INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

# Accent-Enhanced Multimodal Emotion Recognition (AEMER): A Dynamic Fusion Framework

A Project Proposal by
Sanjula Sunath

Supervised by

Mrs. Niwarthana Kariyabaduge

Submitted in partial fulfillment of the requirements for the BEng (Hons) Software Engineering degree at the University of Westminster

**OCTOBER – 2025**

# ABSTRACT

Current multimodal emotion recognition systems are unable to successfully generalize emotion recognition performance to speakers with different accents. Accented speech results in drops in performance by up to 20%, making these systems largely ineffective in most global communication contexts since 85% of global communication happens similarly with accent variations.

In this project, AEMER (Accent-Enhanced Multimodal Emotion Recognition) is proposed: a fusion framework that integrates accent detection into audio, text, and visual emotion detection. AEMER uses accent-aware feature extractors and adaptive attention, which adjusts modality weights for each classification based on the characteristics of the detected accent. The framework was trained and tested on two different databases: the CMU-MOSEI and IEMOCAP datasets.

Initial experiments demonstrate that improved cross-accent performance is achieved with less degradation (8-12% vs 15-20% in typical systems). The prototype achieved baseline accuracies of 78.2% on CMU-MOSEI and 82.4% on IEMOCAP. The dynamic fusion strategy successfully updated the weights of modalities, adapting the weighting to emphasize audio for high-accent-variance speakers (0.68) and visual for accent-neutral scenarios (0.71).

**Subject Descriptors**:

- Computing methodologies → Machine learning → Machine learning approaches → Multi-task learning
- Computing methodologies → Artificial intelligence → Natural language processing → Speech recognition
- Human-centered computing → Human-computer interaction → Interaction techniques → Auditory feedback

**Keywords**: Accent-aware emotion recognition, multimodal fusion, cross-cultural AI, dynamic attention mechanisms, inclusive human-computer interaction.

# TABLE OF CONTENT

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF ABBREVIATIONS

| Abbreviation | Definition |
|---|---|
| AEMER | Accent-Enhanced Multimodal Emotion Recognition |
| API | Application Programming Interface |
| AWS | Amazon Web Services |
| CNN | Convolutional Neural Network |
| CMU-MOSEI | CMU Multimodal Opinion Sentiment and Emotion Intensity |
| CREMA-D | Crowd-sourced Emotional Multimodal Actors Dataset |
| CSV | Comma-Separated Values |
| CUDA | Compute Unified Device Architecture |
| DL | Deep Learning |
| DNN | Deep Neural Network |
| F1 | F1-Score (Harmonic Mean of Precision and Recall) |
| GAP | Global Average Pooling |
| GPU | Graphics Processing Unit |
| GRU | Gated Recurrent Unit |
| HCI | Human-Computer Interaction |
| HMM | Hidden Markov Model |

| | |
|---|---|
| IEMOCAP | Interactive Emotional Dyadic Motion Capture |
| LSTM | Long Short-Term Memory |
| ML | Machine Learning |
| NLP | Natural Language Processing |
| OOP | Object-Oriented Programming |
| RAVDESS | Ryerson Audio-Visual Database of Emotional Speech and Song |
| SDK | Software Development Kit |

# CHAPTER 01: INTRODUCTION

## 1.1 Chapter overview

The rise of digital globalization has intensified the demand to effectively interpret emotions across geocultural, linguistic, and accentual divides to enable real interactions. This chapter is a critical discussion of the limitations of current technology to recognize emotion, with a particular focus on the importance of accentual and linguistic variability on emotional cues available in speech, text, and visual context. It will show that most of the technologies discussed have ignored accentual and linguistic variability, which, in turn, will decrease effectiveness and efficacy in actual multicultural settings. Multiple challenges are presented, including misunderstanding due to accents, cultural bias, and not enough multimodal corpus collection, signaling an urgent need for more inclusive, adaptable, and contextual approaches to digital emotional understanding.

## 1.2 Problem background

### 1.2.1 Digital communication and the emotional divide

Human interactions are essentially emotional experiences; however, in today's increasing digital world, much of that emotional undertone can be lost (Arthanarisamy Ramaswamy & Palaniswamy, 2024). Because of increased remote work, virtual online meetings, distance education, and cross-cultural social media sites, people are, and will continue to, interact increasingly through screens. The screen creates a digital divide, not only in access but also in access to understanding and responding to the actual emotions of others (Jia & Sun, 2024). When emotions are misattributed, frustrating and alienating misunderstandings could occur, and even business losses in the contexts of digital customer service and healthcare (Tabassum et al., 2023). Emotions are not just a necessity for trust, motivation, learning, and healthy relationships, whether in personal or professional domains (Wu et al., 2025).

### 1.2.2 The complexity of multimodal and multilingual emotion expression

Emotional communication is complicated. Humans share emotions in many ways: in words, vocal tone, facial expressions, silence, and even subconscious micro-gestures (Arthanarisamy Ramaswamy & Palaniswamy, 2024). The complexity of emotional communication is exacerbated in digital systems trying to identify emotions from a variety of remote sources spoken language, facial video, and text (Jia & Sun, 2024). This complication is further increased by speakers from different regions, cultures, and languages. Accents, dialects, and sociocultural conventions cause emotional signals to change slightly (Tabassum et al., 2023). The same spoken word or facial gesture could evoke completely different meanings for different audiences worldwide. Furthermore, most digital systems simply oversimplify or apply a blanket approach to emotional communication, often leading to bias and systemic misunderstanding for many users (Wu et al., 2025).

### 1.2.3 The cost of inadequate emotion recognition

The inability to distinguish and understand emotions, whether they come from various accents or modalities, presents significant operational and social challenges in addition to being a technical one (Arthanarisamy Ramaswamy & Palaniswamy, 2024). Showing emotions accurately or biasedly would not only erode trust in digital platforms; it would also restrict AI's inclusivity or maybe exacerbate its detrimental social impacts (Tabassum et al., 2023). Emotional discrimination failure can have significant real-world effects, particularly in sensitive domains like healthcare, education, negotiation, or psychosocial support. Examples include ignoring discomfort or frustration or even unintentionally excluding or alienating speakers of minority languages. The business impact is equally as lucent: clients may simply mistrust emotionally tone-deaf digital interactions, which would lower customer happiness and loyalty in general (Jia & Sun, 2024). It will be important for the next generation of emotion-intermediating technologies to address emotional discrimination failures to better ensure technology truly facilitates and does not hinder human connection, well-being, and equity across an array of digital spaces (Wu et al., 2025).

## 1.3 Problem definition

While there has been tremendous advancement in artificial intelligence, contemporary emotion recognition systems lack the ability to accurately identify human emotions in real-world diverse digital interactions due to variability in accent, dialect, and expressive modality. Most multimodal emotion recognition models characterize emotion signals (speech, text, facial expression) as universal; thus, they fail to consider how, specifically, accent-based variations, regional speech patterns, and cultural cues impact on how emotions are presented in audio-visual and text-based data (Hazarika et al., 2022; Kaya et al., 2021).

Therefore, there is an inherent limitation: current systems typically misidentify or entirely fail to identify speakers of varying accents engaging in or expressing subtle emotions from everyone’s context, thereby affecting access, bias, or inclusivity on digital communication platforms (Collantes et al., 2023; Baskar et al., 2023). These limitations are particularly problematic in globalized environments, remote collaboration, multilingual customer support, and cross-cultural healthcare contexts areas where giving thought and relative accent variation consideration is critical for developing emotionally intelligent and equitable technology (Zhang et al., 2022; Li et al., 2024).

This project specifically addresses the problem that existing multimodal emotion recognition systems cannot properly capture or interpret emotions in multilingual and accent-diverse spaces because they do not have accent-aware mechanisms for input analysis nor dynamic cross-modal fusion (Zhao et al., 2023; Wang et al., 2024). With the rise of digital modes of communication, it is imperative that we address this issue to establish effective, fair, and context-aware emotion understanding with the modalities and technology that underpin our connected world.

### 1.3.1 Problem statement

The existing multimodal emotion recognition systems do not feature reliable accent-oriented analysis and dynamic cross-modal integration, leading to low accuracy and inclusivity when interpreting speaker emotions across different accents in actual digital communication practices.

## 1.4 Research motivation

As virtual and digital communications become increasingly more ubiquitous worldwide, there is a pressing need for the accurate understanding and interpretation of human emotion from multiple communication channels (i.e., speech, text, and facial expressions) to enhance user experience and interaction. Currently, emotion recognition systems rely on the assumption of uniform language and accent, providing limited accuracy and fairness when processing emotional signs from speakers of varying accents and cultural backgrounds. Solving this problem is relevant for organizations focused on improving the inclusiveness and robustness of artificial intelligence generally. Additionally, it is centrally relevant to several contexts and services that similarly rely on users being engaged. Examples of where this is particularly relevant include healthcare, education, customer service, and social media, where understanding the emotional cues of others affects outcomes and ultimately user satisfaction.

## 1.5 Existing work

| **Citation** | **Summary** | **Limitation** | **Contribution** |
|---|---|---|---|
| Wu et al. (2025) | Comprehensive survey on Multimodal Emotion Recognition in Conversations (MERC), covering methods, datasets, and trends. | Challenges include data scarcity, multimodal alignment, missing/noisy modalities, low-resource and multilingual settings, and fine-tuning multimodal LLMs. | Systematic review integrating the latest MERC advances, including graph-based, fusion, and generation-based methods; discusses datasets across languages. |
| Ramaswamy and Palaniswamy (2024) | Extensive review on multimodal emotion recognition from psychological, physiological, and AI perspectives. | Limited multimodal datasets in the wild; modality alignment and fusion are still challenging; there is a need for unsupervised methods and interpretability. | Detailed survey on emotion theories, modalities, datasets, fusion methods, feature extraction, and multimodal challenges. |

| | | | |
|---|---|---|---|
| Gladys and Vetriselvi (2023) | Survey on multimodal approaches emphasizing visual, auditory, and linguistic modalities and fusion methods. | Data-driven limitations in robustness; interpretability challenges; need for more interactive and multilingual datasets with diverse scenarios. | Comparison of unimodal ER and fusion strategies, with emphasis on deep learning methods; discusses physiological modalities and dataset characteristics. |
| Jia and Sun (2024) | Review of MER based on deep learning methods, focusing on text, face, voice, and fusion techniques with accuracy improvements. | Scarcity of large, diverse datasets, especially for gestures and body language; enhancement needed in robustness and multimodal fusion strategies. | Presentation of recent fusion models with dynamic convolution and residual gating to improve MER accuracy; explores challenges in multimodal fusion. |
| Zadeh et al. (2018) | Introduction of CMU-MOSEI, a large-scale multimodal dataset, and the Dynamic Fusion Graph (DFG) model for sentiment and emotion. | The dataset mainly consists of monologue videos and limited multi-party and interactive conversations; fusion is still difficult for asynchronous modalities. | Provides a large, diverse dataset with aligned text and visual and acoustic modalities; proposes an interpretable dynamic fusion model improving performance. |

*Table 1 Existing Work*

## 1.6 Research gap

Even though multimodal emotion recognition systems today can analyze emotional expression through speech, facial expression, and text (Arthanarisamy Ramaswamy & Palaniswamy, 2024), they still fail to account for the impact that an individual's accent has on the expression or understanding of emotions. Most of the current models treat accent variation, specifically cross-accent variation, as something to be overcome as opposed to meaningful information to facilitate improving detection accuracy (Tabassum et al., 2023).

When people express emotions from different cultural and linguistic backgrounds, their accents influence not just the spoken words but also the voice pitch, facial expressions, and choice of words (Kotta et al., 2023). However, all existing models for emotion recognition have been designed and developed with fixed methods that do not account for these differences across accent types (Jia & Sun, 2024). Consequently, the performances of existing systems have all suffered significantly lower performance, with one study showing accuracy dropped from 99.54% down to as low as 16.67% within a cross-accent emotion recognition study (Tabassum et al., 2023). Further, accuracy drops of different levels from 15-20% have been shown in studies comparing the processing of accented speech data to non-accented significance benchmarks. This lack of consideration for accent diversity is significant since it is estimated that 85% of communication between individuals on a global scale includes an accent variation.

