier for SER on IEMOCAP achieving weighted accuracy of 73.3% and unweighted accuracy of 74.1% with less epochs and overall stability.Together these works showcase Wav2vec2 models strong application in emotion recognition.

2.3 Synthetic Data Generation Process

There is a lack of multi-speaker conversational audio recording and labelled transcripts which are not easily accessible due to privacy concerns for ASR tasks. (Cornell et al., 2024) present a synthetic data generation pipeline to address this scarcity. Their method uses large language model Llama-3 Instruct with prompting done with few Spotify podcast transcripts to generate two speaker conversational transcripts. The text generated is then converted to audio using Parakeet TTS model with reverberations added to match Mixer6 data. The generated audio and corresponding transcripts are used to fine-tune Whisper medium with LoRA adapters. The work was tested on two datasets Fisher Corpus and Mixer 6. The study compared three audio generation processes NeMo multi-speaker generation, xTTS-v2 and Parakeet conversational TTS. Results showed Parakeet with LLM transcripts clearly outperformed other synthetic methods, achieving cpWER of 20.4% on Fisher compared to 34.3% with NeMo MSS. CpWER here refers to the WER metric adjusted to be fair towards multiple speaker data by combining all speaker data into one and then evaluating WER. Limitations seen were though the parakeet audio generation process proved to be better than other synthetic methods used in the study could not match the cpWER of in domain real data.

Tran et al. (2025)explore Domain Adaptation with Synthetic data (DAS) a framework to adapt ASR models like Whisper to specific domains using only synthetic data. The study utilises Llama3 70B to generate domain specific texts and an internal TTS system two convert texts to audio. The audio outputs generated are then used to fine-tune Whisper. Only the decoder layers of the model are fine-tuned with LoRA (Low-Rank Adapters). Separate LoRA adapters are trained for different domains like music, weather and sports with each pushing predictions. During inference all the adapters run together producing multiple word predictions along with the base Whisper output. All the outputs are compared and one with best confidence scores is selected. The study evaluates the framework for three domains music, weather, and sports using synthetic training datasets against real datasets collected with Meta Ray Ban Glasses with

minimal processing done. Results show with DAS WER reduced by 11% in music, 17.2% in weather, 10.3% in sports compared to Whisper base. More importantly out of the domain performance showed only 1% drop.

2.4 Fine- Tuning Strategies

The study here (Liu et al., 2024) explores different fine-tuning strategies on whisper ASR tasks for underrepresented language during whispers pre-training. The study mainly uses FLEURS data with different speakers for training, testing and validation while following hugging face preprocessing training set up specified for best results. The first method that is vanilla fine tuning ,where all the parameters are trained achieved best results with WER reducing from 40 to 17 % . The authors also explored freezing bottom layers that usually captures the basics sounds like all languages achieving faster training time, but the WER (word error rate) reduced by 6% from previous method. Reinitialising the encoder layers completely was also explored showing bad results. Lighter training techniques with lesser parameters and reduced training time like bottle neck adapters and LoRA were also tested achieving efficient results that is average WER of 19 and 21 % compared to pre-trained model. Full fine tuning reaches best accuracy but the training hours and hardware requirements were high. Lighter method like bottleneck adapters and LoRA accuracy drops but can be implemented for low resource settings. (Wagner et al., 2025) focus on improving ASR for dysarthric speakers that is individuals suffering with speech difficulties. Standard ASR models perform lower with this speech due to less data and high variability as it differs with each person. Hence the study here proposes a hybrid approach of fine-tuning, speaker personalisation and synthetic data augmentation to improve the ASR performance in speech disabilities. SAPS dataset with around 300 hours of speech data is utilised. Firstly, personalisation is achieved by adding x vectors (speaker specific audio representations) to decoder layers. To address issue of less data, parler-TTS model was fine tuned to generate synthetic dysarthric data and augmented with training data, transcriptions were generated using LLM. The combination of these methods were tested on FFT, LoRA and ADALoRA techniques with ADALoRA showing optimal reduction in WER by 31 % from non- personalised model.

The paper (Liao et al., 2023) proposes two methods, one model where whisper is fine-tuned with audio- text transcripts but here every transcript has prompts specifying domain. And in the second model the encoder is frozen and decoder transcripts are fine-tuned with prompts added. Trained on gigaspeech with GPT-3.5 created domain tags models reaches 33% WER reduction and second method achieves WER reduction of 29%.

(Chen & Rudnicky, 2023) explore different fine-tuning strategies of Wav2Vec2 to improve Speech Emotion Recognition addressing the challenges of small datasets and mismatch seen between pre-training data and emotional speech. They evaluate three approaches. The first is a vanilla fine-tuning method where the pre-trained model was adapted using global average pooling passed through a ReLU layer and finally a linear classifier to classify emotion with SpecAugment applied during training to improve generalisation. The second method, Task Adaptive Pre training (TAPT) is done by training the model again on emotion datasets IEMOCAP -7 hours, four emotions and SAVEE -30 minutes, seven emotions to reduce the domain gap. Finally, the authors propose Pseudo-Label TAPT which uses k-means clustering on frame-level features derived from audio to generate pseudo-label that is emotion clusters to improve emotion recognition. Results showed that V-FT method outperforms previous models. TAPT boosts accuracy and P-TAPT achieve the best results, that is a 7.4% improvement over the previous work done. The models performed even better than human annotators for SAVEE

data. Limitations seen is that performance was unstable as few runs gave bad results, also the training was limited to two datasets. This study (Sampath et al., 2025) examines effective methods for fine-tuning Wav2Vec2 on the MSP-Podcast data for dimensional speech emotion recognition focusing on predicting continuous scale that is activation and valence than predicting fixed categories. Labels were scaled and audio clips were padded per batch or at caching. Several approaches like full fine-tuning, partial fine-tuning of 3 layers, LoRA and caching strategies were studied. Results show partial fine-tuning of the final three layers with mixed precision matched full fine-tuning with achieving better training time that is 67% faster, while caching further improved efficiency with tiny decrease accuracy. Limitations seen were the paddings added were confusing the model and causing performance drops and explored only one model.

2.5 Multimodal Analysis (Text - Audio)

(Zhao, Wang and Wang, 2022) put forward a multimodal emotion recognition approach that uses transfer learning with wave2vec2 for speech and BERT for text. To address issue of lack of labelled data they recommend a multi granularity framework that not only uses frame-level speech embeddings but also looks into phones, syllables and whole words to understand sensitive cues like tone and rhythm. Two fusion methods were examined . One is late fusion in which speech and text are processed separately and then combined for final results. Second is co-attention based early fusion where both features use attention to highlight the important parts. Last one is hybrid approach where both methods were utilised for learning emotions on IEMOCAP data. Hybrid method including both fusion techniques achieved the best accuracy of 76%. Limitations seen were signs overfitting. Traditional multimodal SER methods often suffer from performance issues due to incorrect ASR transcripts .To address this concern (He et al., 2024) explore a multimodal module to learn modality specific and variant representations reducing the gap between acoustic features and text features. This module is supported by two tasks that are AER ASR error detection to locate incorrect words and AEC that is ASR error correction to improve the quality of transcripts. The model showed better results than earlier methods that is 79% UAR accuracy . But when the authors tried omitting one of the methods the performance dropped hence proving the importance of all the methods.

Mental-Perceiver the work here (Qin et al., 2024) specifically looks into depression analysis. Most of the earlier work is on small data so the authors create MMPsy dataset containing 4266 depression interviews in Mandarin language. Thorough preprocessing steps like filtering unreliable responses, removing silence , denoising with Spleeter, generating transcripts with Paraformer , manual corrections were followed. The model combines both audio and text with experiments proving multimodal input outperforms single modality. The unique design architecture using attention throughout with semantic priors and a dual-loss strategy implemented makes it more efficient than baselines like PerceiverIO and ConvLSTM achieving 0.79 UAR and 85% accuracy. Downsides observed were the research was narrow due to language specific data.

