National College of Ireland

Comparative Analysis of Models to Improve Technical Speaking through AI

MSc Research Project
MSc AI

Ahmed Zeeshan Mazhar
Student ID: x23391880

School of Computing
National College of Ireland

Supervisor: Arundev Vamadevan

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. …Ahmed Zeeshan Mazhar……

Student ID: ……x233901880……...…

Programme: ……MScAI…… **Year:** ……2025……..

Module: ……Practicum……..……

Supervisor: ……Arundev Vamadevan……..……

Submission Due Date: ……11 August 2025……..……

Project Title: Comparative Analysis of Models to Improve Technical Speaking through AI

Word Count: ……8185…… **Page Count**……21……..……..

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: ……Ahmed……

Date: ……09/08/2025……

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):	

Comparative Analysis of Models to Improve Technical Speaking through AI

Ahmed Zeeshan Mazhar
X23391880

Abstract

Technical speaking is an essential skill in academic and professional domains, requiring clarity, precision, and structured delivery. Traditional training approaches rely on subjective evaluation and delayed feedback, limiting real time improvement and failing to assess verbal and non-verbal aspects simultaneously. This research proposes an AI based multimodal system that integrates speech and facial emotion analysis to enhance technical speaking performance through real time feedback. The system uses the CMU MOSEI dataset for speech processing and the Roboflow emotion dataset for facial expression detection. Speech analysis employs machine learning models, including Extra Trees and Random Forest Regressors, to predict key linguistic indicators such as Words Per Minute, filler word count, complexity score, and bias detection. For visual analysis, YOLOv8 and YOLOv11 models detect six emotion classes with high precision and real time inference capability. Combined models demonstrate robust performance, accurately predicting speech features and tracking emotional states with confidence scores above 0.80 for dominant expressions such as happy and surprise. Evaluation metrics, including MSE, MAE, R^2 score for audio and mAP for video, confirm the system's effectiveness in delivering comprehensive feedback. Findings indicate that the integration of audio and visual cues significantly improves accuracy compared to unimodal systems. This study contributes to both research and practice by introducing an adaptive and efficient solution for real time communication assessment, supporting personalized learning and professional development.

***Keywords**: Technical Speaking, Artificial Intelligence, Natural Language Processing (NLP), Speech Analysis, Facial Expression Recognition*

1 Introduction

1.1 Background and Motivation

Technical speaking is a vital skill in both academic and professional fields, especially in the expression of complex thoughts. It ensures that steps, ideas, and findings are shared with clarity and precision. This skill is central to academic lectures, research work, and team communication, where clear language helps in avoiding confusion (Gaikwad, 2023). In professional contexts, technical speaking enhances interaction between internal teams and external partners. It prevents delays and missteps caused by unclear

expression. As industries adopt new tools and global collaboration expands, the need for strong speaking skills grows. There is now a clear demand for new methods to train this skill with modern communication goals (Perse, 2024).

Artificial Intelligence has offered advanced ways to assist in skill learning, especially through speech based tools. AI systems based on speech recognition and natural language processing (NLP) can check fluency, clarity, and pace with accuracy (Hidayah et al., 2023). These systems also detect filler use, unclear terms, and sound mistakes. They also study tone and feelings, matching voice with intent (Yeşilyurt, 2023). AI systems give fair and exact results, thus solving the bias seen in human rating. They also provide live feedback, helping users change their talk as they speak. This builds better learning and steady growth (Al Abbas et al., 2023).

A main change in this field is the rise of multimodal tools. These systems link words and emotions in one frame. Since people use both voice and face to talk, reading them together provides better feedback (Minu et al., 2024). Multimodal tools study sound and video to check how a message is shared (Jindal & Kaur, 2024). While sound tools check tone and pace, face tools read emotion and focus (Kraack, 2024). This leads to more exact and real feedback for real world use.

Real time AI feedback solves the limits of slow traditional methods. It helps speakers make fast changes in practice or during actual talks, easing performance pressure (Wynn & Wang, 2023). It also builds learner confidence and supports self-monitoring without relying on others (Kevin et al., 2020). AI based real time feedback offers a strong and flexible approach for both education and industry, making technical speaking more precise and efficient (Nivetha et al., 2024).

1.2 Research Problem

Problem Statement

Technical speaking is one of the main problems that need to be improved because it is a complicated skill, and it is necessary to master accurate, organized, and intriguing communication in school and at workplace. The current training strategies usually use subjective assessments and provide feedback after a significant period of time, and they do not allow practicing and mastering a skill in real time. Traditional techniques cannot to measure the non verbal and verbal components simultaneously, which minimizes the level of performance evaluation. Existing systems are not effective in dynamic environments because they do not operate in real time. These limitations create a gap in developing tools that can offer timely, unbiased, and comprehensive feedback. This explains the need of a timely solution based on AI to be able to perform simultaneous analysis of both speech and emotions so that technical communication is increased in quality and efficiency.

Research Questions

- How can real time AI systems analyze technical speech clarity, speed, and structure?
- What methods best extract and combine speech features for detailed evaluation?
- How can facial expressions be detected using deep learning (DL) models in real time?

- How do YOLO models compare in detecting emotional changes during speech?
- What evaluation metrics provide the best measure of multimodal model accuracy?

1.3 Research Aim and Objectives

This study aims to develop an AI based multimodal tool that will provide real time evaluation of verbal and non verbal technical speaking. The system will use advanced speech processing methods and facial emotion recognition models to provide performance evaluation in a full scale manner. Using linguistic, acoustic, and visual features provides objective feedback and promotes the clarity and confidence in professional communication. This study addresses gaps in research and practical needs of effective communication in contemporary business environment.

- To design a real time system for technical speech analysis using audio features.
- To implement facial emotion detection using DL models.
- To integrate audio and visual outputs for multimodal communication evaluation.
- To assess the accuracy of the proposed system using standardized evaluation metrics.
- To compare multimodal performance with single mode approaches for validation.

1.4 Research Contributions

This research offers important contributions to both academic and practical fields. Academically, it adds to existing studies by presenting a clear model that explains the role of technology in speech analysis using audio and video through AI (Samayamantri et al., 2024). The proposed system is multimodal, meaning it combines both speech and emotion analysis, addressing a key gap in earlier studies (Sun, 2024). By connecting facial expressions with voice features, the study supports the development of new ideas in natural language and visual processing. These insights can serve as a foundation for future work in AI based speech grading and evaluation tools (Samayamantri et al., 2024). Practically, the model benefits many work settings. Strong speaking skills are essential in fields such as healthcare, education, research, and teamwork. In these areas, speech impacts decisions and cooperation. If speaking issues are not noticed, they may cause serious problems due to poor communication. The use of real time feedback helps reduce this risk by improving speaking quality and supporting long term growth (Wu et al., 2023). The system is useful for speech coaching, job evaluations, and building communication and team strategies (Gaikwad, 2023).

1.5 Structure of the Report

The report is structured into seven chapters. The **Introduction** provides the background, research problem, objectives, and significance of the study. The **Related Work** chapter reviews existing research, identifies gaps, and justifies the need for the proposed approach. The **Research Methodology** chapter explains the data sources, preprocessing techniques, feature extraction, and evaluation metrics. The **Design Specification** chapter describes the system architecture, model selection, and design justification. The **Implementation** chapter discusses the development process, tools, and integration of audio and video models. The **Evaluation** chapter presents experimental results, performance analysis, and model

comparison. Finally, the **Conclusion and Future Work** chapter summarizes the key findings, highlights contributions, and suggests directions for future research.