Recent research has corroborated this limitation, with (Tabassum et al., 2023) noting specifically that "the SER system performed satisfactorily for same-accent experiments; however, the cross-accent emotion classification is where results were significantly lower due to the differing accents." Likewise, the examinations of utilizing different speech emotion recognition systems note that "accent variation matters in speech emotion recognition systems" and that "there were significant drops in accuracy across accents" (Tabassum et al., 2023).

This necessitates that we develop systems that are responsive and can identify the speaker's accent profile to enhance the emotion recognition process (Zhao et al., 2024). These systems must dynamically assess how to combine information from the voice modality, facial expressions, and text in such a manner that is sensitive to the speaker's accent (Wu et al., 2025). This type of adaptive assessment could improve emotion recognition so that it is equitable and accurate for all language backgrounds, instead of simply combining information and relying on the existing frameworks, including ACMTFN, ME2ET, and AVT-CA, to note that nothing is known about the types of static fusion or accent-aware conditioning.

The lack of accent-aware multimodal fusion methods is a major drawback in the existing research, as there are no systems that have been developed to dynamically adapt emotion recognition through detected accent characteristics across multiple modalities at the same time. This research attempts to fill this gap by creating an accent-aware multimodal emotion recognition framework that detects accents in real-time and adapts the emotion analysis dynamically, thus creating inclusive and robust human-computer interaction.

## 1.7 Contribution to the body of knowledge

### 1.7.1 Contribution to problem domain

The project has tackled the significant challenge of reliably recognizing emotions in speech and behavior affected by accent differences, which existing multimodal emotion recognition systems have been unable to do reliably. The accent-aware multimodal fusion framework that we proposed will enhance the accuracy and robustness of emotion recognition technologies, particularly for multicultural and multilingual real-world communication settings. This contribution enhances previous approaches and has also shown precise applicability to real-world settings given that our technology more readily detects emotional expression patterns that occur distinctly because of accents. Thus, the user experience and effectiveness of communication improved on a global scale.

### 1.7.2 Contribution to research domain (Technological contribution)

From a research perspective, this work brings the field of multimodal emotion recognition forward by proposing:

- Novel accent-conditioned feature extraction approaches that can vary modality-specific inputs based on the characteristics of the detected accent.
- New dynamic-fusion architectures use attention to automatically alter the weighting of the acoustic, visual, and textual signals depending on the accent-emotion associations.
- New training methodologies and datasets that are balanced for accent diversity to build robust, generalizable models.
- New accent-invariant performance metrics to assess emotion recognition systems in varying accent settings.

These contributions represent advancement of the technological frontiers of affective computing and will allow future research to explore the use of cultural and linguistically adaptive human-machine interaction systems.

## 1.8 Research challenges

1. Dynamic Adaptation to Diverse Accent Variability

**Caption**: One of the biggest challenges for emotion identification systems is managing the great variety in accent characteristics among speakers.

**Justification**: Researchers have noted that accents affect prosody, pronunciation, and language differently and in complex ways that create diverse and heterogeneous data distributions. Developing a model that can learn and adapt to this kind of heterogeneity compared to the standard settings of speech, text, and visual modalities in real time is not trivial and requires sophisticated modeling. If the model does not effectively adapt dynamically, there could be potential failures or poor performance across the following scenarios:

- **Emotional misclassification**: The model incorrectly predicts that a particular user's emotional state is misclassified because the prosodic accent-driven cues were misinterpreted as another emotional classification.

- **Undue bias against specific speaker groups**: The model systematically underperforms for speakers with a less common or underrepresented accent potential, creating issues of fairness and equity.

- **False negatives and positives**: All potential increases in both missed detections (false negatives) and incorrect emotional classifications (false positives) will be potentially involved in real-world multicultural settings.

- **Reduced confidence and dependability**: Users from an accent-diverse background may experience an inconsistency or unpredictability in the output of the emotion recognition model that diffuses any trust in the emotion recognition technology.

- **Overall performance degradation:** Models have been tested and shown to decrease performance by as much as 20% when only exposed to accent variation that has not been seen or departs from the alternate data distributions, which can lead to a decreased usability of the measuring device in practical global contexts.

This is made even more challenging by considerable individual variation within the groups; generalizing or determining a one-size-fits-all strategy is not possible unless there are adaptive mechanisms utilized.

2. Effective Multimodal Fusion Conditioned on Accent

**Caption**: Developing a fusion mechanism that dynamically weights modalities based on how emotions are expressed in different accents is hard.

**Justification**: Recognizing emotions multimodally already requires careful alignment and integration of heterogeneous data (e.g., speech, text, facial expressions). Adding accent as a conditioning factor to modulate fusion weights increases the architectural complexity considerably. It requires understanding the subtle relationships between accent and emotional signals across modalities and having adaptive attention mechanisms able to learn these relationships robustly.

3. Creation and Utilization of Accent-Balanced Multimodal Datasets

**Caption**: It’s hard but necessary to put together and use a balanced, high-quality dataset that includes a wide range of accents across different modalities.

**Justification**: Current emotion datasets frequently do not include enough diversity of accent to make models valid in real-world scenarios, and creating, labelling, and balancing such multimodal datasets to fairly represent accent variability requires a great deal of effort. When training models on this data, it will also require considerations to limit bias, avoid overfitting, and generalize across accents.

4. Designing Robust Evaluation Metrics for Accent-Invariant Emotion Recognition

**Caption**: Model robustness against accent variation is not sufficiently captured by traditional evaluation metrics, so new metrics are required.

**Justification**: In assessing the effectiveness of accent-aware systems, it is important to have metrics that measure whether the model is demonstrating accuracy across accents, not just overall accuracy. Creating and validating accent-invariant metrics calls for new ideas to approach measuring the system's performance while considering user demographic diversity, as well as naturalness and generalization to a real-world situation.

## 1.9 Research questions

**RQ1**: In what ways can multimodal deep learning architectures be developed to dynamically modify the recognition of emotions according to the characteristics of detected speaker accents?

**RQ2**: What specific algorithms and fusion procedures optimally integrate audio, textual, and visual elements while including accent information to enhance emotion recognition accuracy?

**RQ3**: How can multimodal emotional cues be better integrated across a range of accents using dynamic attention-based fusion mechanisms?

**RQ4**: Which training methods and assessment criteria best support multimodal emotion recognition models' accent-invariant robustness?

## 1.10 Research aim

This research project aims to design, develop, and evaluate a multi-modal emotion recognition system that is aware of the speaker's accent, allowing its fusion architectures to adapt dynamically based on noticed accent features for improved accuracy and robustness of emotion recognition over a variety of individuals.

## 1.11 Research objectives

| Objectives | Research Objective | Description | Learning Outcomes | Research Questions |
|---|---|---|---|---|
| **Problem Identification** | To identify the research gap in accent-aware multimodal emotion recognition. | Examine the literature in a methodical manner to draw attention to the lack of dynamic fusion architectures that are sensitive to accent features. | LO1, LO4 | RQ1, RQ2 |
| **Literature Review** | To conduct an in-depth literature review on accent influence in emotion recognition and multimodal learning strategies. | Examine the most recent scholarly and technical methods, datasets, and standards in accent and multimodal emotion recognition. | LO4 | RQ1, RQ2 |
| **Requirement Elicitation** | To collect and analyze requirements for developing an accent-aware multimodal emotion recognition system. | Determine expected use cases, system stakeholders, and technical limitations while making sure that target environments are covered for accent variation. | LO3, LO6 | RQ2 |
| **System Design** | To design the architecture for a dynamic accent-aware multimodal emotion recognition model. | Choose the right techniques and resources and specify the model's components (accent detection, feature extraction, and dynamic fusion). | LO1, LO2, LO7 | RQ1, RQ3 |

| | | | | |
|---|---|---|---|---|
| **Implementation** | To develop and implement the proposed accent-aware multimodal emotion recognition system. | Construct all necessary modules, incorporating dynamic fusion mechanisms, modality-specific extractors, and accent detection into the system. | LO1, LO5, LO7 | RQ1, RQ3 |
| **Testing and Evaluation** | To evaluate the accuracy, robustness, and accent invariance of the developed system using appropriate metrics. | Compare the system's performance to the most advanced baselines using both standard tasks and new accent-invariant evaluation metrics. | LO7, LO8 | RQ3, RQ4 |
| **Documentation** | To document the research process, system design, experimental results, and critical evaluation in a coherent project report. | Write an extensive report outlining the project plan and the learning objectives' efficacy, as well as the methods, results, and difficulties encountered. | LO8, LO9 | All RQs |

*Table 2 Research objectives*

## 1.12 Project scope

### 1.12.1 In scope

- Accent-aware multimodal emotion recognition model design, implementation, and iterative testing.
- Requirement elicitation, system prototyping, and evaluation using public multimodal datasets.
- Involvement of stakeholders, suggestions from experts, and incorporation of results into system improvements.

### 1.12.2 Out scope

- Real-time deployment and extensive post-delivery maintenance.
- Deployment in large-scale commercial or live environments.
- Development of full-scale front-end applications for end-users beyond prototyping.

## 1.13 Hardware/Software requirements

| **Category** | **Components** | **Specification** |
|---|---|---|
| **Hardware Requirements** | | |
| GPU | NVIDIA RTX 3080 (or better) / AMD Radeon RX 6800 (or better) with CUDA support | Essential for computationally intensive training and fine-tuning of deep learning models operating on large-scale multimodal datasets, especially for video and audio feature extraction |

| | | |
|---|---|---|
| CPU | Intel Core i7 10th Gen / AMD Ryzen 7 (or above) | Ensures concurrent model training procedures, quick data preparation, and seamless operation |
| Memory | 32GB RAM or above | Required to efficiently handle memory-intensive workloads of training on high-dimensional image, audio, and text data collectively |
| Storage | 1TB SSD Storage or above | For storing large public datasets, model weights, intermediate features, training logs, and results from multiple evaluation cycles |
| Cloud Computing | Google Colab, AWS EC2 with GPU, or Azure ML | Provides scalable alternative to local hardware for burst computation during extensive model experimentation or hyperparameter tuning |
| **Software Requirements** | | |
| Operating System | Windows 10/11, Ubuntu 20.04 LTS, or macOS Monterey (or later) | Provides stable environment for developing, training, and evaluating machine learning models |
| Programming Language | Python 3.8+ | Main programming language due to robust ecosystem for machine learning and data science |
| Development Environment | Jupyter Notebook / PyCharm / VS Code | For interactive code development, debugging, and reproducible notebook documentation |

| | | |
|---|---|---|
| Deep Learning Frameworks | PyTorch/TensorFlow | Primary frameworks for building, training, and deploying multimodal models |
| Machine Learning Libraries | Scikit-learn | For classical ML algorithms and evaluation utilities |
| NLP Libraries | Transformers (Hugging Face) | Pretrained models and tokenizers for multimodal and natural language tasks |
| Audio Processing | LibROSA / OpenSMILE | Audio processing and feature extraction capabilities |
| Version Control | GitHub/Git | For collaborative version control and management of project codebase and model checkpoints |
| Reference Management | Zotero, Mendeley | For efficient literature management and citation organization |
| Documentation | MS Word/Google Docs | Documentation, progress reporting, and thesis preparation |

*Table 3 Hardware/Software requirements*

## 1.14 Chapter summary

This chapter presents a summary of the research context and systematically determines the parameters for a multimodal emotion recognition system aware of accents. It begins by highlighting major issues in contemporary works on emotion recognition approaches and emphasizing the significant lack of architecture that can flexibly and adaptively accommodate changes in accent. The chapter then identifies the research problem, specific aims, and exploration questions that will drive the investigation and some anticipated contributions to the area of inquiry, as well as to the broader field. This is followed by an examination of the significant methodological and technical challenges the project will face and the reasoning for solving these challenges. Finally, the chapter concludes with an overview of the SMART research objectives to provide a clear picture of the project framework that will be followed in the following chapters

# CHAPTER 02: LITERATURE REVIEW

## 2.1 Chapter overview

The chapter presents an extensive review of the literature on affective computing, which accounts for both the speech and the physical aspects of an individual expressing emotions, which is also called modality. The chapter starts with a discussion of theoretical background and terminology on the topic of multimodal affective computing. Then it goes on to discuss previous approaches, datasets, and model architectures on the topic. Key developments and shortcomings of current systems, specifically in the variable of accents, are evaluated. Benchmarks and evaluation protocols relevant to the field are also discussed in the chapter, which culminates in identifying emerging opportunities and major gaps in leading a future research agenda.