Most of the prior work explored standard human done transcripts but work here (Tran et al. (2025) benchmarks SER using outputs from 11 different ASR models with different WER levels on three different data sets. Preprocessing steps included lowercasing all transcripts, removing punctuation from Whisper outputs and excluding utterances with missing text. SER modelling utilised RoBERTa base text encoders with audio features added for bimodal analysis. Six fusion methods that are early, late, cross-attention, tensor fusion, NL-gate, and

MISA were measured and compared. Results show that SER works well when the WER is around 12%.When combined text and audio analysis the accuracy loss drops from 10 to 6%. The study further proposes a new framework for correcting errors and then combining text and audio for analysis which further improves accuracy to 76% compared to 74% on gold transcripts that are manually created transcripts.

3. Research Methodology

This section explains step by step methodology utilized in this thesis work. The research work deploys two-phase methodology to develop a multimodal emotion recognition framework addressing the challenges in real-time analysis of psychiatric conversations. The methodology includes synthetic data generation and preprocessing, fine-tuning models for domain data, ensemble fusion and evaluation to deliver a robust system for psychiatric conversation analysis.

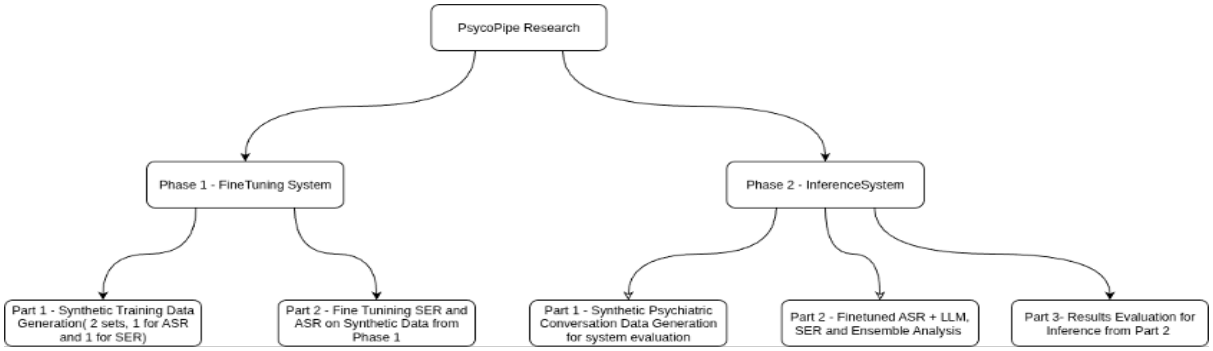


Figure1- Research Methodology overview

3.1 Two-Phase Research Framework

Phase 1: Fine-Tuning System

Part 1: Synthetic Training Data Generation - Builds two separate training datasets one streamlined for ASR model fine-tuning and another for SER model fine-tuning aligning to their tasks. Generated text data is converted to audio samples using a TTS model for generating realistic emotional conversations. Ensuring a balanced representation for diverse emotions in psychiatric conversations.

Part 2: Fine-tuning SER and ASR on synthetic data from part 1 - Wav2Vec2-base model for speech emotion recognition and Whisper-tiny model for automatic speech recognition is fine-tuned using synthetic training data generated in Part 1.The models are gradually trained such its adapts to domain data and validated with performance tests.

Phase 2: Inference System

Part 1: Synthetic Data Generation For System Evaluation - Builds a new set of psychiatric consultation texts with varying conversation lengths and characteristics to measure performance on unseen data and avoid data leakage.

Part 2: Fine-tuned ASR + LLM, SER and Ensemble Analysis - The fine-tuned ASR model transcribes the testing data audio samples. The transcribed text is then analysed using Gemma3 LLM for semantic emotion analysis. At the same time, the SER model is used to detect emotions from acoustic features. Finally, the ensemble fusion method is implemented to combine both modality predictions for enhanced analysis.

Part 3: Results Evaluation - The performance of finetuned models is measured against baseline models using metric WER for ASR tasks and metric accuracy , F1 score , precision for SER tasks. During final inference both the outputs evaluated on test data are combined using weighted algorithms and confidence scores to predict emotions.

3.2 Data Generation and Processing

The data generation module was designed to create multi-speaker domain-relevant varied-emotion psychiatric consultation conversations that serve as the major resource in multimodal emotion training and recognition. As access to real-time psychiatric datasets is limited due to scarcity and ethical concerns. A structured pipeline was implemented such that realistic data was generated consistently.

Training Data Generation - This phase generates two datasets for SER and ASR tasks. For Whisper-tiny finetuning ,Gemma3 LLM is utilised to generate psychiatric , medical terminology specific text samples for different subcategories like patient symptoms, medical-assessments, treatment discussions. There are five main categories with subcategories 150 samples are generated for each subcategory till it reaches 2500 samples to varied conversations. Diverse prompts were used to produce natural, domain relevant concise data for easy audio conversion using TTS models. Length filtering and white space removal was done to achieve quality in data. The data generated is then saved in a structured Json file with metadata including model,timestamp and categories to achieve reproducibility.
For Wave2Vec2 finetuning, LLM generates emotion specific texts mapping to 17 emotions. For each emotion 150 samples are created generating a total of 2500 short emotion rich sentences with balance achieved across each emotion category.These texts are also filtered for quality and saved in Json file with metadata. The generated text dataset for both cases are then converted into the speech as detailed below.

Evaluation Data Generation - For inference fresh dataset is generated. The evaluation data creation uses the same methodology as in Phase 1 but with different conversation lengths, and emotional distributions to avoid overlapping with training data. LLM was prompted to create multi speaker conversation with key variables like gender and number of exchanges specified. All the data produced was stored in two formats .txt for debugging with raw outputs and a structured CSV file with emotion, gender, exact spoken sentence and role mapped. The same CSV was used to generate audio as detailed below.

Text-to-Speech Synthesis - A key step in this research was successful integration of IndexTTS - a GPT-style text-to-speech (TTS) model mainly based on XTTS and Tortoise to create natural sounding emotional speech. After testing multiple TTS systems like Kyutai, Coqui.ai TTS, F5-TTS , IndexTTSv1.5 turned out to be the optimal solution due to its refined and advanced emotional expression features. Index-TTS model is single /multi-speaker model pretrained to produce natural, context-aware and emotionally expressive speech enabling the generation of realistic clinical conversation audio.Further, voice-emotion mapping strategy is done using EARS(Emotional Audio Recognition System) dataset ensuring that synthetic samples match

natural voices consistently.This proved essential for creating training and evaluation datasets that accurately reflect the emotional complexity of real psychiatric consultations. Additionally IndexTTSs unreleased version in their pipeline (IndexTTSv2) will allow for emotion control through text prompts which take voice reference audio for speaker variation and can be used with minimal changes.

3.3 Model Architecture and Fine-tuning Strategies

Speech Emotion Recognition Component

Wave2vec2 base model by Meta was chosen for speech emotion recognition task due to its ability to learn acoustic features from raw audio. The model is pre-trained on 960 hours of unlabelled Librispeech dataset which enables it to learn tone, pitch , and pronunciation styles which are critical in SER. Also the model is better at capturing smallest details like stress on words and overall pitch pattern without the need for manually designed features, making it the right fit for SER tasks.

For fine-tuning Wace2Vec2 a partial tuning strategy was utilised where the classifier linear layer was updated with 22 emotions, the upper layers of encoder were updated with small learning rate and limited epochs while the lower layers were untouched so that gradual and stable learning was achieved. These adjustments were made according to learnings from the experimentations done. After each epoch, results were tracked with accuracy, precision and F1-score metrics. The Fine-tuned model performance is compared to the baseline model.

Automatic Speech Recognition Component

For automatic speech emotion recognition task Whisper tiny model by OpenAI was utilised. Whisper is trained on 680K hours of multilingual audio and text transcripts covering varied accents, noise conditions and languages. It also handles multiple tasks like language identification and speech translation with proven low error rates , which makes it more adaptable to finetune on domain specific data.