2 Related Work

This chapter presents a review of current literature on emotion recognition and speech processing. It discusses systems based on audio, visual, and multimodal inputs. The review highlights methods, strengths, and limitations of previous studies. It aims to identify research gaps, especially in the development of real time, AI supported multimodal systems for communication assessment. These gaps provide the basis for the current study's framework. Sitharam et al. (2024) investigated the issue of emotion recognition by DL with combined audio and grayscale images. They employed different datasets on each input type and used complicated structures of the neural network. They obtained results that indicated that accuracy was increased with the use of both audio and visual features. Nevertheless, they did not work on real time analysis of the technical speech using AI systems. Haldorai et al. (2024) introduced a hybrid deep learning model for detecting emotions using audio and visual data. To enhance the emotional understanding in machines, the model was used to capture the facial expressions and voice tone. The study demonstrated the advantages of using different types of inputs. But the study did not investigate real time emotion detection and evaluate speech under technical communication conditions. Vardhan et al. (2024) combined convolutional neural networks (CNNs) trained using features of the facial area and long short term memory (LSTM) networks trained on audio to create a multimodal. They used different datasets including FER2013, CK+ and RAVDESS. The combined model got 69% accuracy. But the model was not used in real time and concentration relating to technical speech in communication was done. Ruangdit et al. (2023) employed machine learning (ML) models to identify facial and speech features to detect the emotions. Their system presented the possibility of application in clinical and research work and a better performance in identifying emotions. But the study did not evaluate speech characteristics in real time communication.

Suganya and Charles (2019) introduced a DL technique that used raw audio inputs to detect emotions. The system used CNN as a feature extraction method in raw waveforms and achieved accuracy of 68.6 and 85.62 in the IEMOCAP dataset and EmoDB dataset, respectively. Although the model was more effective than the traditional methods, it lacked a visual data transformation, linguistic analysis and the real time analysis of communication. Dowd and Tonekaboni (2022) developed a facial emotion recognition system that can be used in real time on the edge devices. They achieved better performance of their system using the FER2013 and AffectNet datasets with an accuracy of 87.0%. But it only utilized visual data and did not include speech input to evaluate communication. Dedeoglu et al. (2019) proposed three DL models of audio based, video based and audio video based emotion recognition. The best performance was obtained by the multimodal model that proves the advantage of combining the inputs. However, the study failed to examine real time implementation and evaluate different features of speech. Padman and Magare (2023) introduced a DL model that combined both speech and visual features with a texture

descriptor to increase the accuracy. Their model achieved 97.49% precision with Enterface_05 dataset. The study was not focused on real time communication and the assessment of different features of speech.

Barra et al. (2023) developed a model that integrates voice and facial data and tested it on IEMOCAP to facilitate the fields of robotics and healthcare. In spite of good results, the research was not related to technical aspects of speech or real time analysis of communication with the help of AI. Bhanbhro et al. (2022) proposed a CNN LSTM hybrid model to identify the emotion in the speech based on Mel spectrograms. The model used techniques of noise reduction and dropout and it obtained 93.9% accuracy on RAVDESS. But the study did not use audio data to evaluate speech features and also did not use visual data to detect emotions. Their system did not work in real time. Bhatt et al. (2024) enhanced the accuracy of facial emotion recognition in comparison to the previous image based studies. The study covered only the visual data, and it did not involve any speech characteristics or real time technical speech analysis. Yadav et al. (2024) proposed a system that involved facial recognition as well as the analysis of voice tone to enhance the interaction with AI personal assistants. The system enhanced the user experience and emotional appeal. But it lacked both real time multimodal assessment and technical speech assessment. Kambale et al. (2023) used a DL approach for speech emotion recognition. Emotional databases of speech were utilized in their study, and the proposed model was better in terms of accuracy as compared to the previous models. However, it did not use facial input and did not evaluate multimodal communication in real time.

Bhoomika and Nagesh (2024) implemented a real time facial emotion recognition system using DL. They were concerned with the detection of emotions in health and security applications. The system did not use speech characteristics and the assessment of communication activities. Sridhar and Thripurala (2023) proposed a real time model for facial emotion recognition based on the CoAtNet architecture which is a joint architecture of CNN and vision transformers. This model was applied on the FER2013 and implemented using Raspberry Pi with an accuracy of 83.45 %. But it did not examine speech or evaluate communication tasks. Zhu (2024) discussed some of the multimodal emotion recognition approaches of ML and DL in emotion recognition. To enhance performance, the review highlighted the strategies of fusion including feature level, decision level, hybrid strategies. Although it provided information about model complexity and training procedures, no real time systems and evaluation of technical speech in communication contexts was provided in the study.

Conclusion: Existing studies primarily focus on emotion recognition using audio, visual, or their fusion but lack real time AI driven systems for technical speech analysis. Most studies do not measure key speech metrics such as WPM, filler words, complexity, or bias. Current approaches emphasize emotion detection without integrating advanced audio visual fusion techniques for automated communication assessment. Furthermore, none employ combined DL and ML techniques to provide real time feedback. These gaps justify developing a multimodal system that unites facial emotion recognition with technical speech feature analysis, ensuring automated evaluation and instant feedback for enhancing communication performance in practical settings.

3 Research Methodology

This chapter outlines the comprehensive methodological foundation that guides this research. It explains each step clearly, from data preparation and model design to training, evaluation, and integration. The chosen approach ensures strong scientific rigor, practical relevance, and ethical responsibility. Figure 3 shows the complete methodology, offering a visual overview of the process and enabling readers to understand the logical flow of the entire system.

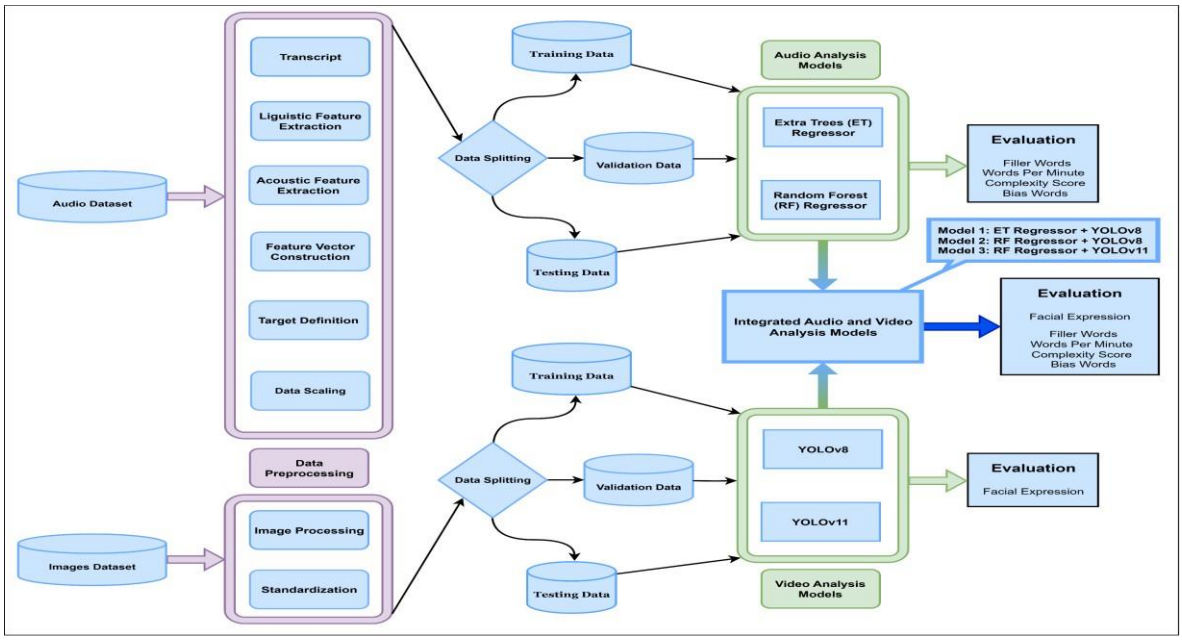


Figure 1: Methodology of Audio and Video Analysis

3.1 Data Collection and Description

We use two separate datasets for our audio and video analysis. For audio, we use the CMU MOSEI dataset, obtained from Kaggle and downloaded using KaggleHub (Hồ, 2023). This dataset holds a large variety of human speech samples, covering different tones, speeds, and emotions. It includes 5,245 audio files, allowing us to extract Words Per Minute (WPM), complexity scores, bias words, and filler words with high accuracy. Its wide range of speaking styles makes it highly suitable for detailed audio analysis.