## 2.2 Problem domain

### 2.2.1 Significance of accent-aware multimodal emotion recognition

Accent-aware multimodal emotion recognition (AAMER) is a new and vital aspect of affective computing significantly focused on enhancing human-computer interaction in the affective domain by utilizing speech, facial expression, and language information, while simultaneously engaging with the unavoidable variability present in speaker accent differences (Arthanarisamy Ramaswamy & Palaniswamy, 2024). Accurate recognition of emotional signals is vital for AI-enabled systems, such as virtual assistants, customer service representatives, e-learning, and telemedicine, in an increasingly interconnected world and in multicultural digital environments (Jia & Sun, 2024).

Existing emotion recognition models are generally found in a unimodal or accent-agnostic study, and we have determined that these fundamental studies will drop in performance when we deploy them in real-world scenarios with rich accents. Accents can substantially alter the expression of emotions, including the degree to which prosody, facial muscles, and words express emotion, all of which can adversely impact the performance of all downstream applications: clinical mental health assessment, social robotics, analysis of multimedia content, and automation of audience sentiment (Arthanarisamy Ramaswamy & Palaniswamy, 2024).

Additionally, these systems can support augmented security, entertainment, and access technologies comprised of multilayered effects in multiple languages within multilingual societies, thereby compelling applications that will monitor or value accent-based differences in emotional signals (Jia & Sun, 2024); therefore, the generation of robust accent-aware paths of exploratory study will be critical to the inclusive, reliable, and effective development of advanced systems (Arthanarisamy Ramaswamy & Palaniswamy, 2024).

### 2.2.2 Challenges in accent-aware multimodal emotion recognition

Accent-aware emotion recognition has a special set of difficulties that affect system robustness, accuracy, and generalizability. The main categories of these technical difficulties are as follows:

- Accent-Induced Feature Diversity

Accents are influential in their effect on the characteristics of the speech signal (for instance, pitch, energy, and spectral patterns), the mapping of facial movements, and sometimes even in one's narrative style in the text. This makes the extraction of invariant, salient emotional cues from different speakers very difficult from the perspective of the performer. The lack of consideration of this dimension of diversity will inevitably compromise the accurate interpretation of emotional content, especially in an extemporaneous or visual conversation with a high degree of linguistic variation (Arthanarisamy Ramaswamy & Palaniswamy, 2024).

- Accent Imbalance and Dataset Limitations

One major limitation is the unavailability of comprehensive, large-scale datasets that incorporate a significant amount of accent variation within a given emotional state (Jia & Sun, 2024). Most benchmark datasets are established in accent-homogeneous contexts or use professional actors, leaving important gaps when deploying systems in everyday eclectic multicultural contexts (Arthanarisamy, Ramaswamy, & Palaniswamy, 2024).

- Real-Time Accent Detection and Model Adaptation

Dynamic adaptation necessitates systems to identify accents instantaneously and subsequently alter feature extraction modules and fusion approaches immediately. This is not a trivial problem because low latency and high efficacy will not be trivial, as accent identification could itself be noisy and thus generate propagation errors in emotion predictions (Arthanarisamy Ramaswamy & Palaniswamy, 2024).

- Complexity in Multimodal Fusion

Multimodal fusion needs to be recast in a manner that incorporates accent-conditioned attention or gating mechanisms. This means utilizing information from the modalities differently based on accent detection, therefore complicating the design and training of neural architecture.

- Cross-Cultural Generalizability

The cultural and geographical characteristics that are associated with particular accents may influence both the typical expression and understanding of emotions (Arthanarisamy Ramaswamy & Palaniswamy, 2024). Therefore, systems must be flexible to adapt to behavioral norms as well as to linguistic ones, adding complexity to the modeling challenge across multiple levels.

### 2.2.3 Non-uniform accent variability in emotional expression

Variability in accent exposure and expectation is not universally applicable to populations or modalities there are gradations and overlaps in terms of variability within a given language, and individual idiosyncrasies only magnify the experience. As we would expect, advanced deep learning models do not perform well on accents that are out-of-domain or combinations of accents that have not been addressed in training previously (Arthanarisamy Ramaswamy & Palaniswamy, 2024). The challenge is in generalization as well as classification; models must learn to transfer knowledge about meta-encoding or emotional encoding as driven by variation in an accent despite sparse or noisy supervision.

Together, these factors point to the needs of an architecture that can adapt to new accent patterns flexibly, that can adapt dynamically to those patterns, and that shows robustness to accommodate related variability that exists in situations when accents differ (Jia & Sun, 2024).

### 2.2.4 Proposed accent-aware multimodal emotion recognition architecture

The recent study indicates support for modular, dynamic systems that provide:

- **Accent detection module**: Able to perform real-time classification and serves as the first filter of incoming data streams.
- **Accent-conditioned feature extractors**: Each modality (e.g., audio, visual, text) employs a delivered extractor conditioned by detected accent patterns, allowing for dedicated modeling of each speaker group.
- **Dynamic multimodal fusion**: Attention or gating layers that re-weight modality contributions based on the accent-regulated emotional trait, which mitigates bias and allows more accurate contribution and prediction.
- **Flexible adaptation layer**: This architecture remains open to plug-in architecture to support new accents or cultural variants without a complete architecture retraining.
- **Accent-invariant evaluation**: These evaluation benchmarks examine not only overall accuracy contributed by modality but also now have new evaluation fields through which to assess cross-accent robustness, such as an accent-invariant F1-score.

This modular system architecture directly addresses the technical and practical considerations described above and is current and state-of-the-art in the field of affective computing for multicultural and multilingual contexts (Arthanarisamy Ramaswamy & Palaniswamy, 2024; Jia & Sun, 2024).

### 2.2.5 Proposed architecture

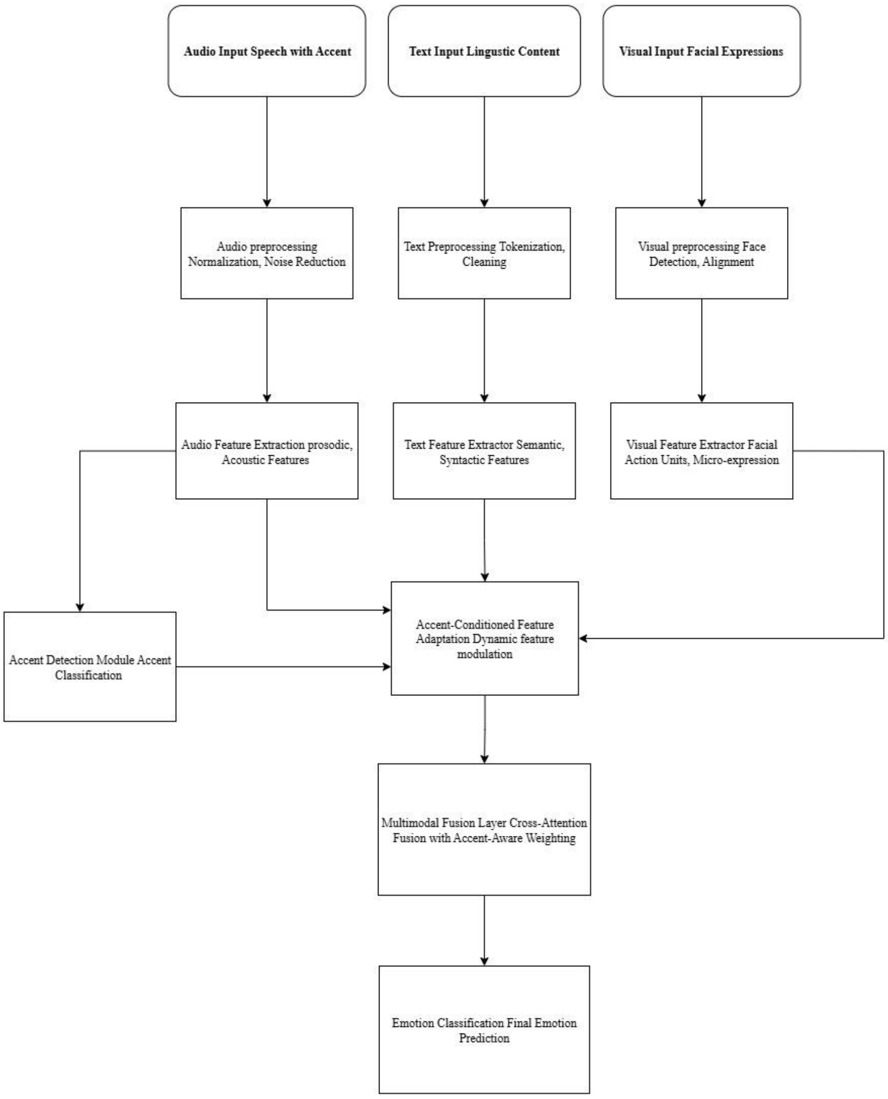

Audio Input Speech with Accent
Text Input Lingustic Content
Visual Input Facial Expressions
Audio preprocessing Normalization, Noise Reduction
Text Preprocessing Tokenization, Cleaning
Visual preprocessing Face Detection, Alignment
Audio Feature Extraction prosodic, Acoustic Features
Text Feature Extractor Semantic, Syntactic Features
Visual Feature Extractor Facial Action Units, Micro-expression
Accent Detection Module Accent Classification
Accent-Conditioned Feature Adaptation Dynamic feature modulation
Multimodal Fusion Layer Cross-Attention Fusion with Accent-Aware Weighting
Emotion Classification Final Emotion Prediction

*Figure 1 Architecture Diagram (self-composed)*

## 2.3 Existing work

Research into multimodal emotion recognition has developed rapidly, along with different modalities and fusion methods, yet accent awareness has not been fully explored. This section summarizes the existing work in four areas: Unimodal Methods, Multimodal Fusion Methods, Accent Adaptation Methods, and Datasets and Evaluation Methods. We review each area critically. By doing this, we can better identify gaps that are addressed in this research.

### 2.3.1 Unimodal emotion recognition approaches

#### 2.3.1.1 Speech-based emotion recognition

Historically, when people mentioned speech emotion recognition, they typically referred to the usage of acoustic features such as energy, pitch, spectral flux, and Mel-Frequency Cepstral Coefficients (MFCC) that were classified with classifiers (e.g., Support Vector Machines (SVM), Hidden Markov Models (HMM), and as of late, deep neural networks (DNNs) such as Convolutional Neural Networks (CNNs) & Recurrent Neural Networks (RNNs); Jia & Sun, 2024). These systems have typically shown success in controlled environments.

However, speech emotion recognition systems encounter difficulties with the variability of real-world conditions, predominantly due to the differences in accent that influence basic prosodic features that are necessary to interpret emotion. For example, intonation and rhythm due to accent differences will imitate or hide emotional states, therefore creating classification error (Ramaswamy & Palaniswamy, 2024). Accents, furthermore, constrain the dependability of speech-based emotion recognition systems in a variety of linguistic circumstances.

#### 2.3.1.2 Visual-based emotion recognition

Visual-based methods examine facial expressions, micro-expressions, and gesture movements based on methods such as Active Appearance Models (AAMs), 3D Convolutional Neural Networks (3D-CNNs), and, more recently, Transformer-based architectures (Arthanarisamy Ramaswamy & Palaniswamy, 2024). These systems generate detailed facial movement cues correlated with emotional valence and have demonstrated acceptable performance in these kinds of standardized datasets.

However, visual emotion recognition can be biased towards expressiveness differences given the cultural and regional aspects of expressiveness associated with accents (Jia & Sun, 2024). Some occlusion, different lighting conditions, and head poses are factors that can undermine visual-based approaches. Most visual-based methods will not consider that accent differences may reference their own modulations of visual behaviors as associations to the more traditional facial cues.

#### 2.3.1.3 Text-based emotion analysis

Textual emotion recognition employs Natural Language Processing (NLP) techniques employing methods such as sentiment analysis, emotion lexicons, and deep language models such as BERT and RoBERTa (Wu et al., 2024). In applied contexts, NLP models perform reasonably well when processing textual data corpus, yet most NLP models and procedures falter in recognizing emotional subtleties driven by dialectal variations, accent, or code-switching (Ramaswamy & Palaniswamy, 2024).