Fine-tuning of Whisper was done by implementing a training strategy where only the last few decoder and upper encoder layers were updated, while lower encoder layers and remaining decoder layers were frozen to preserve general features. Hyper parameters were adjusted according to research such that the model learns gradually and does not plateau quickly. To avoid the risk of overfitting, regularisation, label smoothing and dropout was incorporated along with performance monitoring mechanisms. The fine-tuned Whisper model was evaluated against the base Whisper model using metric word error rate (WER). This comparison showed the fine-tuned model performed better in recognising psychiatric terms and dealing with varied audio inputs than the base model. Both the trained models will be further utilised in combined emotion analysis.

Large Language Model Integration

Gemma-3 is a LLM designed by Google DeepMind , trained on large text data across various domains which gives it the ability to understand semantics, detect emotions and recognise

domain specific information. Though the model is smaller in size ,compared to GPT-4 models it displays robust performance in classification, summarisation and reasoning tasks. In the integration process the transcripts generated are passed through gemma-3 with prompts requesting emotion analysis.Gemma3 model interprets the semantic representations of the transcripts generated for evaluation data. The prediction outputs with labelled emotion and psychiatric cues are stored in CSV file for further inference.

3.4 Inference Pipeline

The inference pipeline combines SER and ASR +LLM components utilising the ensemble late fusion method to extract multimodal insights from psychiatric consultations in real time. The process starts with a new set of data generated for evaluation utilising Gemma3 stored in both CSV and raw text formats. These text samples are further converted into audio .wav files using the IndexTTS system mapping each spoken line with a .wav audio path in a CSV file. The fine-tuned Whisper model is used to transcribe these audio files. These transcripts are passed through Gemma3 for emotion analysis. Emotions and cues extracted are stored in the same CSV output file with labels and confidence scores for this process. At the same time fine-tuned Wave2Vec2 analyses the audio samples created and emotions classified are stored in the same CSV output file with confidence scores labelled as SER process. For ensemble fusion , both SER predictions and ASR with LLM predictions are combined. Since the both emotions generated are from different modalities the emotion labels may differ accordingly i.e , joy for happiness .To handle this varied emotions are mapped to common standard labels to maintain consistency. For combining predictions the ensemble uses the outputs from the SER and ASR+LLM in an LLM prompt to detect the final emotion based on the emotion + description and top 3 emotion prediction by SER and their confidence scores. This provides the model more context of what was said and what probable tone it was said in. Instead of hardcoding model file paths and other important settings in each script, a configuration management script is utilised to maintain changes.

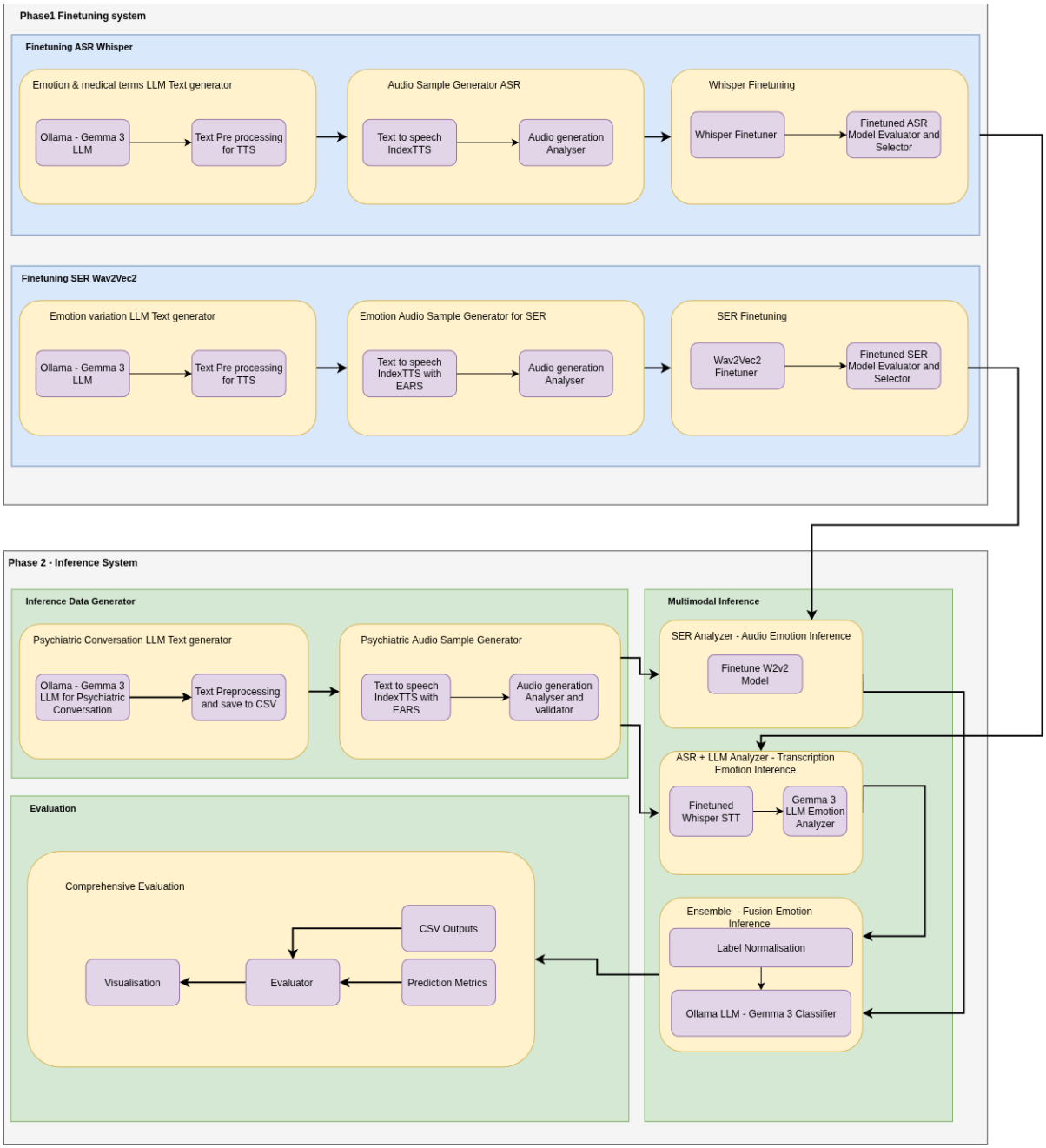
3.5 Evaluation Metrics

In this study multiple evaluation metrics were used to measure the performance of different components of the multimodal emotion recognition framework. For Whisper ASR finetuning evaluation the main metric used is Word Error Rate(WER) which measures the number of incorrect words compared to the real sentence. A lower WER indicates better accuracy. Along with this training loss and validation accuracy were calculated to monitor models learning progress.For Wave2Vec2 SER finetuning evaluation metric used is accuracy , precision and F1score.These metrics were compared against the baseline model performance to measure the improvements of fine-tuning. Finally in the ensemble pipeline both the component outputs were combined and measured with model specific confidence scores for final prediction and evaluated against the actual emotion, the emotion category and similarity was used to compute the emotion distance and each model was scored based on the emotion distance.

4.Design and Implementation

This section provides a detailed explanation of the system architecture and the details of how the 2 phase research methodology was implemented. It describes the step by step processes, tools, and methods used to implement the methodology.

4.1 System Architecture



4.1 Development Environment and Technology Stack

Development was done primarily on the Linux desktop with the below specifications to allow for faster model training and inference workloads. Some of the high level evaluation was also done on a MacBook pro machine for portability.

Component	Specification
Operating System	Linux 6.8.0-78-generic (Ubuntu-based)
Python Version	3.12.11
Virtual Environment	psyco
GPU	NVIDIA GeForce RTX 4090 (24GB VRAM)
CUDA Version	12.7 (Driver: 565.57.01)
CPU	AMD Ryzen 9 7900X 12-Core Processor
CPU Cores	12 physical cores, 24 logical threads
System Memory (RAM)	61GB DDR5 (51GB available)
Storage	180GB NVMe SSD (164GB used, 6.5GB available)

Python was the choice of the development language and the following libraries were mainly used, along with the tools mentioned below.