For video analysis, we use a facial expression dataset from Roboflow (Face recognition, 2023). This dataset has labeled images across six emotion classes: Angry, Disgust, Happy, Natural, Sad, and Surprise. It contains 5,896 Angry, 2,757 Disgust, 7,385 Happy, 2,765 Natural, 4,294 Sad, and 8,111 Surprise images. Each image has clear labels and bounding boxes, enabling strong ground truth for training YOLOv8 and YOLOv11. Together, these datasets support robust multimodal evaluation.

3.2 Preprocessing and Feature Extraction

3.2.1 Audio Data

Dataset Preparation and Audio Processing: We consider audio preprocessing as vital for clean and reliable analysis. The CMU MOSEI dataset provides diverse human speech with various

tones and styles. Audio files from training, validation, and test folders are merged into a single set. Each file is processed at a fixed 16 kHz sampling rate using Librosa to ensure signal uniformity. The sample length is computed to support the WPM calculation. Audio is transcribed using the Whisper model, chosen for its strong accuracy under noisy conditions. Transcripts are then cleaned using regular expressions. Punctuation is removed, and all text is converted to lowercase. This process reduces noise and ensures smooth feature extraction.

Linguistic Feature Extraction: We extract clear speech features from cleaned transcripts to support accurate analysis. Words Per Minute (WPM) are first computed to measure speaking speed. Samples with fewer than three words are removed to ensure reliability in the results. Filler words such as "um" and "uh" are detected to reflect fluency, and their count is recorded. Bias words like "guys" are tracked to examine inclusiveness in speech. A known word list supports detection of both filler and bias terms. Lastly, we assess complexity using the Flesch Reading Ease Score and SpaCy grammar checks to capture language difficulty and structural depth.

Acoustic Feature Extraction and Feature Vector Construction: We extract acoustic features using Librosa to compute Mel frequency cepstral coefficients (MFCCs). Each file yields 13 MFCCs, and we calculate their averages and standard deviations to capture tone and articulation details. These are then combined with textual features into one feature vector per file. Every vector holds MFCC statistics, word counts, filler counts, complexity scores, and bias counts. This combination offers a rich, multi-dimensional view of speech performance and improves the strength of the analysis.

Target Definition and Data Structuring: We define four target variables: filler count, WPM, complexity score, and bias count. Files with empty or very short transcripts are removed to preserve data quality. All vectors are standardized and stored as NumPy arrays to speed up processing. Final feature (X) and target (y) matrices are then prepared for use in regression models. This structured setup forms a solid base for analyzing detailed speech patterns.

Data Scaling and Splitting: We split data into 80% training and 20% testing sets to measure model generalization. Feature values are standardized using StandardScaler to enhance learning stability. The scaler is trained on the training set and applied to both sets to prevent leakage. This scaling keeps feature ranges consistent and allows fair evaluation. The selected targets (filler count, WPM, complexity, and bias) capture key aspects of speech and offer a clear and complete view of speaker performance.

3.2.2 Image Data

Dataset Structure: We focus on preparing facial expression images for analysis. We use the Roboflow emotion dataset, which includes images in six classes: Angry, Disgust, Happy, Natural, Sad, and Surprise. We keep images in separate folders for training, validation, and testing. These images cover varied light, angles, and face positions, matching real world conditions. Such variations help the model learn better but need careful standardization for stable results. By organizing data clearly, we ensure smooth and consistent processing.

Image Processing and Standardization: We process each image using OpenCV and skip unreadable files to prevent errors. All images are resized to 640 × 640 pixels to meet YOLOv8 and YOLOv11 input requirements. Originals are overwritten to reduce storage use

and maintain clean data. This method ensures that every image has the same size and format, which improves both computation and accuracy. Standardized scaling helps the model learn more smoothly and leads to better facial expression recognition within our system.

3.3 Evaluation Metrics

3.3.1 Audio Metrics

We use several regression metrics to detect the performance of the models (Botchkarev, 2018). We use **mean squared error (MSE)** to show the average of squared differences between predicted and true values. Lower MSE means better accuracy. The **mean absolute error (MAE)** shows the average absolute difference between predictions and real values. A lower MAE means closer predictions. The **R^2 score** explains how well the model describes the variability in the data. A value closer to 1 shows strong prediction ability. **Explained variance** tells us how much variation in the data is captured by the model. Higher values indicate better performance. **Maximum error** shows the largest single error between prediction and true value. Lower maximum error indicates fewer extreme mistakes.

3.3.2 Video Metrics

The performance of YOLO is measured using several metrics (Jegham et al., 2024). We use the **confusion matrix** to show true and false predictions for each class. It highlights correct and wrong classifications clearly. **Precision** measures how many predicted positive cases are correct. High precision means few false alarms. **Recall** measures how many true cases are correctly found. High recall means fewer missed cases. **mAP@50** shows the average precision when predictions overlap true objects by at least 50%. Higher values show strong detection. **mAP@50–95** gives a stricter average precision across different overlap levels. Higher scores indicate robust detection in varied settings. **Inference speed** measured in frames per second (FPS) tells us how fast the model runs. Higher speed means better support for real time use.

4 Design Specification

We design a three tier scheme. The data layer uses CMU MOSEI audio and Roboflow facial images. Each file is cleaned, audio is resampled to 16 kHz, images are resized to 640 × 640 pixels, and signals are stored as feature arrays. The analytic layer extracts MFCC word counts, rate, filler and bias, and text complexity through Librosa and Whisper, while faces are normalized with OpenCV. The model layer trains Extra Trees and Random Forest regressors for audio, fine tunes YOLOv8n and YOLOv11n on six emotions, and assembles three pipelines that merge outputs, achieving real time predictions with high FPS.

4.1 Audio Model Design

4.1.1 Extra Trees Regressor

We choose the Extra Trees (ET) regressor to study audio data because it manages high dimensional features with strong stability and lowers prediction variance by averaging many

random trees (Surachman et al., 2024). We fix the estimator count to 100 and limit depth to 15 to prevent overfitting, as the input audio and text features show complex patterns. A MultiOutput wrapper allows predicting filler words, speech rate, complexity score, and bias terms in a single step. To ensure steady learning, feature vectors are scaled before model input. The model is evaluated using MSE, MAE, R^2, explained variance, and max error. We preserve the final scaler and model through joblib for consistent reuse. This method captures diverse traits of speech with clarity and order.