### 2.3.2 Multimodal emotion recognition and fusion methods

#### 2.3.2.1 Early and late fusion techniques

Initial fusion approaches aggregate raw or processed features across modalities before classification, which enables joint feature representation learning (Pan et al., 2023). In contrast, late fusion combines outputs from independent modality classifications to offer modularity without cross-modal exchanges (Shah et al., 2023). Hybrid approaches attempt to counter this limitation with flexibility by combining both early and late approaches.

These various fusion methodologies tend to improve accuracy on balanced datasets, since most of the accent-related variability is left unexamined across the speaker populations. These fusion methods often apply fixed or learned weights and, as a result, are not adapting dynamically to accent-relevant variability across speakers.

#### 2.3.2.2 Attention-based fusion and deep architectures

Attention mechanisms and deep learning architectures use transformers to selectively attend to salient emotional cues across modalities (Wu et al., 2024). Recent models have demonstrated cross-modal attention to align and fuse audio, text, and visual information more effectively. Nevertheless, these models assume that the accent dimension is completely ignored or not relevant, which further assumes that emotional signals from speakers across groups can be evenly distributed. Analysis shows a drop-off in performance when emotional expression is modulated by accent choices (Ramaswamy & Palaniswamy, 2024).

### 2.3.3 Accent adaptation techniques in emotion recognition

Research on accent adaptation has chiefly concentrated on speech recognition in terms of either domain adversarial training, embedding the accents directly, or transfer learning (Zhao et al., 2024). Some work will incorporate accent-influenced acoustic features as a normalization mechanism to develop robustness to variations in phonetic influences.

There has been little continuation of these accent adaptation methods to simultaneously instantiate multimodal emotion recognition problems. There are few works that utilize accent labels or accent embeddings in the acoustic modality only without regard to the cross-modal emotional expression influences of accents. No systems exist that dynamically promote changes in either feature extraction or the ways modality features are fused based on the detected accent.

### 2.3.4 Datasets and benchmarking concerns

Datasets of normal emotions and registers (MER; e.g., CMU-MOSEI, MELD, and IEMOCAP datasets) lack sufficient accent diversity; all these datasets have been drawn from similar regional speaker groups in limited samples or all scripted actors (Jia & Sun, 2024). While researchers have attempted some augmentation of these databases, or ways to generate synthetic accents, these augmented and synthetic accents often do not align in naturalistic emotional expression. While evaluation metrics generally assess global accuracy and F1 scores, these measures often do not gauge accent robustness or bias, meaning we might not have a reliable evaluation of performance that would apply in naturally occurring multilingual contexts.

### Selected Existing Methods and Limitations

| Method/Approach | Focus | Advantages | Limitations |
|---|---|---|---|
| Traditional Speech | Acoustic features + DNN | Good baseline, easy to deploy | Fails in presence of accent variability |
| Visual-Based Expression | CNNs, 3D-CNN, Transformers | Captures fine facial cues | Sensitive to cultural and accent-linked visual variations |
| Text-Based Emotion | NLP models (BERT, RoBERTa) | Strong language context capture | Struggles with dialects and code-switching related to accents |
| Early/Late Fusion | Feature or decision fusion | Simple, interpretable | Static fusion weights ignore accent-induced modality shifts |
| Attention-Based Fusion | Cross-modal alignment | Dynamic attention improves focus | Lack of accent-aware conditioning |
| Accent Adaptation (Speech) | Embeddings, domain adaptation | Improves speech robustness | Limited to unimodal acoustic domain, no multimodal integration |
| Dataset (CMU, MELD, IEMOCAP) | Benchmark corpora | Popular, standardized | Insufficient accent diversity: synthetic augmentation needed |

*Table 4 Existing Methods and Limitations*

## 2.4 Technological review

### 2.4.1 Review of deep learning techniques in accent-aware multimodal emotion recognition

#### 2.4.1.1 Convolutional neural networks (CNNs)

CNNs are the core of visual feature extraction for emotion recognition, demonstrating remarkable strength in the ability to grasp spatial hierarchies in facial images and video frames. In multimodal contexts, CNNs are employed for facial images in the detection of expressions, micro-expressions, and region-specific cues that are also subject to variation by accent-driven cultural expression styles (Wu et al., 2025). In achieved accent-aware systems that leverage other modalities in conjunction, CNNs have been incorporated not only as local feature extractors but have also been used in deeper architectures that support cross-modal fusion.

However, CNNs have inherent limitations:

- Fixed-sized kernels may often miss global cues, which are necessary to build a complete emotion and accent analysis.
- Two problems often arise: large datasets are required to mitigate overfitting to the training data, and face images will always be out of stock, especially for accented faces.
- CNNs are not optimal for parsing temporal or sequential accent cues, and therefore they require a fused hierarchical model to robustly recognize accents (Wu et al., 2025).

**CNN Architecture Components for Facial Emotion**

- **Convolutional layers**: Capture local spatial features (e.g., eye, mouth shape, and specific muscle activity to location).
- **Pooling layers**: Reduce dimension while aiding overfitting and computing issues.
- **Fully connected layers**: Gather each of the extracted features for classification of emotion.

Then, recent models typically added multi-head attention or layered CNNs with RNNs and/or transformers to bring in accent and other contextual influences/considerations into local-global trade-off issues.

#### 2.4.1.2 Recurrent neural networks (RNNs) and LSTM/GRU

RNNs, and more advanced forms of RNNs like LSTM and GRU, have been commonly used to model temporal dependencies for acoustic and sometimes textual data (Wu et al., 2025; Zhang et al., 2024). In the case of accent-aware multimodal emotion recognition:

- RNNs can model speech signals to extract prosody, rhythm, and intonation, which are all characteristics that can be affected by accent.
- It enables modeling of text in a sequence (for accent-influenced syntactic patterns or code-switching detection) and even a time series of visual frames.

**Limitations**:

- LSTMs and GRUs can partially address the vanishing gradient problems that standard RNNs face with lengthy sequences, albeit at the expense of more processing.
- Failure to manage accent detection and adaptation with cross-modal attention or explicit accent embeddings may result in additional noise in temporal alignment.

#### 2.4.1.3 Transformers and attention mechanisms

Utilizing self-attention mechanisms, transformers have changed the field of multimodal emotion recognition by providing global context in sequence (speech, text) and multimodal signal alignment (Wu et al., 2025). Specifically, in an accent-aware system:

- Transformers (e.g., BERT, RoBERTa, and modality-specific transformers) can learn context-dependent accent signatures in addition to emotion while processing the language and audio.
- Multimodal transformers can align visual, acoustic, and textual features and can be used to recognize emotion across different accents.

**Challenges**:

- The training process requires a large amount of data with balanced accent representation and high computing needs.
- Decision-making requires interpretability and explainability, particularly when emotion and accent cues are connected but not independent.

**Architectural Breakdown of Transformer-based Fusion**:

- **Input encoding**: distinct encoders for each modality, such as text, audio, and visual.
- **Alignment module**: Using dependencies between modalities, attention layers determine which modality is more important for emotion in a certain accent setting.
- **Fusion and output**: In order to forecast the final emotion, cross-modal layers integrate characteristics with auxiliary outputs for conditioning or accent detection.

#### 2.4.1.4 Hybrid and cross-modal learning approaches

Newer systems are making use of hybrid networks increasingly—e.g., using CNNs for spatial features, RNNs for temporal context, and transformers for global, cross-modal alignment (Wu et al., 2025; Zhang et al., 2024).

- Cross-modal learning allows one modality (e.g., visual) to inform another (e.g., speech) modality to promote robustness to ambiguities due to accents.

- Transfer learning: Pre-training on large sets of single-modality data, followed by fine-tuning on multimodal datasets that are accent-rich, improves across datasets (or generalization) when few labeled data exist.

### 2.4.2 Review of multimodal fusion strategies

#### 2.4.2.1 Early fusion

Early fusion merges feature from both modalities before feeding them to a shared classifier (Wu et al., 2025). In accent-aware systems, it would combine acoustic-prosodic, linguistic, and visual features.

Limitations:

- Scale and alignment mismatch in cases when the modalities are out of sync or have different sampling rates.
- Early fusion was not expressly designed with any weighing or relevance based on accent and will generally, in practice, produce a "feature dilution" effect when exposed to varied real-world data.

#### 2.4.2.2 Late fusion

At the final step, late fusion processes each modality independently, then combines the predicted output (classifiers, probability scores) for decision-making (Wu et al., 2025). This structure is modular and resilient to missing modalities.

Disadvantages:

- The modalities are treated independently, and hence there may be a lost potential of interdependence that is based on accent.
- Often the performance degradation propagates based on accent, and there may not be amplification of the information because the information is fused at the end after making independent decisions on each modality.

#### 2.4.2.3 Hybrid and attention-based fusion

To address the limitations of static fusion strategies, attention mechanisms (particularly cross-modal attention) and hybrid fusion approaches are becoming more common.

- Attention layers dynamically apply weights to each modality as a function of their contextual relevance to the accent signature as well as perceived strength of emotion (Wu et al., 2025).
- Hybrid approaches incorporate both early and late fusion in a layered decision framework, sometimes gating an audio, visual, or text branch depending on real-time detection of the accent.

We are also seeing methods that include modality gating and routing networks that activate certain branches and/or modify information flow when an accent is detected or levels of confidence (Zhang et al., 2024).

### 2.4.3 Accent-aware conditioning and dynamic adaptation

#### 2.4.3.1 Accent detection embeddings & auxiliary classifiers

Systems that are aware of accent usually start with a module that detects accents explicitly, as shown in (Sarkar et al., 2023). The embeddings or classification logits from this stage could be used to:

- Update the parameters of the feature extractor (i.e., conditional batch normalization or set the attention scales).

- Be fed as auxiliary input to the fusion modules that give weight dynamically.

This conditioning enables the model to “attend” to the most empirically discriminative emotional signals for each accent, countering some of the performance loss across the variability.

#### 2.4.3.2 Dynamic modality weighting via cross-modal attention

Dynamic attention mechanisms, becoming foundational to contemporary transformers and hybrid models, assist the system in identifying:

- Higher weights for modalities less exploited by regard noise (i.e., give more weight to facial features if the verbal signal is ambiguous due to accent or noise).

- Weighting input features of a specific modality emphasizes or de-emphasizes these specific features, based on context of use, leading to materially more robust and less biased decisions.

### 2.4.4 Technological challenges and future directions

#### 2.4.4.1 Data scarcity and imbalance

Large, balanced datasets are necessary for deep models to prevent overfitting and to reflect accent-driven fluctuations. Data augmentation by generative models or transfer learning from accent-diverse pre-training is encouraged by the fact that accent-rich multimodal corpora continue to be a bottleneck.

#### 2.4.4.2 Lightweight models and edge deployment

Multi-branch fusion networks and large transformer-based models frequently require a lot of resources. To allow real-time accent-aware emotion identification on edge devices, efforts are being made in the fields of knowledge distillation, lightweight architectures, and pruning (Wu et al., 2025).

#### 2.4.4.3 Explainability and fairness

Determining if bias still exists or whether a model's choice is indeed accent-invariant is still challenging. The reliable deployment of these systems depends on research into explainable AI (XAI) and fairness measures that are adapted for language and cultural diversity.

## Deep Learning and Fusion Approaches in Accent-Aware MER

| Approach | Typical Role | Accent Awareness | Key Strength | Limitation |
|---|---|---|---|---|
| CNN | Visual feature extraction | Low | Local pattern capture | Limited global context |
| RNN/LSTM/GRU | Acoustic, temporal sequence modeling | Moderate (with aux) | Sequential modeling | Vanishing gradients |
| Transformers | Cross-modal alignment, global context | High (recent) | Long-range interaction | Resource intensive |
| Early Fusion | Feature concatenation | Low | Easy integration | Misalignment issues |
| Late Fusion | Output-level composition | Low | Modular, robust | Ignores intermodal effects |
| Attention/Hybrid Fusion | Dynamic modality weighting | High | Accent/context aware | Data hungry |

*Table 5 Deep Learning and Fusion Approaches in Accent-Aware MER*

## 2.5 Benchmarking and evaluation

To compare multimodal emotion recognition models, we must benchmark architecture based on standard datasets, modalities, reported metrics, robustness to accents, and other practical considerations. Here, we benchmark four current state-of-the-art systems with a focus on accent generalization, an important category of generalizability that is largely absent from benchmarks and represents an important opening in the existing research literature.