Component	Version
Python	3.12.11
Virtual Environment	psyco
Ollama	0.11.4

Tool	Version
VS Code	Latest
Git	2.34.1

Major Python Libraries

Library	Version	Purpose
PyTorch	2.8.0+cu128	Deep learning framework
Transformers	4.55.4	Hugging Face model library

OpenAI Whisper	20250625	Speech recognition
NumPy	2.2.6	Numerical computing
Pandas	2.3.1	Data manipulation
Matplotlib	3.10.5	Visualization
Librosa	0.11.0	Audio analysis
JIWER	4.0.0	Word Error Rate
Ollama	0.5.3	Python client for Ollama

4.2 Phase 1 - Implementation

Training Data Generation

The text generator generate_psychiatric_text_samples.py uses Ollama's local LLM interface ollama with the latest Gemma-3 model to create short, conversational synthetic psychiatric conversation data for Whisper fine tuning. The script defines five psychiatric content categories such as patient symptoms, Medical assessment, treatments, Clinical terminology to achieve diversity in data for domain vocabulary.Gemma3 is guided with structured prompts to generate short sentences with less than 20 words on subcategories. Several processing steps are applied to remove generation artifacts and achieve clean and structured data for audio conversion. Length filtering of sentences is done to limit the sentence length to< 20 words as it's easier for parsing and synthesising. Duplicate sentences are removed and only the unique sentences are used. To avoid class imbalance in data , 150 samples are generated for each subcategory. The data generated is stored in psychiatric_text_samples.json with each sentence mapped to meta data category, timestamp and count.

The data generated is stored in psychiatric_text_samples.json with each sentence mapped to metadata including category, timestamp and count. These texts are converted to high-quality audio files using IndexTTS v1.5 in generate_psychiatric_audio_indextts15.py.Audio files are generated at 16kHz sampling rate, with voice variations applied to achieve more realistic data resemblance. The .wav files are stored in the directory psychiatric_audio_outputs_indextts15 with unique names and category mapped. Final output with audio file paths for each sentence is stored in psychiatric_text_samples.json. The script processes 2,893 psychiatric text samples in batches of 10, generating high-quality audio files of 4.57 hours that is suitable for ASR fine-tuning with diverse speaker characteristics and voice variations.

Similarly generate_emotion_text_samples.py uses ollama with Gemma 3 to produce short , natural emotion specific sentences. 22 emotion class labels are defined ,150 samples each emotion is created by dynamically prompting LLM for varied emotion rich sentences with a target of 2550 sentences. The results are organised in emotion_text_samples.json. For audio synthesis generate_emotion_audio_for_SER_tuning.py uses the same TTS as in the previous step to generate 2.14 hours of audio. Emotion to voice mapping is done using voices(both male and female) from EARS(Emotionally Annotated Recordings Of Speech) data set to condition the audio generated. The 16kHz .wav audio files generated are stored in

emotion_audio_outputs for every emotions .The emotion_text_samples.json file is updated with audio files path for the texts.

Model Fine-tuning Process

Whisper fine-tuning for ASR

The whisper_finetune_Fast.py script fine tunes the Whisper tiny model for psychiatric ASR using a selective layer unfreezing. The implementation utilises Pytorch for model training,Hugging Face Transformers for Whisper integration, soundfile to load and normalise the audio and scikit-learn library to achieve train-test split. Pre-trained Whisper tiny model by OpenAI is loaded and initialised for training. Audio with transcripts from audio_generation_metadata.json in psychiatric_audio_outputs_indextts15 is processed where audio is resampled to 16kHz and normalised , texts cleaned to maintain clinical semantics. For training whisper tiny selective layer unfreezing strategy is applied as first 3 encoder layers are frozen to retain the knowledge of general speech acoustics with top encoder layers being trainable to adapt to psychiatric conversations and in decoder higher 2-3 decoder layers are unfreezed to learn psychiatric specific vocabulary and other layers remain frozen. Evaluation was performed every 25 steps with early stopping, and checkpoints were applied to retain the best training parameters. The best training configuration used an effective batch size of 64 , with learning rate at 3e-5 ,trained for 2 epochs. Evaluation loss is saved every 10 steps and at the end of training hugging face Trainer saves the best performing model. All the outputs are saved in whisper-ft-tiny directory. The script then loads both finetuned Whisper tiny and Whisper tiny original models , transcribes the psychiatric audio conversations ,computes Word Error Rate (WER) and compares both model performance.

Wave2Vec2 fine-tuning for SER

The script fast_ser_train.py implements fine-tuning Wave2vec2 for speech emotion recognition. The implementation utilises PyTorch for model training,Hugging Face Transformers for Wav2Vec2 integration, Torchaudio for loading audio files and processing and scikit-learn for train-test splits and measuring evaluation metrics. The script dynamically discovers 17 emotion labels from subfolder names under emotion_audio_outputs. The audio samples are resampled at 16kHz,normalised and padded or trimmed to fixed length of 15 seconds to maintain standard samples.

For training the SERFineTuner class uses transformers library to initialise a Wav2Vec2ForSequenceClassification model with a classification head sized to match the number of emotion labels. A full fine-tuning strategy is applied where the entire Wav2Vec2 encoder layer is updated with the classification head to adapt the model for emotional speech. The Hugging Face Trainer handles the model training with an effective batch size of 16, learning rate of 2e-5 and 5 epochs. Performance is measured at the end of each epoch using accuracy, precision and weighted F1 score as metrics. The best model selection employs a sophisticated top-3 selection algorithm based on dual criteria: highest validation accuracy $\geq$85% and lowest loss difference $\leq$0.3 to prevent overfitting. While training metrics including training loss, validation loss, accuracy and F1 measures are stored in training_metrics.json. Visualisations of loss curves, accuracy patterns, and epoch timings are saved in the training_plots in directory. At the end of training, the fine-tuned model along with emotion label mappings are stored in ser_w2v2_fast .

5.2 Phase 2 - Implementation

Evaluation Data Generation

For evaluation texts, the generator script (generate_psy_text_csv.py) uses the same strategy as in training data generation to generate multi-speaker psychiatric conversations with explicitly defining roles (doctor and patient) and number of exchanges through structured prompts. The conversations generated are stored in .txt and .csv files. The CSV file will have data marked with labels for role , emotions , conversations and gender for each conversation with a filename mapped to that particular conversation. For audio synthesis generate_inference_audio_indexTTS.py uses the similar process as in training data generation. The CSV file is updated with the respective audio paths mapped to the exact line according to the naming.

ASR with LLM Analysis

whisper_asr_analysis_with_LLM.py integrates Whisper for ASR and LLM Gemma3 for transcript interpretation with emotion detection. WhisperASRAnalyzer class utilises a hugging face library to load Whisper tiny original model and fine-tuned Whisper tiny model. Audio files from audio outputs are loaded with librosa and processed. Both the models utilise Transformers package to run separately to tokenise the audio files and generate transcripts. The CSV file is now updated with Base_ASR_transcription and ASR_transcription(for finetuned model).WER is measured for both transcripts to track performance. For semantic analysis analyze_emotion_and_psychiatric builds a structured prompt including role and transcripts from CSV to Gemma3 to perform analysis and output two lines EMOTION: and ANALYSIS: for given role (Doctor/Patient), original emotion label and transcript. If Ollama is not available fallback selects an emotion and gives analysis as unavailable.

SER Analysis

speech_emotion_recognition.py performs Speech Emotion Recognition (SER) using a fine-tuned Wav2Vec2 model. SpeechEmotionRecognizer class is the core component which loads the finetuned Wave2vec2 model checkpoint and collects the emotion label mappings from emotion_mappings.json. Audio samples are preprocessed , resampled in preprocess_audio using librosa. predict_emotion derives acoustic features and evaluates class probabilities through softmax function, emotion mapping is applied to make sure the raw outputs are interpreted into the 22 distinct emotion classes. The CSV file is updated with SER results adding two new columns SER_predicted_emotion and SER_prediction_confidence .