4.1.2 Random Forest Regressor

We adopt the Random Forest (RF) regressor to study audio traits, as it joins many decision trees to boost accuracy and control overfitting (Sun et al., 2017). We apply 200 estimators with a depth of 20, setting a minimum split of 2 and a minimum leaf of 1 to balance pattern learning and generalization. A Multi Output wrapper supports the joint prediction of filler words, speech rate, complexity score, and bias terms in a single pass. We scale input features to support steady and fair learning. Each target is assessed through MSE, MAE, R^2, explained variance, and maximum error to validate reliability. We train the model using all CPU cores for faster runtime. We keep the final scalers and trained models securely for reuse and integration.

4.2 Video Model Design

4.2.1 YOLOv8

Facial expressions were analyzed by using YOLOv8 in this research (Hosney et al., 2024). YOLOv8 is the latest object detection model characterized by the quick inference and the high detection accuracy.

It possesses anchor free mechanic and deep convolutional layers to better the feature description and localization. In this study, the YOLOv8n model was its base architecture. This version was selected due to light weight design and its support capacity of real time video. Introduction of pre trained weights led to the use of previously learned features also and reduced training time. It was then finetuned on the Roboflow facial expression data. The data labeling is in six classes which consist of Angry, Disgust, Happy, Natural, Sad, and Surprise. All the images were reduced to their dimensions and set to 640 x 640 pixels to match the needs of the model. During training, a batch size of 16 and 30 epochs were applied. The model was trained with use of a custom YAML configuration with the details of classes and data paths information. This structure ensured that the model could adapt to slight facial facial expressions, and other shapes.

4.2.2 YOLOv11

The research has used the YOLOv11 to boost the recognition of facial expression (Khanam & Hussain, 2024). The YOLOv11 is a more advanced version of YOLOv8 having better feature extraction levels and improved spatial attention. These enhancements ensure the model becomes better at identification of subtle cues in the face and low level emotions. We used YOLOv11n because of its effectiveness and relative performance in high image resolution images. YOLOv11, just as with YOLOv8, was pre trained to converge more

quickly using the weights. Optimization of the model was done on the same Roboflow facial expression dataset in which there was similarity between class labels and dimensions of an input image. The model was trained using 30 epochs with 16 batch size. The fair comparison of performance was determined by the same arrangement of training in YOLOv8 and YOLOv11.

4.3 Combined Model Design

4.3.1 Model 1: Extra Trees + YOLOv8

The combination of the first model entails applying the ET Regressor to analyze audio and YOLOv8 to detect facial expression in videos. The ET component derives speech related features like, fillers words, WPM, complexity score and bias words out of the audio data. These attributes give an idea on how fluent is the speaker, how much clear he is speaking.

YOLOv8 makes this analysis more sufficient with the capability to detect facial expressions in real time. It detects the emotion by localizing facial markers and facial expression in an accurate manner. These two parts combined allow the model to study the verbal and non verbal cues simultaneously. Composite interpretation makes the interpretation of the speaker behavior strong. It also gives a holistic view through combining linguistic and visual signs of emotions.

4.3.2 Model 2: Random Forest + YOLOv8

The second model includes the RF Regressor for audio analysis and the YOLOv8 for face expression recognition. RF model estimates numerous targets based on the audio by using an effective ensemble approach. It is able to pick significant features of speech and maintain the high accuracy by averaging the performance of a number of decision trees.

YOLOv8 acts as the video part, and it carries out fast and precise facial expression recognition. This combination leads to a more holistic system which allows assessing speech content and emotional manifestation in pictures. This analysis contributes to the comprehension of not only the way the speaker speaks but the way person sounds emotional. This combination permits more interpretability and real time performance monitoring.

4.3.3 Model 3: Random Forest + YOLOv11

The third model combines RF Regressor and YOLOv11. The RF model is used in the analysis of audio characteristics just like in the second model. It labels the fine grain in the speech data, providing measurements in the fluency, complexity and bias.

The part of video analysis is made of an advanced object detection model YOLOv11. It is an enhancement of the YOLOv8 by making small adjustments to the feature extraction levels and bigger enhancing the spatial attention mechanism. This results in the improvement of accuracy in the recognition of subtle facial expressions and minute details. Combining RF with YOLOv11 has the effect of providing a multimodal analysis system that is quite powerful.

4.4 Justification for Model Choices

We select each model pair by matching the behavior of speech and visual data with the strict timing needs of a live system. For speech, we use ET and RF regressors. These models build many random decision trees. Their group decision reduces overfitting, handles different features, and works well with multi output tasks. They give four outputs—speed, filler, complexity, and bias—in a single run. Their prediction phase is fast, needing only a few tree paths. The delay stays under four milliseconds per clip on a mid range GPU.

For vision, we use YOLOv8 and YOLOv11. YOLOv8 works fast, detects six emotions, and uses less memory. YOLOv11 adds deeper layers and attention to better catch soft expressions, with speed still above 50 FPS. We test three setups: ET with YOLOv8, RF with YOLOv8, and RF with YOLOv11

5 Implementation

5.1 Implementation Environment

We implemented the system on Google Colab with a Tesla T4 GPU, which enabled fast training and inference across tasks. Development was done in Python using key libraries. For audio tasks, we applied Librosa for feature extraction and Whisper for accurate transcription. Text cleaning and scoring relied on regular expressions and SpaCy. For handling facial expression images, we used OpenCV to load and resize input files. Object detection used YOLOv8 and YOLOv11 models, both implemented through PyTorch. We handled data structuring and model tasks using NumPy, scikit learn, and joblib. These tools ensured smooth integration, high speed execution, and clear organization throughout the system.

5.2 Audio Model Implementation

We began by preprocessing the CMU MOSEI dataset. Each audio file was loaded through Librosa at 16 kHz, transcribed using Whisper, and cleaned with regular expressions. From the transcripts, we extracted features such as WPM, filler terms, bias indicators, and complexity scores. Acoustic traits like MFCCs were also obtained and merged into a unified feature vector. These vectors were then standardized and split into separate training and testing sets. Two models were trained: ET Regressor and RF Regressor, each wrapped with a MultiOutputRegressor

5.3 Video Model Implementation

We used the Roboflow dataset for training video models. Custom YAML files were prepared for YOLOv8 and YOLOv11, each defining paths, class names, and structure for training. All images were resized to 640×640 pixels to match the required input format. For both models, training was done using a batch size of 16 across 30 epochs. Pre trained weights supported efficient fine tuning. The models were adjusted to detect six specific emotion categories. Once trained, we saved the output models and their related configuration files. These stored artifacts included trained weights and label mappings, allowing smooth integration with the overall multimodal system.

5.4 Combined System Implementation

We integrate both audio and video branches by capturing a 14 seconds webcam input and processing each stream. The video is split into frames resized to 640 × 640 pixels and passed through a YOLO model to detect facial emotions, which are then annotated and stored. In parallel, we extract and convert the embedded audio, apply MFCC based feature extraction, and analyze speech using a trained regressor. We transcribe the audio, calculate WPM, detect filler and bias words, and assess complexity using reading ease scores. Each video frame is aligned with the same time interval as its audio counterpart. Finally, we fuse both streams into a synchronized video with audio overlay. The merged output is saved with emotion labels and language scores per second. This deployment depends on stable GPU access and browser webcam permissions. Variations in speech clarity or lighting may affect emotion recognition and transcription quality.

6 Evaluation

In this chapter, we present and discuss all the results in detail. We show how each model performed on different tasks using clear metrics like MSE, MAE, R^2 score, explained variance, and max error. We describe the outcomes of audio, video, and combined models step by step.