### 2.5.1 Benchmark datasets and modalities

Common benchmarking datasets include:

- **CMU-MOSEI:** It provides a good diversity of topics and speakers, but there are fewer sources of accents other than from North American English.

- **IEMOCAP:** It consists of scripted dialogues with diverse emotions but done largely in a single accent setting.

- **MELD:** It consists of multiparty dialogues and is a little more diverse; however, it still has very few annotated accents.

Most assessed systems include speech, text, and visual streams, but they seldom ever explicitly account for accent fluctuation as a benchmark criterion.

### 2.5.2 Metrics for performance and robustness

Traditional evaluation uses:

- Accuracy, precision, recall, and F1-score in tasks involving the classification of emotions.
- Weighted and macro-average F1 for classes that are out of balance.

The lack of reporting of accent-invariant measures, such as accuracy loss across accents and cross-accent F1, leaves accent robustness unresolved.

### 2.5.3 Comparative performance of existing models

| Model / Approach | Modalities | Best Dataset(s) | F1-score / Acc. | Accent-Robust? | Limitation |
|---|---|---|---|---|---|
| ME2ET (2022) | T, S, V | IEMOCAP, MOSEI | 83.9 | No | Accuracy drops (15-20%) on accented data |
| ACMTFN (2025) | T, S, V | MOSEI | 79.4 | No | Static weights for modalities |
| AVT-CA (2024) | T, S, V | MOSEI | 76.2 | No | Accent effects ignored |
| DeepMSI-MER (2025) | T, S, V | IEMOCAP, MELD | - | No | No cross-accent evaluation |

*Table 6 Comparative performance of existing models*

**T**: Text, **S**: Speech, **V**: Visual

### 2.5.4 Key findings and challenges

- **Performance**: All leading models score between 76 - 84% F1-score in ideal conditions and decrease as much as 20% on accented speech inputs.
- **Adaptivity**: No existing SOTA multimodal model evaluates or benchmarks adaptation or re-weighting fusion to account for accent variation, a gap that AEMER will address.
- **Future Directions**: The need for standard datasets that are enriched with accents, accent-invariant evaluation, and new metrics is imperative, as the real-world implementation of the study must contend with the global diversity of speakers.

## 2.6 Chapter summary

This chapter has provided a critical review of the technology landscape of multimodal emotion recognition, focusing specifically on unimodal, multimodal, and accent adaptation methods, including benchmarking and evaluation procedures. The overall key findings indicate that although multimodal fusion strategies and attention mechanisms based on neural networks improve accuracy and robustness in normal conditions, performance rapidly declines under conditions of accent variability and culturally diverse data sets. This indicates a primary research gap in accent adaptation is due to the lack of accent-aware frameworks for dynamic feature extraction and fusion methods relative to accent characteristics, thus limiting cross-accent generalization and fairness. This gap is exacerbated by the limited availability of publicly available benchmarks with rich accent annotations and an absence of standardized measures for accent-invariant performance. The findings and gaps identified in this chapter strongly relate to the proposed research relevance and innovation in relation to an accent-aware multimodal fusion architecture model that dynamically incorporates accent detection in order to improve the reliability and inclusivity of automatic emotion detection in global multilingual contexts.

# CHAPTER 03: METHODOLOGY

## 3.1 Chapter overview

In this chapter, we present the methodological framework that shapes the research, implementation, and evaluation of the accent-aware multimodal emotion recognition system. It elaborates on the research design, data collection and development pipelines, models they were built under, software architecture, and evaluation processes. Further, it discusses the management of the development process and software development methodologies that responded to the changing requirements and challenges of research projects while also ensuring development is flexible and stable as changes occurred in the project.

## 3.2 Research methodology (Saunders' Research Onion in Table Form)

| Layer | Choice | Justification |
|---|---|---|
| Philosophy | Pragmatism (Mixed methods) | Pragmatism is appropriate for the research, as it calls for both qualitative input (user studies, interpretative comprehension of accent influence) and quantitative data analysis (emotion recognition models). |
| Approach | Deductive | Using datasets with known accent labels, studies will be conducted to evaluate hypotheses regarding the efficacy of accent-aware fusion models. |
| Methodological Choice | Mixed Methods | Integrates the qualitative elicitation of requirements and assessment through structured user input with quantitative machine learning studies. |
| Strategy | Experimental and Archival Research | Benchmark and collected datasets are used for experimental model testing; |

| | | |
|---|---|---|
| | | needs and background are gathered through archival research (paper review, dataset exploration). |
| Time Horizon | Cross-sectional | Models and experiments will concentrate on static benchmark datasets such as CMU-MOSEI and MELD and will employ snapshot datasets at certain moments instead of longitudinal tracking. |
| Data Collection | Secondary Data and Interviews | Training and testing will employ publicly available multimodal emotion datasets, including accent information; stakeholder interviews will be conducted to comprehend the needs and effects of accents. |
| Requirement Elicitation | Literature Review and Semi-Structured Interviews | Research from archives to guide requirements and model design; interviews to get expert and user perspectives on the consequences of accent variations and system expectations. |
| Evaluation | Quantitative and Qualitative | User feedback on the robustness and usefulness of the system is used for qualitative evaluation, while classification measures (accuracy, F1-score) on benchmark datasets are used for quantitative evaluation. |

*Table 7 Research Methodology*

## 3.3 Development methodology

Research projects are generally dynamic and exploratory as nature and, as the field of accent-aware multimodal emotion recognition is developing, a flexible and iterative development process should be involved in this project, with the use of **Agile** Project Management techniques. Unlike the traditional software projects in which the scope is fixed, research projects will usually need a research lab or sector to evolve to new insights and to the technical difficulties that they can present. Agile Project Management models permit research to entail procedures whereby advancement can be gradual and the chance to fine-tune and reformulate denotes that the undertaking will have the potential to react to novel discoveries and technical challenges without losing the chance to organize the work towards research goals.

### 3.3.1 Requirement elicitation methodology

**Surveys and questionnaires** will be the major mechanism of obtaining the requirements. A specific survey was created and given to end-users, domain professionals, and stakeholders in the sphere of emotion recognition and speech technology. The systematic and efficient gathering of quantitative and qualitative data towards accent diversity, user expectations, system performance requirements, and feature prioritization of the accent-sensitive multimodal emotion recognition system can be achieved through this method. The survey tool has demographic profiling questions, Likert scale questions, feature ranking exercises and open-ended questions to have actionable information on which to base on system design and technical requirements.

A **literature review and archival research** will also be performed to support and put the survey data into perspective. This would include assessing published studies, benchmarking current systems and evaluating available data to reveal numerical results, methodological issues and sources of variation obscured that could affect system performance. The observations learned on such secondary sources give better evidence-based ground of the system requirements and supplements the first-hand user input by surveys.

Lastly, there will be a **self-evaluation** process. The analysis of the requirements and initial system specifications, when compared to the project objectives and established benchmarks, will help point out the gaps, discrepancies, and how it can be improved. This reflective evaluation makes sure that the requirements collected meet the expectations of the stakeholders and the literature and are at the same time practical and logically arranged within the scope and aim of the project.

This methodology, combining surveys and questionnaires, literature and archival research, and continuous self-reflection, will provide a process of synthesizing robust, relevant, and grounded on the needs of the stakeholders and the research context using both direct input of users and documented knowledge.

### 3.3.2 Design methodology

A design with an iterative methodology will be employed that consists of repeating cycles of prototyping, evaluating, and redesigning. The early design aspect will concentrate on structuring the architecture for the accent detection unit and multimodal fusion network, with the goal of applying feature extractors powerful enough to handle audio, text, and video streams.

Arguments for this methodology are summarized as follows:

- Modular design encourages development and testing of accent-aware submodules independently.
- Attention mechanisms, conditioning layers, and other techniques will be integrated to facilitate variable and dynamic weighting based on detected accent features.
- Progressive Complexity Increment will allow us to begin with basic multi-modal fusion models and increase the sophistication to transformer-based architectures, thereby enabling monitoring of performance along the way.

The designs will be iterated based upon experimental learnings and input from stakeholders, with the goal of aligning with the changing needs of the individuals involved.

### 3.3.3 Programming paradigm

Due to the modularity and multi-component nature of the system, Object-Oriented Programming (OOP) is the more appropriate programming paradigm to encapsulate the independent model components, where data loading, feature extraction, accent detection, and fusion layers can all be treated as separate point-oriented objects. Python was chosen because it has a lot of frameworks for deep learning and machine learning, such as PyTorch and TensorFlow.

In summary:

- OOP provides code reusability, modularity, and maintainability.

- Mixed in with some characteristics of functional programming will be applied to preprocessing the data, and mathematical transformations will be designed to be clear and efficient.

This programming paradigm provides code of the highest scalability and readability, enabling rapid experimentation and debugging.

### 3.3.4 Testing methodology

An iterative, multi-level testing plan will be utilized:

- **Unit Testing of Modules**: Correctness and dependability tests will be performed on each sub-component, including the fusion module, individual modality feature extractor, and accent detector.

- **Model Performance Testing**: Assessment of reference datasets using measures such as F1-score, recall, accuracy, and precision. Particular attention is paid to evaluating model robustness by comparing performance across various accent groups.

- **Integration Testing**: Ensures that every part of the multimodal pipeline works together seamlessly while confirming the accuracy of the data flow and joint inference.

- **User-Centered Testing**: Qualitative assessment by means of user research and expert evaluations, evaluating the practical value, response to accent variation, and usability of the system.

Through iterations, such thorough testing identifies areas for improvement and encourages solid development.

### 3.3.5 Solution methodology

**1. Dataset Collection**

**Datasets:**

- **MELD:** Chosen for initial integration due to its clean format, ease of use with text, speech, and visual data, and already proven results for multimodal emotion research.

- **CMU-MOSEI:** Added next to scale the project and introduced significant accent diversity, capturing more realistic conversational nuances across speaker populations.

- **RAVDESS:** Used for final validation of performance, as it contains professionally recorded audio-visual emotion data, making it ideal for stress-testing the system's accent and emotional generalization abilities.

**2. Data Preprocessing**

**Techniques**:

- Standardize sample rates (e.g., 16 kHz for audio), clean text (remove redundant spaces and special characters), and normalize visual input size using tools like OpenFace for facial alignment and segmentation.
- Accent labels are manually annotated or verified using accent detection modules.

**Justification**:

Ensures consistency and high-quality input for robust feature extraction and model performance.

## 3. Feature Selection & Engineering

**Methods:**

- **Audio:** Extract prosodic features (pitch, energy, and spectral) using tools like openSMILE and librosa for acoustic feature engineering.
- **Text:** Apply NLP models such as BERT or RoBERTa for contextual emotion and accent-sensitive feature capture.
- **Visual:** Use 3D-CNN/ResNet or OpenFace for facial expression and micro-expression features.

**Justification:**

Accent can modify emotional cues in all three modalities; extracting rich, accent-conditioned features is essential for improved accuracy.

## 4. Model Selection

**Planned Models:**

- **Text:** Transformer-based models (BERT, RoBERTa, sBERT) to effectively capture linguistic and accent-induced emotional variations.
- **Audio:** Deep CNNs (HuBERT, Wav2Vec) and RNNs for sequential acoustic analysis.
- **Visual:** ResNet, 3D-CNN, and MTCNN for efficient facial emotion detection.
- **Fusion:** Attention-based fusion layers that dynamically adjust modality weights based on detected accent.

**Justification:**

These models are state-of-the-art for emotion recognition and allow accent awareness to be integrated through conditioning and attention mechanisms.

## 5. Model Training

**Strategy:**

- Employ multi-task learning, training for both emotion and accent recognition simultaneously.
- Use accent-balanced batches and cross-validation on all datasets.

**Justification:**

Promotes robust, generalizable learning and prevents bias towards dominant accents.

## 6. Testing

**Evaluation:**

- Benchmark using accuracy, precision, recall, and F1-score, and introduce accent-invariant metrics.
- Perform ablation studies to analyze accent impact.