Ensemble fusion process

The script advance_ensemble_emotion_detection.py implements an LLM-based ensemble prediction system combining both Whisper with LLM predictions and SER predictions using Gemma3 for intelligent decision making. LLMEnsemble class receives both the predictions and normalises the emotion labels by mapping diverse emotions from both models into standard sets. employs a system/user prompt architecture where the LLM analyzes multiple

emotion predictions and determines the most accurate final emotion.The system extracts top 3 SER predictions (SER1, SERconfidence1, etc.) with confidence filtering (ignores predictions < 0.2), ASR+LLM predictions from ASR_detected_emotion, and integrates context including ASR_transcription, speaker role and llm_psychiatric_analysis.The script adds five new columns to CSV files namely ensemble_predicted_emotion, ensemble_confidence, ensemble_quality, ensemble_method and ensemble_reasoning.

4.3 Configuration Management

The config_loader.py file is a configuration management tool for the PsycoPipe project. All the runtime parameters are handled from a single config.json file to ensure consistency across the project. Following a standardized configuration access process simplifies managing model paths, evaluation and other parameters, promising seamless integration across training, inference and evaluation stages within PsycoPipe.

6. Evaluation

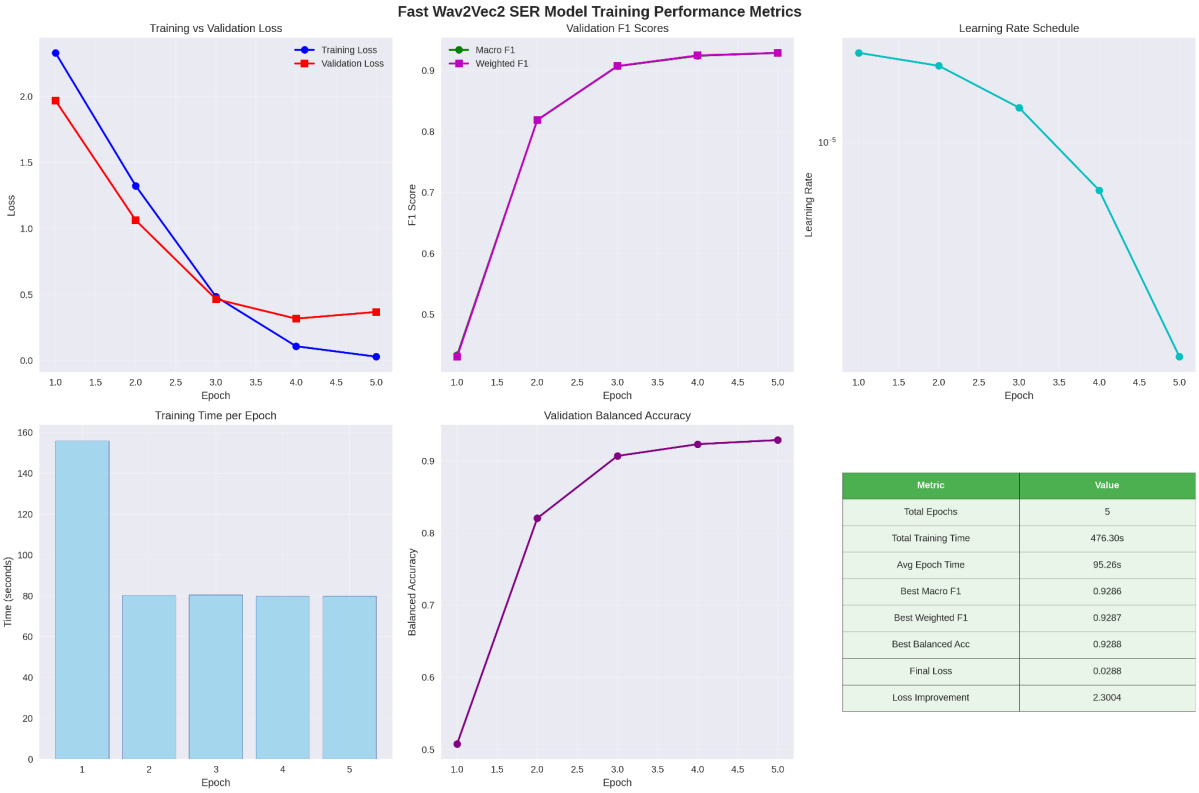
The evaluation spans across the 2 phases. In phase 1, we evaluate the process of fine-tuning and the performance of both the fine-tuned models. In phase 2, we continue to do an overall analysis of the individual model outputs and the detailed analysis of the ensemble performance.

6.1 Fine-tuned SER Performance

The fine-tuned Wav2Vec2-base model for Speech Emotion Recognition (SER) achieved a significant performance with measured results saved in ser_w2v2_fast/training_metrics.json. Effective learning seen as the training loss dropped clearly from 2.33 to 0.029 . The curve shows quick learning in early stages and then gradually improves by fine-tuning in later epochs and Validation loss decreases over epochs (1.97 → 1.06 → 0.46 → 0.32 → 0.37), reaching its lowest at epoch 4 with minor overfitting visible in epoch 5.

The model attains the best validation accuracy of 92.3% and F1 score of 92.9% showing balanced classification throughout all emotions. Most learning occurred within the first 3–4 epochs with decreasing returns in later epochs. The final epoch added little improvement, signalling model convergence.

Total training time was reduced to 8 minutes which approximates to under 2 minutes per epoch. Finetuning results for SER show that fine-tuning was successful with finetuning results obtaining high accuracy and F1 scores while effectively adapting to the SER task.