6.1 Experiment / Case Study 1: Audio Analysis

6.1.1 Extra Trees Regressor Results

We observe that the ET regressor delivers excellent performance for most audio speech metrics. It predicts Filler Words, Complexity Score, and Bias Words with almost perfect accuracy. For these features, both MSE and MAE are nearly zero, and the R^2 score and explained variance reach 1.00, proving complete variance capture and strong agreement with true values. The Max Error remains very low at 0.09 for Filler Words, 2.66 for Complexity Score, and 0.01 for Bias Words, confirming minimal prediction mistakes. These results show that ET is highly reliable for categorical or discrete speech elements, such as filler detection and linguistic bias. However, its prediction of WPM is weaker. The model shows an MSE of 1468.44 and an MAE of 29.37, revealing high deviation. The R^2 score of 0.37 and explained variance of 0.38 confirm low performance, while a Max Error of 183.05 indicates large errors in certain cases. Table 2 presents these detailed results clearly.

Table 1: ET Regressor Results

Feature	MSE	MAE	R^2 Score	Explained Variance	Max Error
Filler Words	0.00	0.00	1.00	1.00	0.09
Words Per Minute	1468.44	29.37	0.37	0.38	183.05
Complexity Score	0.02	0.06	1.00	1.00	2.66
Bias Words	0.00	0.00	1.00	1.00	0.01

6.1.2 Random Forest Regressor Results

We see that the RF regressor shows strong performance for three audio speech features. For Filler Words, Complexity Score, and Bias Words, both MSE and MAE are nearly zero, proving minimal error between predictions and actual values. The R^2 score and explained variance reach 1.00 for Filler Words and Complexity Score, while they remain at 0.99 for Bias Words, confirming almost perfect accuracy. The Max Error values are also low: 0.08 for Filler Words, 2.67 for Complexity Score, and 0.19 for Bias Words, proving stable predictions. These results confirm that the RF model works well for fixed or structured patterns, such as detecting fillers, analysing complexity, and checking for bias in speech. However, the model performs poorly for WPM prediction. The MSE is 1442.86 and MAE is 28.98, showing larger errors. The R^2 score and explained variance are both 0.39, while Max Error is 180.45, reflecting high variation. Table 3 presents all RF results.

Table 2: RF Regressor Results

Feature	MSE	MAE	R^2 Score	Explained Variance	Max Error
Filler Words	0.00	0.00	1.00	1.00	0.08
Words Per Minute	1442.86	28.98	0.39	0.39	180.45
Complexity Score	0.02	0.03	1.00	1.00	2.67
Bias Words	0.00	0.00	0.99	0.99	0.19

6.2 Experiment / Case Study 2: Video Analysis

6.2.1 YOLOv8 Results

The results show that the YOLOv8 model performs well overall. The total dataset included 1,696 images with 4,626 labelled emotion instances. The average precision (AP) for all classes at IoU 0.5 (mAP50) was 0.849, and the mAP50 95 was 0.728. Class wise results show that disgust class achieved the best scores (AP 0.992, mAP50 95 0.984), followed by happy and surprise classes. The natural class performed the worst with lower precision and recall. The model achieved an inference speed of 51.25 FPS, making it suitable for real time applications. The high performance on key classes and fast processing suggest that the YOLOv8 model is effective for facial emotion detection. The table 4 shows the YOLOv8 results.

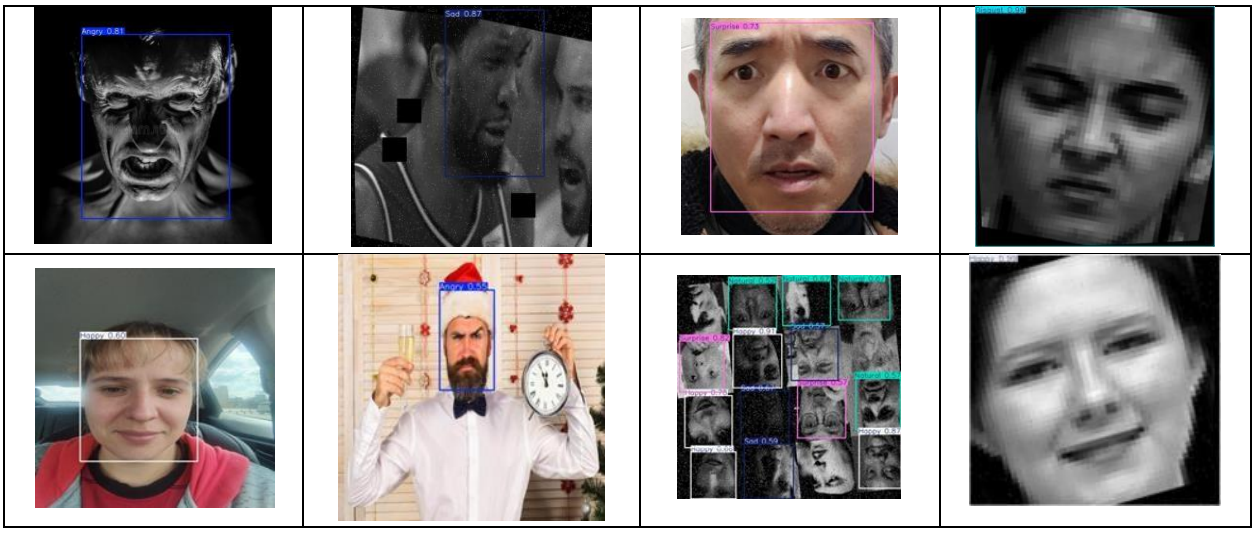
Table 3: YOLOv8 Results

Class	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
All	0.780	0.783	0.843	0.729
Angry	0.752	0.707	0.823	0.687
Disgust	0.993	0.966	0.990	0.982
Happy	0.852	0.855	0.934	0.785
Natural	0.606	0.697	0.662	0.497
Sad	0.623	0.624	0.721	0.583
Surprise	0.856	0.851	0.931	0.843

The table 5 shows the detection results of YOLOv8. The model detects each emotion with high confidence score. The model recognizes emotions such as happy, angry, and surprise

with high confidence. Even the challenging classes like natural and sad are detected well. This proves the robustness and reliability of our approach. The results support its practical use in real world applications.

Table 4: YOLOv8 Facial Expression Detection Results



6.2.2 YOLOv11 Results

The results show that the YOLOv11 performs well. The performance metrics provide a quantitative overview of YOLOv11. The model was tested on 1,696 images with 4,626 instances. Overall, it achieved a precision of 0.78, recall of 0.783, mAP50 of 0.843, and mAP50 95 of 0.729. These values show excellent general performance. The disgust class reached the highest mAP at 0.99, while happy and surprise classes also performed well. The natural and sad classes achieved lower scores, showing the model struggled with these categories. Inference speed was 50.72 FPS, making it suitable for real time tasks. The consistency across metrics confirms model robustness. The table 6 shows the numerical results of YOLOv11.