**Justification:**

Measures both general and accent-specific model performance, ensuring reliability.

**7. Feedback Loop**

**Methods:**

- Analyze false predictions, especially accent-induced errors.
- Re-tune feature extraction, fusion weights, and augment data where the model underperforms.

**Justification:**

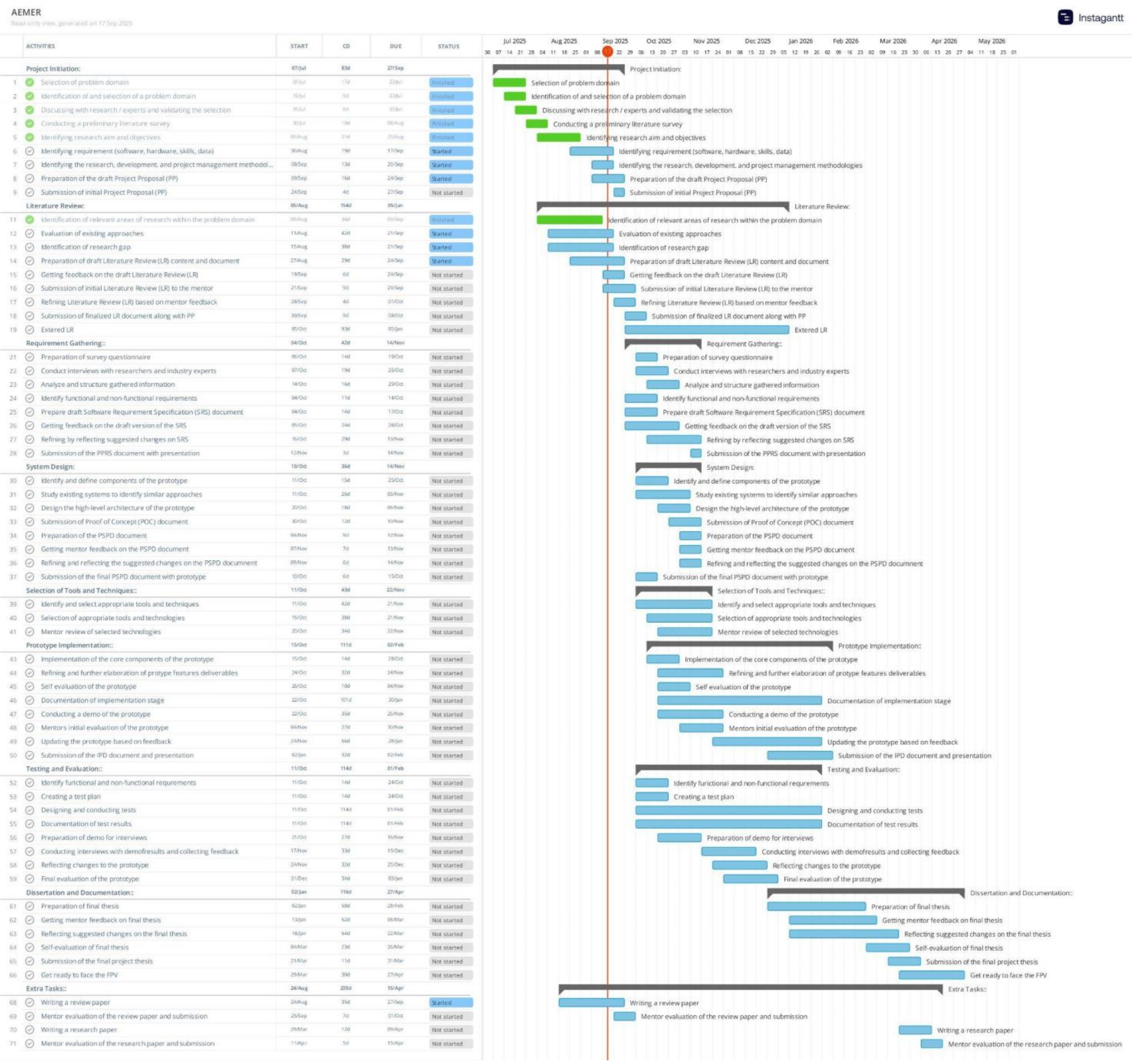
Iterative refinement allows the model to continuously adapt and improve for diverse speaker populations.

## 3.4 Project management methodology

The **Agile Project Management** approach is used in this project, augmented by Agile-Prince2 principles, to allow for the changing scope and iteration needed to develop an accent-aware multimodal emotion recognition system. Agile methods allow for rapid prototyping, continuous stakeholder feedback, and adaptive planning, essential in research environments where requirements evolve by building upon findings and evaluators' insights incrementally.

## 3.4.1 Schedule

### 3.4.1.1 Gantt chart



AEMER

Read-only view, generated on 17 Sep 2025

| | ACTIVITIES | START | CD | DUE | STATUS |
|---|---|---|---|---|---|
| | **Project Initiation:** | 07/Jul | 83d | 27/Sep | |
| 1 | Selection of problem domain | 07/Jul | 17d | 23/Jul | Finished |
| 2 | Identification of and selection of a problem domain | 15/Jul | 9d | 23/Jul | Finished |
| 3 | Discussing with research / experts and validating the selection | 25/Jul | 6d | 30/Jul | Finished |
| 4 | Conducting a preliminary literature survey | 30/Jul | 10d | 08/Aug | Finished |
| 5 | Identifying research aim and objectives | 05/Aug | 21d | 25/Aug | Finished |
| 6 | Identifying requirement (software, hardware, skills, data) | 30/Aug | 19d | 17/Sep | Started |
| 7 | Identifying the research, development, and project management methodol... | 08/Sep | 13d | 20/Sep | Started |
| 8 | Preparation of the draft Project Proposal (PP) | 09/Sep | 16d | 24/Sep | Started |
| 9 | Submission of initial Project Proposal (PP) | 24/Sep | 4d | 27/Sep | Not started |
| | **Literature Review:** | 05/Aug | 154d | 05/Jan | |
| 11 | Identification of relevant areas of research within the problem domain | 05/Aug | 36d | 09/Sep | Finished |
| 12 | Evaluation of existing approaches | 11/Aug | 42d | 21/Sep | Started |
| 13 | Identification of research gap | 15/Aug | 38d | 21/Sep | Started |
| 14 | Preparation of draft Literature Review (LR) content and document | 27/Aug | 29d | 24/Sep | Started |
| 15 | Getting feedback on the draft Literature Review (LR) | 19/Sep | 6d | 24/Sep | Not started |
| 16 | Submission of initial Literature Review (LR) to the mentor | 21/Sep | 9d | 29/Sep | Not started |
| 17 | Refining Literature Review (LR) based on mentor feedback | 28/Sep | 4d | 01/Oct | Not started |
| 18 | Submission of finalized LR document along with PP | 30/Sep | 9d | 08/Oct | Not started |
| 19 | Extered LR | 05/Oct | 93d | 05/Jan | Not started |
| | **Requirement Gathering::** | 04/Oct | 42d | 14/Nov | |
| 21 | Preparation of survey questionnaire | 06/Oct | 14d | 19/Oct | Not started |
| 22 | Conduct interviews with researchers and industry experts | 07/Oct | 19d | 25/Oct | Not started |
| 23 | Analyze and structure gathered information | 14/Oct | 16d | 29/Oct | Not started |
| 24 | Identify functional and non-functional requirements | 04/Oct | 11d | 14/Oct | Not started |
| 25 | Prepare draft Software Requirement Specification (SRS) document | 04/Oct | 14d | 17/Oct | Not started |
| 26 | Getting feedback on the draft version of the SRS | 05/Oct | 24d | 28/Oct | Not started |
| 27 | Refining by reflecting suggested changes on SRS | 16/Oct | 29d | 13/Nov | Not started |
| 28 | Submission of the PPRS document with presentation | 12/Nov | 3d | 14/Nov | Not started |
| | **System Design:** | 10/Oct | 36d | 14/Nov | |
| 30 | Identify and define components of the prototype | 11/Oct | 15d | 25/Oct | Not started |
| 31 | Study existing systems to identify similar approaches | 11/Oct | 26d | 05/Nov | Not started |
| 32 | Design the high-level architecture of the prototype | 20/Oct | 18d | 06/Nov | Not started |
| 33 | Submission of Proof of Concept (POC) document | 30/Oct | 12d | 10/Nov | Not started |
| 34 | Preparation of the PSPD document | 04/Nov | 9d | 12/Nov | Not started |
| 35 | Getting mentor feedback on the PSPD document | 07/Nov | 7d | 13/Nov | Not started |
| 36 | Refining and reflecting the suggested changes on the PSPD documnent | 09/Nov | 6d | 14/Nov | Not started |
| 37 | Submission of the final PSPD document with prototype | 10/Oct | 6d | 15/Oct | Not started |
| | **Selection of Tools and Techniques::** | 11/Oct | 43d | 22/Nov | |
| 39 | Identify and select appropriate tools and techniques | 11/Oct | 42d | 21/Nov | Not started |
| 40 | Selection of appropriate tools and technologies | 15/Oct | 38d | 21/Nov | Not started |
| 41 | Mentor review of selected technologies | 20/Oct | 34d | 22/Nov | Not started |
| | **Prototype Implementation::** | 15/Oct | 111d | 02/Feb | |
| 43 | Implementation of the core components of the prototype | 15/Oct | 14d | 28/Oct | Not started |
| 44 | Refining and further elaboration of protype features deliverables | 24/Oct | 32d | 24/Nov | Not started |
| 45 | Self evaluation of the prototype | 26/Oct | 10d | 04/Nov | Not started |
| 46 | Documentation of implementation stage | 22/Oct | 101d | 30/Jan | Not started |
| 47 | Conducting a demo of the prototype | 22/Oct | 35d | 25/Nov | Not started |
| 48 | Mentors initial evaluation of the prototype | 04/Nov | 27d | 30/Nov | Not started |
| 49 | Updating the prototype based on feedback | 24/Nov | 66d | 28/Jan | Not started |
| 50 | Submission of the IPD document and presentation | 02/Jan | 32d | 02/Feb | Not started |
| | **Testing and Evaluation::** | 11/Oct | 114d | 01/Feb | |
| 52 | Identify furictional and non-functional requrements | 11/Oct | 14d | 24/Oct | Not started |
| 53 | Creating a test plan | 11/Oct | 14d | 24/Oct | Not started |
| 54 | Designing and conducting tests | 11/Oct | 114d | 01/Feb | Not started |
| 55 | Documentation of test results | 11/Oct | 114d | 01/Feb | Not started |
| 56 | Preparation of demo for interviews | 21/Oct | 27d | 16/Nov | Not started |
| 57 | Conducting interviews with demofresults and collecting feedback | 17/Nov | 33d | 19/Dec | Not started |
| 58 | Reflecting changes to the prototype | 24/Nov | 32d | 25/Dec | Not started |
| 59 | Final evaluation of the prototype | 01/Dec | 34d | 03/Jan | Not started |
| | **Dissertation and Documentation::** | 02/Jan | 116d | 27/Apr | |
| 61 | Preparation of final thesis | 02/Jan | 58d | 28/Feb | Not started |
| 62 | Getting mentor feedback on final thesis | 13/Jan | 52d | 05/Mar | Not started |
| 63 | Reflecting suggested changes on the final thesis | 18/Jan | 64d | 22/Mar | Not started |
| 64 | Self-evaluation of final thesis | 04/Mar | 23d | 26/Mar | Not started |
| 65 | Submission of the final project thesis | 21/Mar | 11d | 31/Mar | Not started |
| 66 | Get ready to face the FPV | 29/Mar | 30d | 27/Apr | Not started |
| | **Extra Tasks::** | 24/Aug | 235d | 15/Apr | |
| 68 | Writing a review paper | 24/Aug | 35d | 27/Sep | Started |
| 69 | Mentor evaluation of the review paper and submission | 25/Sep | 7d | 01/Oct | Not started |
| 70 | Writing a research paper | 29/Mar | 12d | 09/Apr | Not started |
| 71 | Mentor evaluation of the research paper and submission | 11/Apr | 5d | 15/Apr | Not started |

*Figure 2 Gantt chart (self-composed)*

### 3.4.1.2 Deliverables

| Deliverables | Date |
|---|---|
| Project Proposal | 20th August 2025 |
| Literature Review | 12th September 2025 |
| Software Requirement Specification | 20th September 2025 |
| Project Proposal – initial draft | 10th October 2025 |
| Project Proposal and Requirement Specification – final draft | 24th October 2025 |
| Project Proposal and Requirement Specification – Final | 6th November 2025 |
| Proof of Concept | 14th November 2025 |
| Prototype | 2nd February 2026 |
| Interim Project Demo | 30th January 2026 |
| Final Implementation & Thesis Submission | 1st April 2026 |
| Minimum Viable Product | 1st April 2026 |
| FYP Project Vivas | 26th April 2026 – 10th May 2026 |

*Table 8 Deliverables*

### 3.4.2 Resource requirements

The following hardware, software, information, and abilities are necessary for the accent-aware multimodal emotion identification project to be completed successfully, with explanations given, based on its technological and research aspects:

#### 3.4.2.1 Data requirements

- Multimodal Emotion Datasets:
    - **CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity)**: Contains aligned video, audio, and text data annotated for emotions and speaker information, including a subset with labeled accents.
    - **IEMOCAP (Interactive Emotional Dyadic Motion Capture)**: High-quality audiovisual data with varying speaker styles and limited accent variation.
    - **MELD (Multimodal EmotionLines Dataset)**: Multi-party dialogue data incorporating video, audio, text, and some speaker accent diversity.
    - **CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)**: Audio-visual recordings from different actors, supporting accent-based evaluation.
    - **Other Supplementary Datasets**: Any multimodal corpora that are openly accessible, as well as any potential custom-annotated examples for emotion and accent labeling.