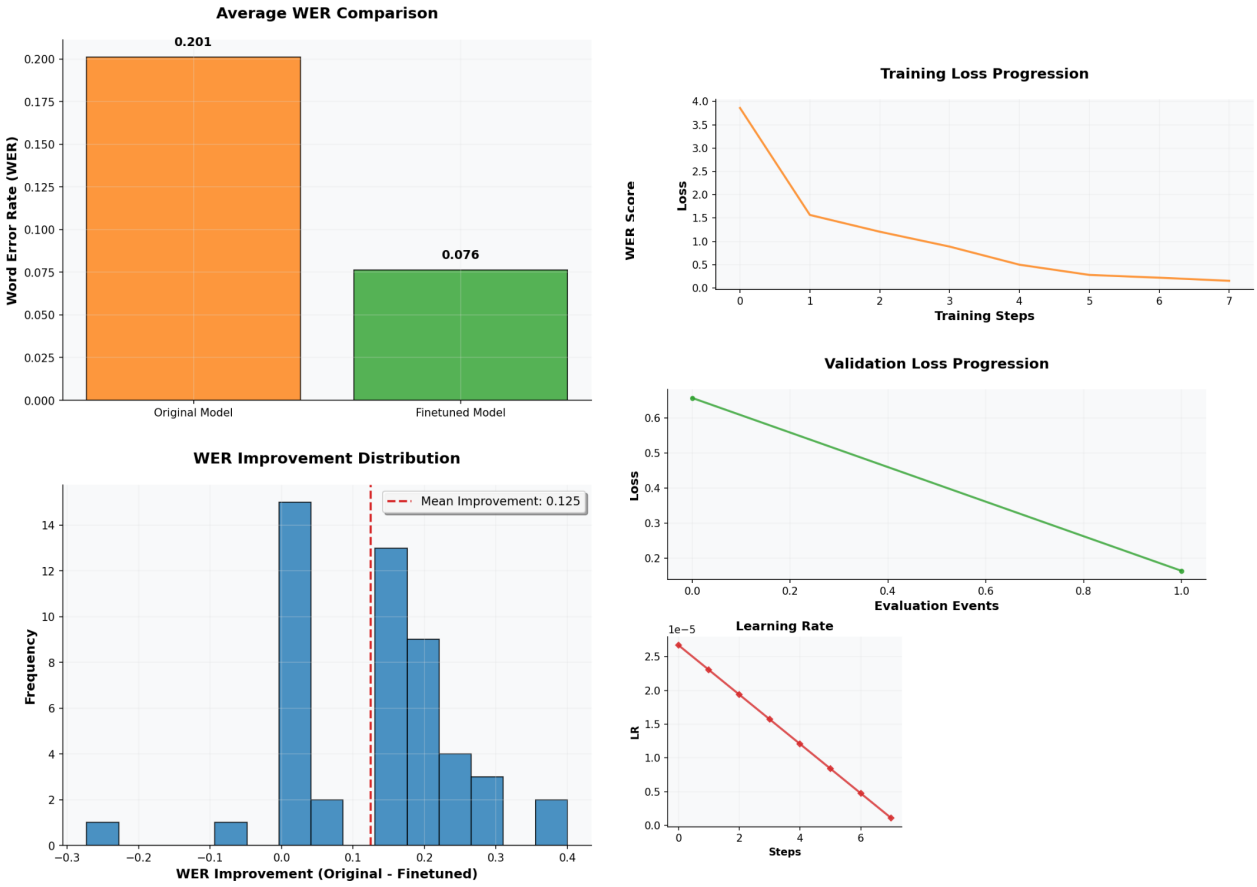


Metric	Value
Total Epochs	5
Total Training Time	476.30s
Avg Epoch Time	95.26s
Best Macro F1	0.9286
Best Weighted F1	0.9287
Best Balanced Acc	0.9288
Final Loss	0.0288
Loss Improvement	2.3004

6.2 Fine-tuned ASR Performance

The fine-tuned Whisper Tiny model for Automatic Speech Recognition (ASR) achieved good performance with evaluation results saved in whisper-ft-tiny/run_summary.json.
Effective learning was seen throughout as training loss decreased from starting values to 0.1529. The loss curve indicates the training was stable and consistent, not converging quickly.The model achieved Final Validation Loss of 0.164 and Final Validation Accuracy 97.7% . Final Training Loss was 0.153 over 2 epochs.The model showed major improvement in WER with original model WER at 0.201 and Fine-tuned model WER at 0.076 with WER improvement of 0.125 or 62.0% .

The training utilized 2,893 audio files totalling 4 hours 34 minutes of psychiatric conversation audio, with an average duration of 5.7 seconds per file and maximum duration of 25.6 seconds.Total training time was 46.24 seconds (less than 1 minute), averaging 23 seconds per epoch over 2 epochs. The training time was cut down by using a batch size of 64 on RTX4090.



6.3 Multimodal Framework Evaluation Summary

The complete analysis evaluates ASR and emotion detection performance across multiple modalities. It processes conversation CSV files containing original text, ASR transcriptions, and emotion predictions. For measuring Whisper tiny Word Error Rate (WER) is calculated using jiwer library to compare base and fine-tuned models with both standard and filtered metrics. Emotion analysis utilises semantic similarity matrices and category-based classification to evaluate SER, ASR with LLM and ensemble predictions. The system generates visualizations for WER distribution plots and overall performance dashboards while creating detailed CSV reports with similarity scores, confidence levels and category matches.

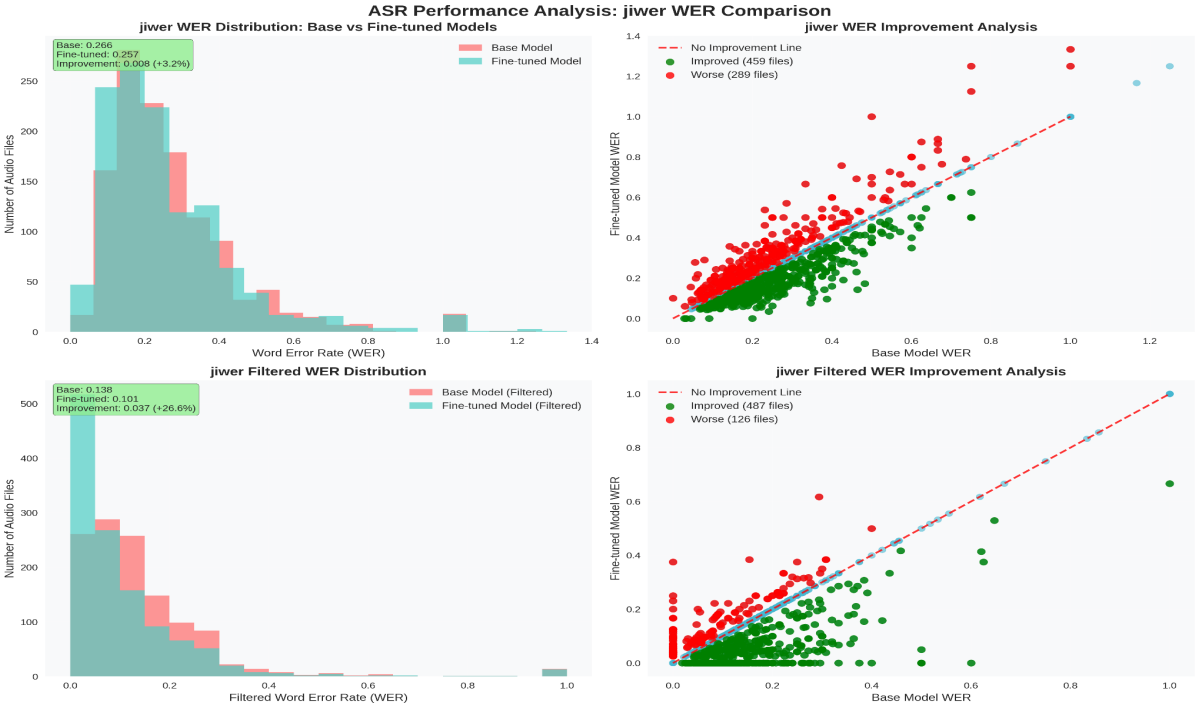
WER Analysis with Advanced Filtering

Standard WER– Uses jiwer library for baseline Word Error Rate calculation(includes fillers)

Filtered WER– Performs clinical speech normalization by removing fillers (um, uh, huh), changing short forms into full words (it's -> it is) and fixing repetitions (it's... it's –> it is it is)

WER Performance

Metric	Base Model	Fine-tuned Model	Improvement
Standard WER	26.58% ± 17.08%	25.74% ± 18.59%	+0.85%
Filtered WER	13.76% ± 14.25%	10.10% ± 14.18%	+3.66% (26.6%)



Emotion Distance Analysis

***Semantic Similarity Matrix*:** A 17×17 similarity matrix showing how close emotions are that are 0 = same and 1 = very different .

***Bidirectional Mapping*:** Matches psychiatric emotions (concerned, anxious) with standard SER emotions (desire, fear) using forward and reverse mapping.