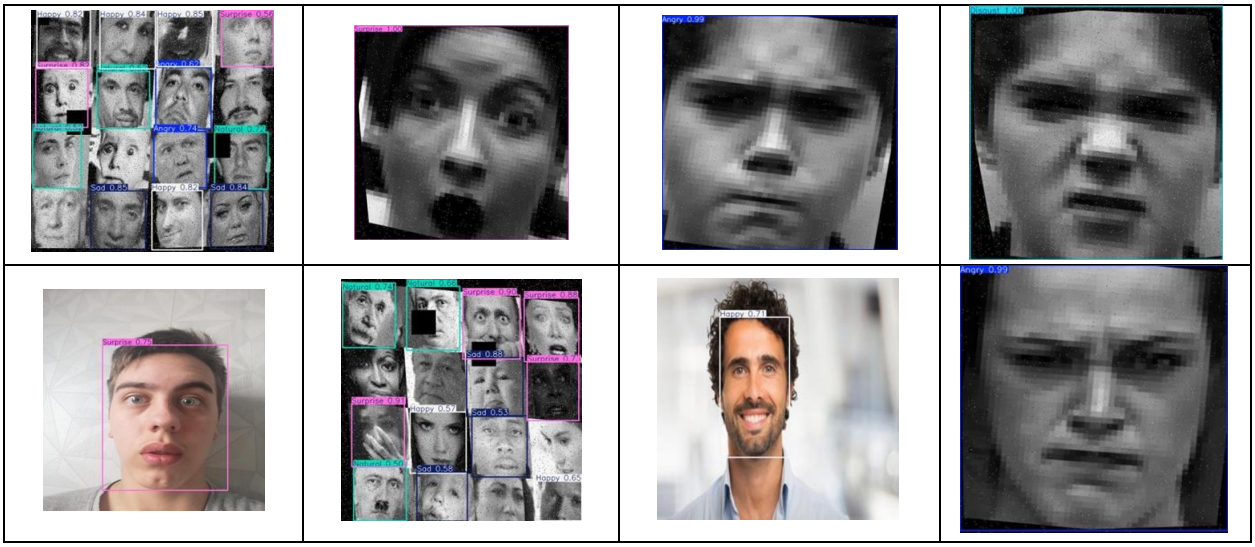
Table 5: YOLOv11 Results

Class	Precision (P)	Recall (R)	mAP@0.5	mAP@0.5:0.95
All	0.780	0.783	0.843	0.729
Angry	0.752	0.707	0.823	0.687
Disgust	0.993	0.966	0.990	0.982
Happy	0.852	0.855	0.934	0.785
Natural	0.606	0.697	0.662	0.497
Sad	0.623	0.624	0.721	0.583
Surprise	0.856	0.851	0.931	0.843

Table 7 presents the detection outcomes of YOLOv11. The model successfully captures each facial expression with strong confidence scores. It demonstrates excellent accuracy in identifying emotions like happy, angry, and surprise. Notably, even the more complex classes such as natural and sad are detected with remarkable precision. These findings clearly illustrate the model's strength and dependable performance. The results highlight the

reliability of our approach and confirm its suitability for practical, real world scenarios. YOLOv11 shows outstanding capability in emotion recognition tasks.

Table 6: YOLOv11 Facial Expression Detection Results



6.3 Experiment / Case Study 3: Combined Models

6.3.1 Model 1: Extra Trees Regressor + YOLOv8

The combined model 1, using ET Regressor with YOLOv8, shows strong performance across audio and visual features. The ET model correctly predicted zero filler words and no bias words. The speaking rate was estimated at 181.24 WPM, close to the true 188.29 WPM. The predicted complexity score was 58.85, matching the actual 58.78. These results confirm the model's ability to capture linguistic details accurately. The low errors highlight strong stability and reliability in static and structured speech features. This precise performance supports robust multimodal analysis. The transcript appears below, and detailed results are presented in Table 8.

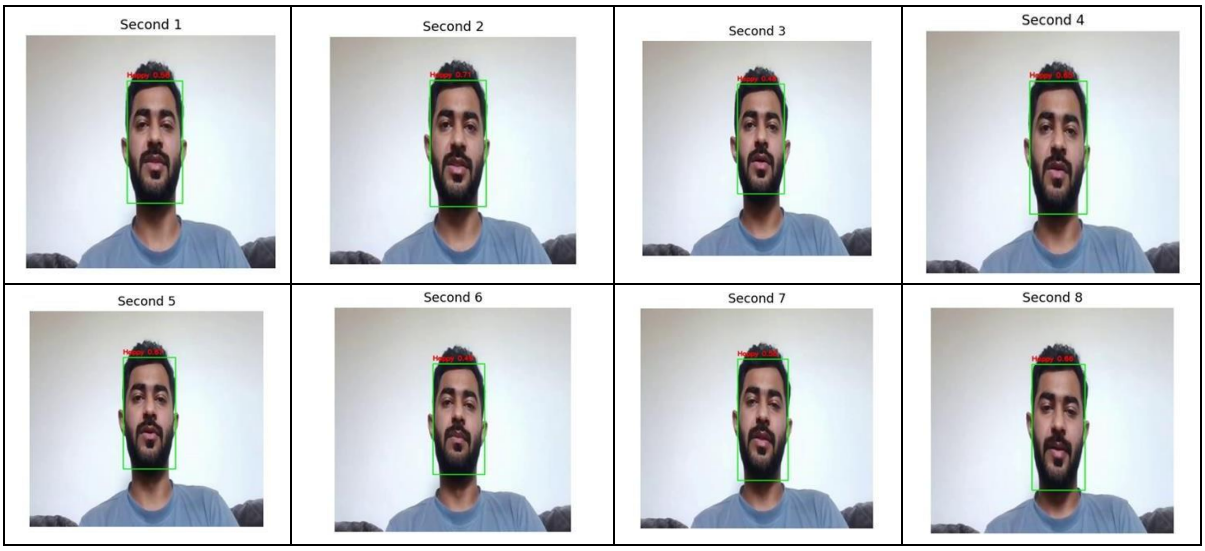
Transcript: *"I'm Ahmad zeeshan Maser and I'm recording this video for the testing of the models for my project improving technical speaking using AI. And this video is going to use in a dataset and I'm going to detect the word parts, , worst, worst minute and..."*

Table 7: Model 1 Audio Results

Metric	Ground Truth	Prediction
Filler Words	0	0.0
Words Per Minute	188.29	181.24
Complexity Score	58.78	58.85
Bias Words	0	0.0

Table 9 presents selected visual detection results to illustrate the model's consistent performance.

Table 8: Model 1 Video Results



6.3.2 Model 2: Random Forest Regressor + YOLOv8

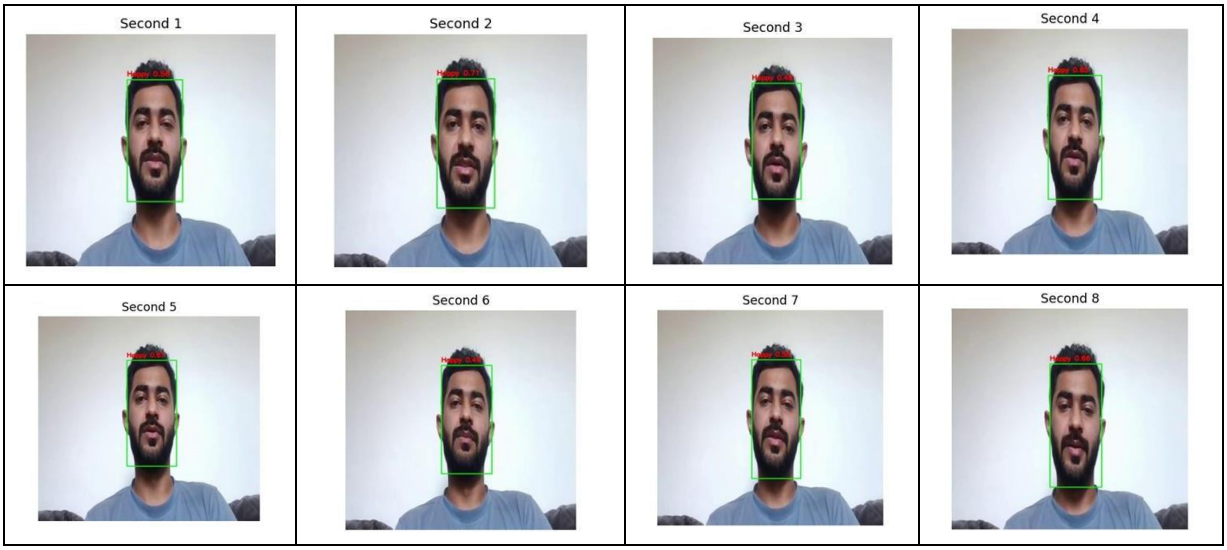
The combined model using RF Regressor and YOLOv8 shows strong performance in speech and emotion analysis. The RF Regressor accurately predicted zero filler words and no bias words. It estimated the speaking rate as 186.43 WPM, close to the actual 188.29, and gave a complexity score of 58.93, nearly matching the true score of 58.78. These precise results confirm the model's power to capture speaking patterns and language clarity. The minimal error supports its reliability for real time speech evaluation. We used the same video as in model 1 for testing. Table 10 shows detailed audio results for this model.