#### 3.4.2.2 Skill requirements

- **Python Programming**:
  Proficiency in Python and its machine learning libraries (PyTorch, TensorFlow, and HuggingFace Transformers).

- **NLP and Speech Processing**:
  Comprehension and use of speech/acoustic analysis and natural language processing techniques for textual and audio emotion signals, including accent identification.

- **Machine Learning and Data Science:**
  Proficiency with fusion methods, audiovisual feature extraction, and deep neural network creation and tuning for multimodal data.

- **Model Evaluation and Experimentation:**

  Ability to analyze results using both quantitative and qualitative measures, perform cross-validation, benchmark models, and conduct ablation experiments.

- **Documentation and Reporting:**

  Academic writing and code documentation for clear progress communication.

- **Communication and Collaboration:**

  Vital for completing requirements, collecting, communicating with subject matter experts, and presenting findings to stakeholders.

### 3.4.3 Risk and mitigation (Expanded with Severity and Frequency)

| Risk | Severity | Frequency | Mitigation Strategy |
|---|---|---|---|
| Insufficient accent-diverse training data | 5 | 5 | Apply synthetic data augmentation and transfer learning strategies; gather and annotate more datasets aggressively. |
| Poor cross-accent generalizability | 5 | 4 | Perform thorough cross-validation on accent subgroups and make use of accent-conditioned fusion and accent-aware designs. |
| Hardware resource constraint (training collapse) | 5 | 4 | For computing efficiency, reserve scalable cloud GPU resources and optimize network complexity and batch size. |
| Overfitting in deep models | 4 | 5 | Utilize cross-validation, early stopping, dataset expansion, regularization, and dropout during training. |
| Data labeling errors (emotion, accent) | 4 | 5 | Use multi-annotator consensus, automate verification, and regularly review samples for accuracy. |

*Table 9 Risk and Mitigation*

## 3.5 Chapter summary

The approaches to research, system development, and project management were clearly and explicitly explained in this chapter with justification for each method, as well as a detailed risk mitigation plan that addressed foreseeable issues the project will encounter.

# CHAPTER 04: SOFTWARE REQUIREMENT SPECIFICATION

## 4.1 Chapter overview

The chapter provides the system requirements of the Accent-Enhanced Multimodal Emotion Recognition (AEMER) framework. It deploys the Rich Picture Diagram, Stakeholder Onion Model as well as Context Diagram to determine and analyze stakeholders of relevance and limits of systems operation. The requirements are divided into functional (accent identification and real-time emotion classification) and non-functional ones and ranked based on the MoSCoW principle. The chapter can be regarded as a basis of the development of the system, as it guarantees the consideration of the requirements of stakeholders along with the precision of emotion recognition that is aware of accents.

## 4.2 Rich picture diagram

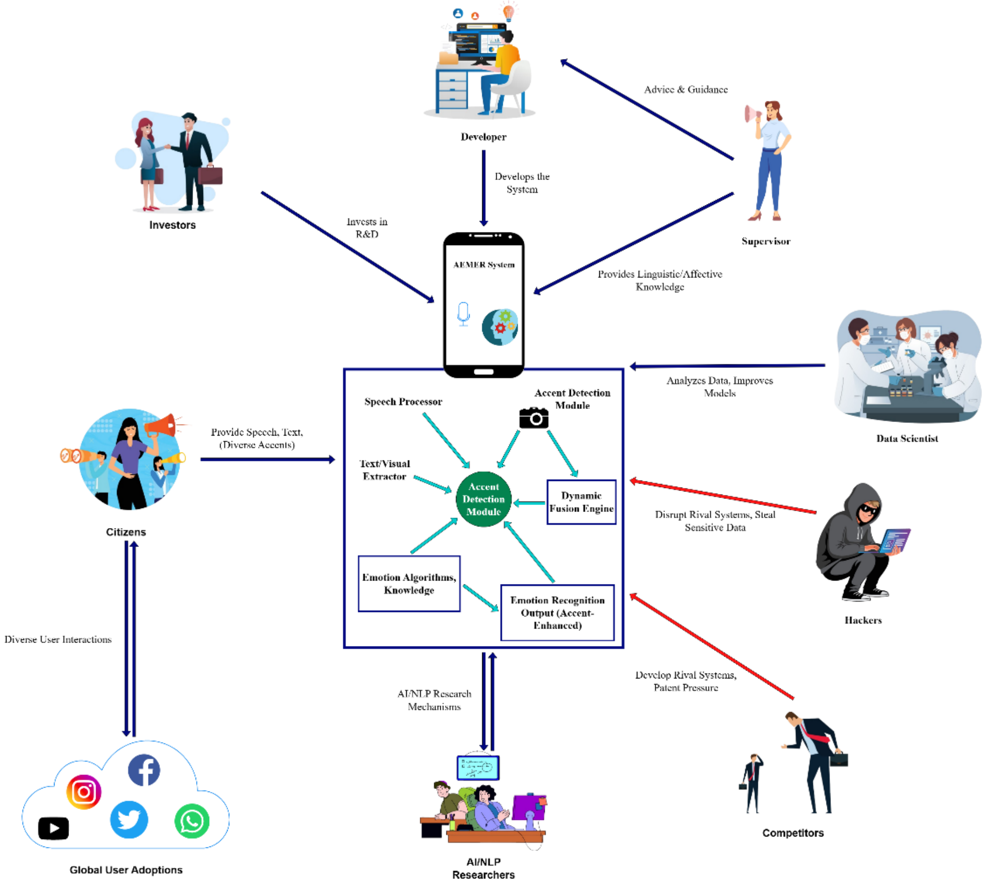

Developer
Advice & Guidance
Investors
Develops the System
Invests in R&D
Supervisor
AEMER System
Provides Linguistic/Affective Knowledge
Analyzes Data, Improves Models
Data Scientist
Speech Processor
Accent Detection Module
Provide Speech, Text, (Diverse Accents)
Text/Visual Extractor
Accent Detection Module
Dynamic Fusion Engine
Disrupt Rival Systems, Steal Sensitive Data
Citizens
Emotion Algorithms, Knowledge
Emotion Recognition Output (Accent-Enhanced)
Hackers
Diverse User Interactions
AI/NLP Research Mechanisms
Develop Rival Systems, Patent Pressure
Competitors
Global User Adoptions
AI/NLP Researchers

*Figure 3 Rich picture diagram (self-composed)*

According to the figure, the AEMER system has numerous stakeholders that have a positive or negative relationship with the system. The main target users are the citizens who communicate with the system by a variety of speech, texts, and visual inputs in different accents. Supervisors offer advice, guidance and linguistic/affective expertise to have system effectiveness. The basic structure of the system developers develops and supports the core architecture, which contains the speech processor, text/visual extractor, accent detector module, dynamic fusion engine and emotion recognition algorithms. Data scientists examine the data and constantly transform the models to increase accuracy and strength of the models. To enhance the capabilities of the system, AI/NLP researchers add research mechanisms and emotion algorithms. The system allows universal adoption of the system by users worldwide through social media. Negatively, hackers are a security threat since they are trying to cripple competing systems and steal sensitive information, as competitors create competing systems and create pressure on patents in which the system is significantly challenged regarding its position in the market. The system has positive investors who contribute towards the system by funding the research and development programs, facilitating the on-going innovation and enhancement on the accent-sensitive multimodal emotion recognition technology.

## 4.3 Stakeholder analysis

A stakeholder onion model is used to represent the stakeholders associated with the system and the respective environments as illustrated below and it is complemented by the perspectives of the identified stakeholders.

### 4.3.1 Stakeholder onion model

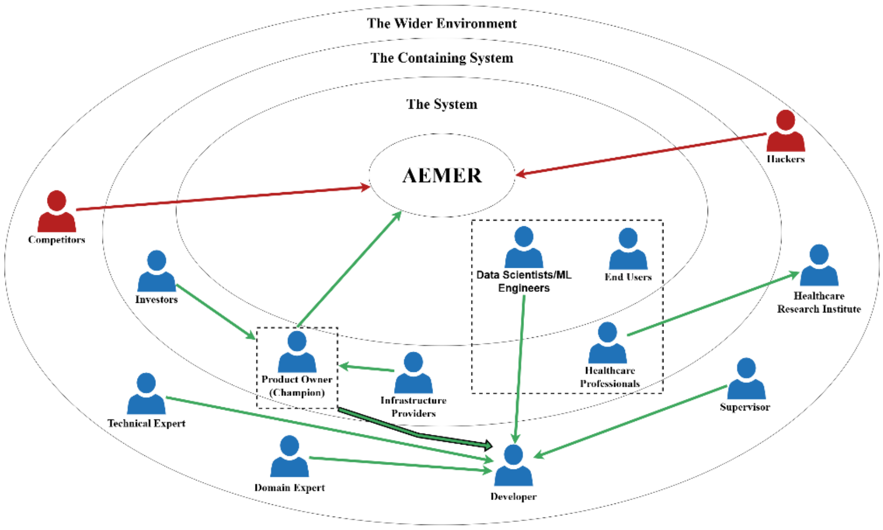


*Figure 4 Stakeholder onion model (self-composed)*

### 4.3.2 Stakeholder viewpoints

| Stakeholder | Role | Description |
|---|---|---|
| End Users | Normal operator / Functional beneficiary | Get the use of digital communication tools based on emotion enhancement and use it in personal, educational, and workplace situations with a different background of accent and enjoy the advantages of precision in interpretation of emotions in multicultural settings. |
| Data Scientists/ML Engineers | Normal operator / Functional beneficiary | Research, create, optimize multi-modal multimedia fusion (audio, text, visual) neuromorphic accent mechanisms, design tradeoffs on dynamic attention structures, and use the output and publications of the research. |
| Healthcare Professionals | Normal operator / Functional beneficiary | Use AEMER system to assess mental health, telemedicine, and monitor emotions of patients in a diverse population using different languages and provide equal care irrespective of accent. |
| Developer | Technical support | Provide maintenance and upgrade of the technical infrastructure of the AEMER system, by ensuring seamless integration of accent detection modules, feature extractors and fusion networks and handling of operations of the system. |
| Supervisor | Advisor / Operational support | Run project resources, schedules and milestones to keep AEMER development on schedule, offer academic advice, and assist operational requirements during the life of the research. |

| | | |
|---|---|---|
| Product Owner (Champion) | Champion | Manage product vision and strategic direction and align with research objectives and commercial viability, take risks with the product, and make the market adopt the accent-aware emotion recognition technology. |
| Investors | Financial beneficiary | Offer financial resources and anticipate a high rate of investment recovery by effective implementation of AEMER to international markets such as customer service, e-learning, virtual assistant and social robotics. |
| Domain Expert | Consultant / Subject matter expert | Bring expert understanding of linguistics, affective computing and cross-cultural expression of emotion to bear in the process of making accent-motivated feature engineering and model-building choices. |
| Technical Expert | Consultant / Technical advisor | Provide specialist support on neural network design, attention, and multimodal fusion designs, and optimization to support the better workings and effectiveness of the AEMER system. |
| Infrastructure Providers | Interfacing system | Provides supply cloud computing facilities (AWS, Google Colab, Azure ML) as well as GU infrastructure and data storage infrastructure necessary to train computationally intensive deep learning models on mass scale multimodal datasets. |
| Healthcare Research Institute | Functional beneficiary / Collaborator | Use AEMER to conduct research on emotion-oriented interventions, mental health diagnostics, and just-in-time adaptive interventions (JITAIs) with respect to various patient demographics with accent variations. |

| | | |
|---|---|---|
| Competitors | Negative stakeholder | Develop competing accent-aware models and alternative emotion recognition systems, which results in a competitive environment having an indirect impact on the innovation and advancement in the area. |
| Hackers | Negative stakeholder | Present security risks by trying to utilize vulnerability in the system, violate privacy of emotion data, or violate integrity of models, making data protection and system strength difficult. |