Category-Based Classification: Groups emotions into eight categories (high_positive, low_positive, high_negative, low_negative, cognitive, neutral, therapeutic, patient_distress)

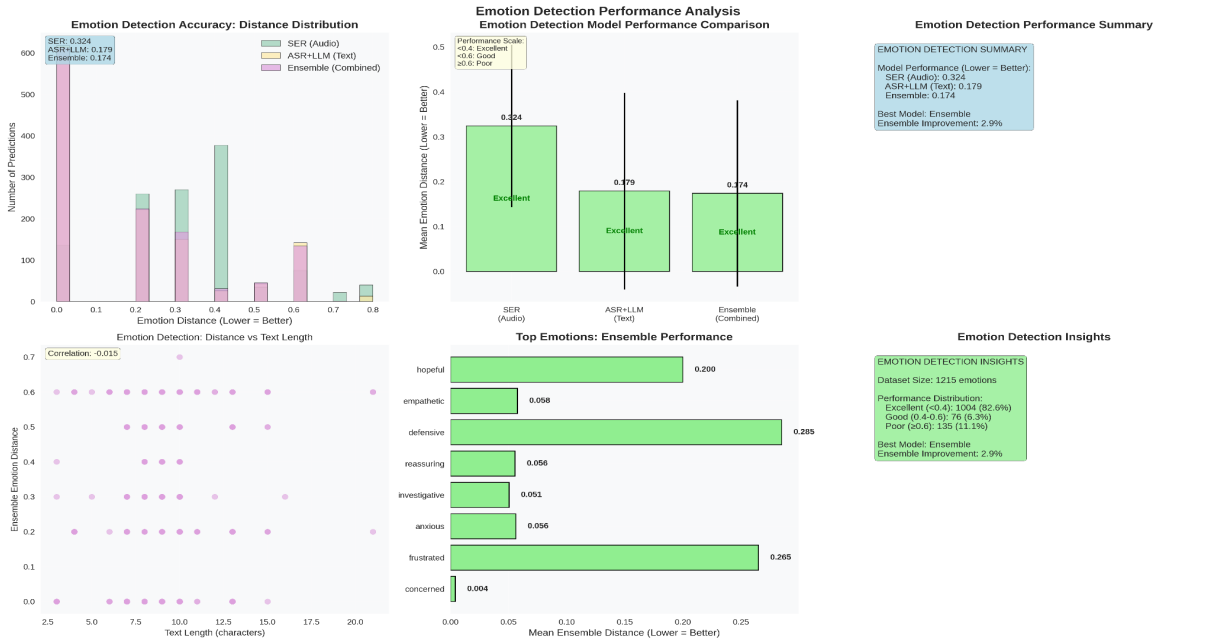
Distance Calculation: For distance calculation it integrates direct match, reverse match and category based fallback with normalized emotion labels.

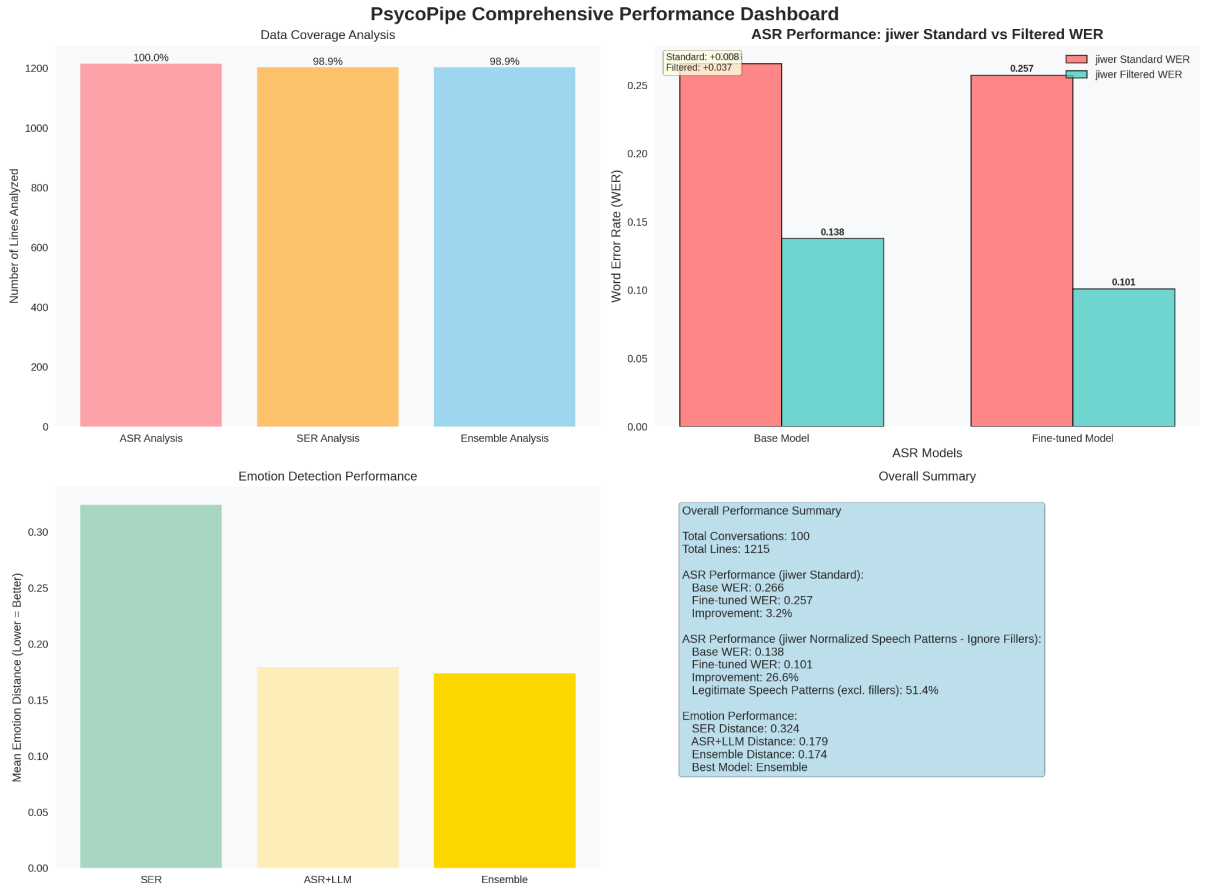
Emotion Detection Evaluation

Model	Mean Distance	Std Deviation	Performance
SER	**0.3240**	0.1806	Average
ASR+LLM	**0.1791**	0.2190	Good
Ensemble	**0.1738**	0.2076	Best

Mean Distance - Lower is better

The evaluation shows that the ensemble model outperforms other individual approaches in emotion detection. The standalone SER model had a high mean distance of 32.40% while integrating ASR with an LLM reduced this to 17.91% .The ensemble proved to be the best by further improving the performance with a mean distance of 17.38% .





Experiments

Experiment 1 : Finding the right Text to speech(TTS) Model

Finding the right TTS model required the following considerations, realistic speech synthesis, emotion control, voice variations and fast performance. The research started with exploring TTS models like tortoiseTTS, H5TTS, coqui among others. After trying many TTS models we first settled on Kyutai, this helped us progress with our project by allowing high quality audio synthesis with emotion variations. However we hit the bottle neck when we found that the Kyutai team did not opensource/ provide binaries to create voice embeddings for custom datasets. This came in as a huge limitation as the entire EARS dataset could not be used and only the ones Kyutai provided embeddings worked well.

To overcome this more research was done to settle on IndexTTS by Bilibili. This is a very advanced TTS and the v2 version of the same is set to release soon with features that allow text description based emotion control. Since only v1.5 was available, it was used . The v1.5 also provides excellent emotional audio synthesis using the EARS dataset by generating voice and emotion style transfer. It also uses the cuda GPU to the fullest to generate fast audio.

Experiment 2 : Fine-Tuning ASR and evaluating WER

This work required the verification of the effectiveness of the fine-tuning of the ASR models for psychiatric audio data. Our initial tests during the finetuning, including the WER on eval dataset showed good results and upon running the same model on the inference dataset the results changed drastically. This led to in-depth analysis of the results, evaluated a large sample of the data where the transcription varied between finetuned and Original model results, this revealed the wrong penalties set on the finetuned model which included transcription of fillers, words like it’s vs its and repetitions. These are potentially psychiatric cues that might help the analysis better. So we needed to create a normalisation for evaluating both the model performances. Hence a separate filtering process was included to filter these speech nuances. Filtering helped set a fair ground where such special cases were not rewarded or penalised. This became part of the final evaluation framework.

Experiment 3: Evaluating the Emotion Model Performance

Once the final set of inferences were generated with SER, WER+LLM and Ensemble, a robust evaluation framework was needed to ensure reliable performance assessment. Due to the fact that the system had a large number of varied emotion variations in the inference dataset to simulate real world scenarios and the EARS emotions was categorised into 17 emotions. Also the LLM predictions from the transcribed data were not within the 17 emotions as well. Hence went on to create a semantic similarity based model to evaluate the models. Different categories of emotions were defined and also created an emotion distance matrix to compute how correct or wrong each method was in predicting the emotion. Like Joy and happiness are the same, crying and pain are similar, happy and disgusted are far. This helped us effectively analyse the emotion outputs and also forms a part of the current evaluation framework.