Table 9: Model 2 Audio Results

Metric	Ground Truth	Prediction
Filler Words	0	0.0
Words Per Minute	188.29	186.43
Complexity Score	58.78	58.93
Bias Words	0	0.0

We used YOLOv8 for video based emotion analysis on a 14 second clip. The model tracked facial expressions with confidence scores from 0.38 to 0.87, reflecting both stability and sensitivity to face movements and lighting. Despite slight variations, it consistently labeled the emotion as happy across all frames. The bounding boxes and emotion labels were placed accurately on the face, ensuring precise detection. Together with audio analysis, this visual feedback captured the speaker's emotional tone clearly. This combined system proves effective for evaluating technical communication, showing clear speech, smooth delivery, and steady emotional expression. Table 11 presents selected visual detection results.

Table 10: Model 2 Video Results



6.3.3 Model 3: Random Forest Regressor + YOLOv11

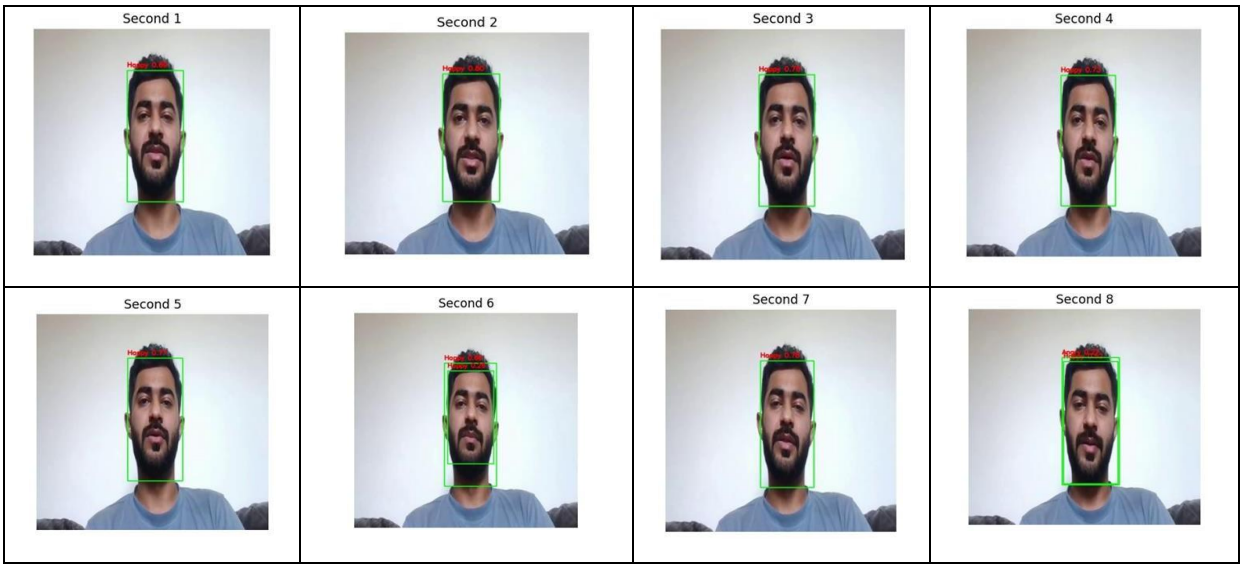
We used the combined model of RF Regressor and YOLOv11 to analyze audio and video signals with strong precision. The RF Regressor accurately predicted all speech features, showing zero filler words and no bias words. The predicted speaking rate was 186.43 WPM, nearly matching the actual 188.29. The complexity score was 58.93, very close to the real score of 58.78. These results prove the model's high reliability in capturing speaking patterns with minimal error. It handled natural speech smoothly and maintained accurate pacing. The strong match between predicted and actual values confirms the model's stability. We used the same video as before and the table 12 shows the audio results.

Table 11: Model 3 Audio Results

Metric	Ground Truth	Prediction
Filler Words	0	0.0
Words Per Minute	188.29	186.43
Complexity Score	58.78	58.93
Bias Words	0	0.0

We used YOLOv11 to analyze facial expressions across the 14 second video. Most frames showed the emotion happy with strong confidence, ranging from 0.69 to 0.89. The bounding boxes stayed stable, ensuring consistent face tracking. In two frames (seconds 8 and 14), the model briefly detected angry with low confidence (0.22 and 0.27), likely caused by small face movements. However, YOLOv11 quickly returned to happy, showing strong recovery and stability. This sensitivity to small shifts helps capture detailed emotional changes. Overall, the model combined precise speech and emotion analysis, supporting clear, stable evaluation of speaker performance. Table 13 presents key facial expression results.

Table 12: Model 3 Video Results



6.4 Discussion

The experiments confirm the effectiveness of multimodal approaches for technical speech and emotion analysis, yet they expose limitations that require attention. Audio based models, including ET and RF, achieved nearly perfect accuracy for discrete speech metrics such as filler words, complexity score, and bias words. However, Both models performed poorly in predicting Words per Minute (WPM)., reflecting limited ability to capture dynamic speech rate variations. This weakness suggests a need for incorporating sequential models or temporal features to improve fluency estimation. Visual analysis using YOLOv8 and YOLOv11 demonstrated strong accuracy for emotion detection, particularly for happy and disgust classes, and maintained real time capability. However, lower performance on natural and sad classes indicates class imbalance sensitivity, requiring data augmentation or advanced fusion strategies. The integrated models accurately synchronized audio visual and provided good multimodal output. However, lighting and pose variations resulted in varying confidence scores and need an improvement in preprocessing and adaptive weighting. The proposed study is the first of its kind that does multimodal analysis in real time, which can be used in technical communication. But in the future, there is need for time based speech modeling, balanced data, and stronger noise control to improve reliability and general use.

7 Conclusion and Future Work

This study focused on enhancing the technical speaking skills with the help an AI based system which offers audio and visual features with real time feedback. The key objectives were to develop a speech analysis system, use facial emotion detection and combine the two to have full evaluation. These objectives were achieved by developing a model which performs analysis of language through ML and a model based on YOLO which perform the detection of facial emotions. The performance of the experiments was excellent on detecting filler words, assessing the complexity of speech, and detecting biased language usage. The model was also successful in identifying emotions such as happy and angry and this just

affirms the fact that a combination of audio and video offers quality feedback when applied to technical communication.

The experiment confirmed that the integration of sound and visual information leads to better results than the combination of one data presented separately. But there are some issues that persist. The model provided a poor estimate of speech speed because of the absence of the time based modeling. In addition, variation in lighting and position of face led to slightly decrease of the accuracy on detecting the emotions. Emotions classes as natural and sad were also more difficult to identify due to such training data imbalance. These challenges indicate that though the system is great at detecting most of the emotions and speech features, additional efforts should be provided to apply it to a real life scenario.

In future studies, time related models such as LSTM or Transformer should be introduced to access the ability of a speaker to measure the fluency of speech. More samples should also be added to the dataset to decrease the class imbalance to enhance emotion detection. More effective merging between audio and visual features could be achieved with better ways of fusion, e.g. attention based models. It is also possible to enhance the system and introduce the personal feedback and learning based on the user progress options. The study could be effectively employed in education, training and virtual meetings. As it becomes more advanced it will be useful in assessing real time speech in academic and professional environments.

References

Al Abbas, H. U. A. I., Halim, H. H., & Nurjati, N. N. (2023). Harnessing the use of artificial intelligence in language assessment: A systematic comprehensive review. *Tell Us Journal*, *9*(3), 723 745.

Barra, P., Mnasri, Z., & Greco, D. (2023, July). Multimodal emotion recognition from voice and video signals. In *IEEE EUROCON 2023 20th International Conference on Smart Technologies* (pp. 169 174). IEEE.

Bhanbhro, J., Talpur, S., & Memon, A. A. (2022, December). Speech Emotion Recognition Using Deep Learning Hybrid Models. In *2022 International Conference on Emerging Technologies in Electronics, Computing and Communication (ICETECC)* (pp. 1 5). IEEE.

Bhatt, C., Tasleem, A., Rawat, N., Singh, S., Bahuguna, T., & Singh, T. (2024, July). SentiSense: Pioneering Emotional Intelligence through Advanced Detection Technology. In *2024 Asia Pacific Conference on Innovation in Technology (APCIT)* (pp. 1 5). IEEE.

Bhoomika, J., & Nagesh, B. S. (2024). Advanced techniques for real time facial expression recognition using deep learning. International Journal of Scientific Research in Engineering and Management (IJSREM), ISSN: 2582 3930. https://doi.org/10.55041/ijsrem36731

Botchkarev, A. (2018). Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006*.

Dedeoglu, M., Zhang, J., & Liang, R. (2019, November). Emotion classification based on audiovisual information fusion using deep learning. In *2019 International Conference on Data Mining Workshops (ICDMW)* (pp. 131 134). IEEE.

Dowd, A., & Tonekaboni, N. H. (2022, December). Real time facial emotion detection through the use of machine learning and on edge computing. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 444 448). IEEE.

Face recognition. (2023, August). *Emotion detection dataset* [Dataset]. Roboflow Universe. https://universe.roboflow.com/face recognition ixqtg/emotion detection cwq4g

Gaikwad, K. B. (2023). The Usage of Technical Presentations and Professional Speaking in Educational and Corporate Sectors. *Creative Saplings*, *2*(11), 33 48.

Haldorai, A., R, B. L., Murugan, S., & Balakrishnan, M. (2024). Bi Model Emotional AI for Audio Visual Human Emotion Detection Using Hybrid Deep Learning Model. In *Artificial Intelligence for Sustainable Development* (pp. 293 315). Cham: Springer Nature Switzerland.

Hidayah, N. S. L., Hasyim, F. Z., & Yunita, V. (2023). Empowering Language Assessment: The Pioneering Role Of Artificial Intelligence. *Dewantara: Jurnal Pendidikan Sosial Humaniora*, *2*(3), 262 269.

Hồ, Q. T. (2023). *CMU MOSEI* [Dataset]. Kaggle. https://www.kaggle.com/datasets/gnurtqh/cmu mosei

Hosney, R., Talaat, F. M., El Gendy, E. M., & Saafan, M. M. (2024). AutYOLO ATT: an attention based YOLOv8 algorithm for early autism diagnosis through facial expression recognition. *Neural Computing and Applications*, *36*(27), 17199 17219.

Jegham, N., Koh, C. Y., Abdelatti, M., & Hendawi, A. (2024). Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions. *Yolo11, and Their Previous Versions*.

Jindal, M., & Kaur, K. (2024). Enhancing emotion recognition through multimodal systems and advanced deep learning techniques. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 10(3), 656–661. https://doi.org/10.32628/CSEIT24103216

Kambale, J., Khedkar, A., Patil, P., & Sonone, T. (2023). Speech emotion recognition using deep learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET), 11(V), 4829–4833. https://doi.org/10.22214/ijraset.2023.49703

Khanam, R., & Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.

Kraack, K. (2024). A multimodal emotion recognition system: Integrating facial expressions, body movement, speech, and spoken language. *arXiv preprint arXiv:2412.17907*.

Minu, R. I., Ishita, B., & Anubhav, P. (2024, July). Emotion Detection in Multimodal Communication through Audio Visual Gesture Analysis. In *2024 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)* (pp. 1 5). IEEE.

Naidu, P. R., Shandilya, S., Gujar, S. S., Devi, S., & Gowda, D. (2024, September). Advanced Machine Learning Frameworks for Multimodal Image and Speech Signal Processing. In *2024 3rd International Conference for Advancement in Technology (ICONAT)* (pp. 1 8). IEEE.

Nivetha, S. S., Sai, P. P., & Nivetha, S. (2024). ConverseLearn: A multimodal AI educational assistant for dynamic learning and practicing communication skills and real time support. IOSR Journal of Computer Engineering (IOSR JCE), 26(6, Ser. 3), 10–20. https://www.iosrjournals.org/iosr jce/papers/Vol26 issue6/Ser 3/C2606031020.pdf

Padman, S. N., & Magare, D. (2023). Multi modal speech emotion detection using optimised deep neural network classifier. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *11*(5), 2020 2038.

Perse, R. J. P. (2024). Speaking Skills: A Panacea for College Students in the Classroom and in the Workplace.

Ruangdit, T., Sungkhin, T., Phenglong, W., & Phaisangittisagul, E. (2023, October). Integration of Facial and Speech Expressions for Multimodal Emotional Recognition. In *TENCON 2023 2023 IEEE Region 10 Conference (TENCON)* (pp. 519 523). IEEE.

Samayamantri, L. S., Singhal, S., Krishnamurthy, O., & Regin, R. (2024). AI driven multimodal approaches to human behavior analysis. In *Advancing Intelligent Networks Through Distributed Optimization* (pp. 485 506). IGI Global.

Sitharam, A., Rashmi, K. R., Prakash, N. E., & Rangareddy, J. (2024, November). Multi Modal Emotion Detection using Deep Learning Techniques. In *2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS)* (pp. 1 6). IEEE.

Sridhar, K. V., & Thripurala, S. (2023, August). Real time facial emotion detection system using multimodal fusion deep learning architecture. In *2023 International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)* (pp. 1 6). IEEE.

Srivastava, K., Sinha, A. R., & Shukla, A. K. (2023). Breaking barriers: Fostering effective technical communication training across corporate and academic sectors. Journal of Survey in Fisheries Sciences, 10(1). https://doi.org/10.53555/sfs.v10i1.1888

Suganya, S., & Charles, E. Y. A. (2019, September). Speech emotion recognition using deep learning on audio recordings. In *2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer)* (Vol. 250, pp. 1 6). IEEE.

Sun, J. (2024). Research and application analysis of multimodal emotion recognition methods based on speech, text, and facial expressions. Highlights in Science, Engineering and Technology, 85, 293–297. https://doi.org/10.54097/agvjvq19

Surachman, L. M., Abdulraheem, A., Al Shuhail, A., & Kaka, S. I. (2024). Acoustic impedance prediction based on extended seismic attributes using multilayer perceptron, random forest, and extra tree regressor algorithms. *Journal of Petroleum Exploration and Production Technology*, *14*(7), 1923 1931.