7.Conclusion and Future Work

This research successfully illustrates the usefulness of multi-modal late-stage ensemble fusion techniques for psychiatric conversation analysis by achieving significant improvements in emotion detection accuracy and clinical relevance. The research also works on finding techniques to generate realistics synthetic datasets as privacy-sensitive data would be hard to gather in the real-world. The work also demonstrates the effectiveness of fine-tuned tiny STT models in ASR/SER for clinical domain conversations. Evaluation of the performance of the models also required detailed analysis on clinical relevance and outputs. This led to adapting of filtered WER and emotion distance calculation.

The study utilises synthetic data generated for complete analysis though the data generated closely matches natural realistic psychiatric conversations for future work we need to have a balanced approach by including both synthetic data and real world clinical data with clinicians and psychiatric experts evaluations . As this study explores Whisper tiny and Wave2vec2 a wider range of models needs to be explored for better performance and adaptation. Additionally including FER (Facial Emotion Recognition) along with industry standard practices followed would help build a more robust and clinically reliable system.

References

[1] Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I., 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint* arXiv:2212.04356. Available at: https://doi.org/10.48550/arXiv.2212.04356

[2] Ferraz, T.P., Zanon Boito, M., Brun, C. and Nikoulina, V., 2024. Multilingual Distilwhisper: Efficient distillation of multi-task speech models via language-specific experts. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of,pp.10716–10720.Available at: https://doi.org/10.1109/ICASSP48485.2024.10447520

[3] Gokcimen, T., Das, B. and Das, R., 2024. Evaluating the performance of Turkish automatic speech recognition using the generative AI-based Whisper model. *2024 9th International Conference on Computer Science and Engineering (UBMK)*, Antalya, Turkiye, pp.121–125. Available at: https://doi.org/10.1109/UBMK63289.2024.10773523

[4] Zhou, H., Song, X., Fahy, B., Song, Q., Zhang, B., Peng, Z., Wadhawan, A., Jiang, D., Verma, A., Ramesh, V., Prasad, S. and Franceschini, M., 2025. Adapting Whisper for Streaming Speech Recognition via Two-Pass Decoding. Available at: https://doi.org/10.48550/arXiv.2506.12154

[5] Qu, X., Sun, Z., Feng, S., Chen, C. and Tian, T., 2024. Breaking the Silence: Whisper-Driven Emotion Recognition in AI Mental Support Models. 2024 IEEE Conference on Artificial Intelligence (CAI), Singapore, pp.290–291. Available at: 10.1109/CAI59869.2024.00063

[6] Chou, H-C., 2024. A Tiny Whisper-SER: Unifying Automatic Speech Recognition and Multi-label Speech Emotion Recognition Tasks. 2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Macau, Macao, pp.1–6. Available at: 10.13140/RG.2.2.11583.32166

[7] Baevski, A., Zhou, H., Mohamed, A. and Auli, M., 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477. Available at: 10.48550/arXiv.2006.11477

[8] Kounades-Bastian, D., Schrüfer, O., Derington, A., Wierstorf, H., Eyben, F., Burkhardt, F. and Schuller, B., 2024. Wav2Small: Distilling Wav2Vec2 to 72K parameters for low-resource speech emotion recognition. arXiv preprint arXiv:2408.13920.Available at: 10.48550/arXiv.2408.13920

[9] Zhao, H., Huang, N. and Chen, H., 2024. Knowledge enhancement for speech emotion recognition via multi-level acoustic feature. Connection Science, 36(1), pp.1–18. Available at:10.1080/09540091.2024.2312103

[10] Li, Y., Bell, P. and Lai, C., 2024. Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques. In: 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, China, 27–30 Jan. 2024. IEEE, pp.518–525. DOI**:** 10.1109/SLT61566.2024.10832143

[11] Huang, X., Wang, F., Gao, Y., Liao, Y., Zhang, W., Zhang, L. and Xu, Z., 2024. Depression recognition using voice-based pre-training model. Scientific Reports, 14, p.63556. DOI:10.1038/s41598-024-63556-0

[12] Wang, N. and Yang, D., 2025. Speech emotion recognition using fine-tuned Wav2Vec2.0 and neural controlled differential equations classifier. PLOS ONE, 20(2), p.e0318297. DOI:10.1371/journal.pone.0318297

[13] Liu, Y., Yang, X. and Qu, D. (2024) 'Exploration of Whisper fine-tuning strategies for low-resource ASR', EURASIP Journal on Audio, Speech, and Music Processing, 2024(29). DOI:10.1186/s13636-024-00349-3

[14] Wagner, D., Baumann, I., Engert, N., Lee, S., Noeth, E., Riedhammer, K. and Bocklet, T. (2025) 'Personalized fine-tuning with controllable synthetic speech from LLM-generated transcripts for dysarthric speech recognition', arXiv preprint arXiv:2505.12991. DOI:10.48550/arXiv.2505.12991

[15] Liao, F-T., Chan, Y-C., Chen, Y-C., Hsu, C-J. and Shiu, D-S., 2023. Zero-shot domain-sensitive speech recognition with prompt-conditioning fine-tuning. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Taipei, Taiwan, pp.1–8. DOI:10.48550/arXiv.2307.10274

[16] Chen, L-W. & Rudnicky, A., 2023. Exploring Wav2vec 2.0 fine tuning for improved speech emotion recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. DOI:10.1109/ICASSP49357.2023.10095036

[17] Sampath, A., Tavernor, J. and Mower Provost, E., 2025. Efficient finetuning for dimensional speech emotion recognition in the age of transformers. IEEE, pp.1–14. DOI:10.48550/arXiv.2503.03756

[18] Zhao, Z., Wang, Y. and Wang, Y., 2022. Multi-level Fusion of Wav2vec 2.0 and BERT for Multimodal Emotion Recognition. Proceedings of Interspeech 2022, pp.4725–4729.DOI:10.21437/Interspeech.2022-10230

[19] He, J., Shi, X., Li, X. & Toda, T., 2024. MF-AED-AEC: Speech Emotion Recognition by Leveraging Multimodal Fusion, ASR Error Detection, and ASR Error Correction. Proceedings of IEEE ICASSP 2024, pp.11066–11070. DOI:10.1109/ICASSP48485.2024.10446548

[20] Qin, J., Liu, C., Tang, T., Liu, D., Wang, M., Huang, Q., Xu, Y. and Zhang, R., 2024. Mental-Perceiver: Audio-Textual Multimodal Learning for Mental Health Assessment. arXiv preprint arXiv:2408.12088. DOI:10.48550/arXiv.2408.12088

[21] Li, Y., Bell, P. & Lai, C., 2024. Speech emotion recognition with ASR transcripts: A comprehensive study on word error rate and fusion techniques. 2024 IEEE Spoken Language Technology Workshop (SLT), Macao, pp.518–525. DOI: 10.1109/SLT61566.2024.10832143

[22] Cornell, S., Darefsky, J., Duan, Z. and Watanabe, S., 2024. Generating data with text-to-speech and large-language models for conversational speech recognition. Proceedings of SynData4GenAI 2024, pp.6–10. DOI:10.21437/SynData4GenAI.2024-2

[23] Tran, M., Pang, Y., Paul, D., Pandey, L., Jiang, K., Guo, J., Li, K., Zhang, S., Zhang, X. and Lei, X. (2025) 'A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic data', ICASSP 2025 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10887883

[24] Joshi, Raviraj & Singh, Anupam. (2022). A Simple Baseline for Domain Adaptation in End to End ASR Systems Using Synthetic Data. 244-249. 10.18653/v1/2022.ecnlp-1.28. DOI:10.18653/v1/2022.ecnlp-1.28