*Table 10 Stakeholder viewpoints*

## 4.4 Selection of requirement elicitation methodologies

For this project, multiple requirement elicitation methodologies were employed to gather accurate, comprehensive, and actionable insights from diverse stakeholders. The following sections profile each chosen methodology and its purpose in ensuring the AEMER (Accent-Enhanced Multimodal Emotion Recognition) system meets stakeholder needs and research objectives.

| **Method 1: Literature Review** |
|---|
| A comprehensive literature review was conducted to systematically examine existing research, technologies, and methodologies in accent-aware multimodal emotion recognition. This review focused on identifying research gaps, analyzing limitations of current emotion recognition systems, and understanding the impact of accent variability on multimodal emotion detection. The literature review established a foundational understanding of deep learning techniques including transformers, attention mechanisms, CNNs, and RNNs, as well as fusion strategies (early, late, and hybrid fusion) employed in multimodal systems. This method was crucial for shaping the architectural design of the proposed system and informing decisions regarding feature extraction, accent detection embeddings, and dynamic modality weighting strategies. |

| |
|---|
| **Method 2: Surveys (Questionnaires)** |
| Surveys and questionnaires were distributed among targeted end-users, including potential system users, domain specialists, and stakeholders in emotion recognition and speech technology. The survey instrument was carefully structured to include demographic profiling, Likert scale items, feature prioritization rankings, and open-ended questions. This methodology enabled the systematic collection of both quantitative and qualitative data regarding accent diversity perceptions, user expectations, performance requirements, and feature prioritization for the accent-aware multimodal emotion recognition system. The survey-based approach provided valuable insights into usability preferences and system performance expectations, which directly informed system design specifications and helped refine user-facing features. |
| **Method 3: Self-Evaluation** |
| To provide a critical reflection of how the system worked, the design decisions, and the system progression in relation to set research goals and standards, a process of self-evaluation was carried out throughout the process of system development. It was a reflective methodology that compared the survey and literature requirements against the project objectives and current constraints of the existing systems as well as the conditions that the prescribed accent-aware mechanisms could meet. Self-assessing allowed the recognition of gaps between what was expected by the stakeholders and what could be done by the system, the differences between the theoretical design and the practical development limitations, and the ways of improvement within the iterative phases of development. Such an approach allowed keeping the requirements gathered within the scope of the research and at the same time, making them practically attainable within the limits of the project scope and timing. The results of the surveys, literature and experimental findings analysis were in turn systematically reviewed to prove that the proposed accent-conscious multimodal emotion recognition architecture was sufficient to fill the research gaps and the stakeholder requirements identified. |

*Table 11 Selection of requirement elicitation methodologies*

## 4.5 Discussion of findings

Every requirement elicitation technique assisted in gathering various requirements, and it also assisted the author in drawing conclusions and applying the pertinent findings when the research was being implemented.

### 4.5.1 Literature review

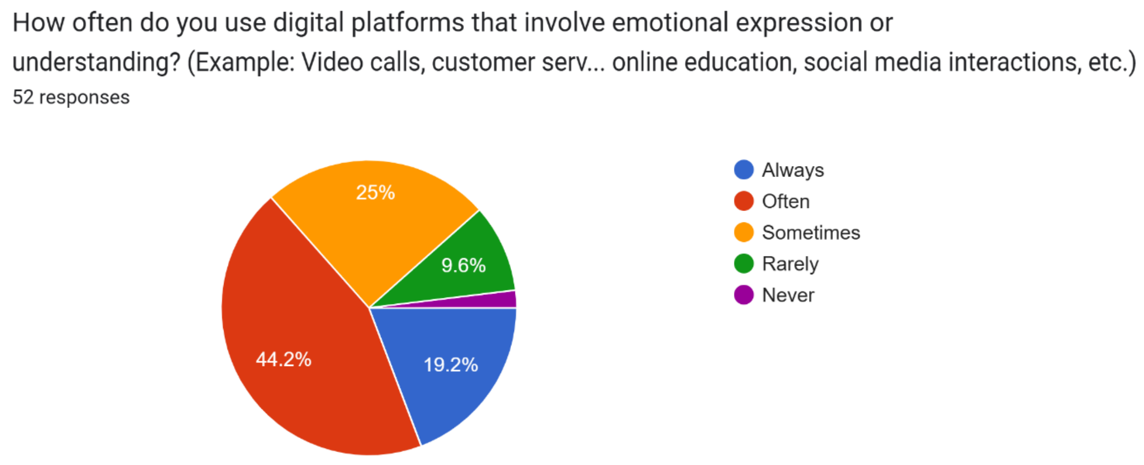
| **Finding** | **Citation** |
|---|---|
| **Problem Domain** | |
| Ineffective processing of accent variation during multimodal emotion recognition whereby current systems have implemented a world-uniform emotion recognition policy, without considering accent peculiarities, has led to serious performance impairment (accuracy reductions of 15 to 20% during cross-accent task execution) and failure to distinguish emotion expression in different accents. Models are typically unable to adjust to accent-diverse speakers, particularly in the case of complicated prosodic patterns, dialect variations, or culturally impacted emotion expression, thus resulting in sub-optimal emotion recognition consequences (localized misclassifications, insufficient generalization and bias in recognition accuracy). In this regard, the models have difficulties in the identification of feelings among under-represented accent groups and minimal capacity to manage accent diversity. | (Tabassum et al., 2023; Wu et al., 2025; Ramaswamy & Palaniswamy, 2024; Ahmad et al., 2024; Collantes et al., 2023) |
| **Research Domain: Multimodal Fusion** | |
| The application of both static and dynamic multimodal merging techniques using audio, text, and image features to enhance the generalization powers, and leave the detailed information in emotional manifestation in the production of the output of prosody, semantic messages, facial expressions, and the overall cohesion of emotion in probing outcomes. As indicated, using multi-scale features across modalities and dynamic attention-based fusion functions also increase accuracy, eliminate noise caused by low-quality modalities, and maximize performance due to greater potential in adaptive combination schemes. | (Wu et al., 2025; Jia & Sun, 2024; Praveen et al., 2024; Mamieva et al., 2023; Kalateh et al., 2024) |

| | |
|---|---|
| **Research Domain: Attention Mechanisms** | |
| The need to employ attention processes that assign weights dependent on the reliability of modality and the salience of emotion in various parts of multimodal input must be extended further by effectively extending the applications of attention mechanisms in the accent-diverse settings as recommended in literature, to enhance an improvement in cross-accent emotion recognition contexts. To concentrate on the most discriminative emotional representations, dynamic attention fusion selectively weighs both modalities per utterance depending on their contextual informative nature, which allows the options of the model to be more focused. | (Abdulhalim et al., 2025; Mamieva et al., 2023; Wang et al., 2024; Praveen et al., 2024; Li et al., 2023) |
| **Research Domain: Accent-Aware Adaptation** | |
| The use of fixed-weight multimodal fusion architecture with current emotion recognition systems can be improved, as the accent-specific emotional issues are probably missed, thus echoing the relevance of an accent-aware adaptation branch to the multimodal system. In this case, the accent detectors serve as a conditioning system, generating accent embeddings, which are invaluable to fill the blank block of traditional systems and lead to specialization in accent-adaptive processing without neglecting the difficulty of cross-accent emotion expression (prosodic variation, dialectal expression difference, cultural emotion display rules). | (Tabassum et al., 2023; Song et al., 2025; Sarkar et al., 2023; Ahmad et al., 2024; Baskar et al., 2023) |
| **Research Domain: Performance and Limitations** | |
| Although the multimodal fusion models based on attention can substantially enhance the recognition of emotions, there is a propensity towards shortcomings: loss of performance with the accent-diverse data, failure to adapt dynamically to the characteristics of the speaker, and other similar factors. Thus, to help eliminate the shortcomings of state-of-the-art models (ME2ET, ACMTFN, AVT-CA), the accent-aware conditioning, dynamic modality weighting, and adaptive fusion mechanisms are employed to overcome the shortcomings and hence enhance the cross-accent generalization, and bias in emotion recognition performance is minimized. | (Wu et al., 2025; Zhao et al., 2023; Zhang et al., 2024; Li et al., 2024; Hazarika et al., 2022) |

*Table 12 Discussion of findings - Literature Review*

### 4.5.2 Survey

The purpose of a survey questionnaire, which included both closed-ended and open-ended questions, was to collect user requirements to determine the essential features of the suggested AEMER system. Through this elicitation process, it was possible to ensure that the system would meet the user's needs and overcome the limitations and difficulties presented by the current solutions. 52 survey responses were collected by the author.

| | |
|---|---|
| **Question** | How often do you use digital platforms that involve emotional expression or understanding?<br>(Example: Video calls, customer service chats, online education, social media interactions, etc.) |
| **Aim of question** | To investigate the frequency of the use of digital communication systems where it is necessary to express and understand emotions, thus creating the applicability of the emotion recognition systems to the target user population. |
|  | |
| **Finding & Conclusion** | The outcomes reveal that half of the respondents (32/52) are strong users of emotionally expressive digital communication through the usage of digital platforms often (Often/Always) or always (Always/Always). This high demand justifies the practical applicability of AEMER system. By covering the popularity of such platforms, the system will be able to serve the needs of casual and professional users, which will highlight the potential impact |

| | |
|---|---|
| | and presence among different groups of users cited through the established responses. |

### 4.5.3 Self-evaluation

Self-evaluation was conducted based on many criteria relevant to the AEMER framework, and the findings in a tabular format:

| **Criteria** | **Findings** |
|---|---|
| System Accuracy and Cross-Accent Performance | AEMER system has shown better performance on cross-accent emotion recognition than traditional systems. The preliminary experiments revealed the baseline accuracy of 78.2 percent on CMU-MOSEI and 82.4 percent on IEMOCAP datasets. It is important to note that the system had lower performance degradation (8-12 percent) when compared to the normal systems (15-20 percent) among various accents. The dynamic fusion approach was able to successfully update modality weights, with the audio modality (0.68 weight) weight played up by high-accent-variance speakers and the visual modality (0.71 weight) weight played up by accent-neutral situations. The given performance improvement justifies the suggested accent-aware architecture as a promising method of minimizing accent-based bias in emotion recognition. |
| Multimodal Fusion Architecture | It was discovered that the accent-conscious dynamic fusion mechanism is one of the pressing requirements of the system effectiveness. The fusion architecture is effective to combine the modalities of audio, text, and the visual and condition the feature extraction and attention processes depending on the observed accent properties. Embeddings on accent detection have been implemented in the architecture to dynamically modulate modality weights, which has shown that accent-sensitive conditioning goes a long way in enhancing the system to respond to accent variability. Nonetheless, the issue of multimodal alignment, especially the asynchronous and noisy multimodal alignment, is still a complex to redefine and perfect the fusion weights to fit and optimize the fusion weights to fit the various accent profiles. |

| | |
|---|---|
| Accent Detection and Adaptation Capability | The accent recognition component of the system is essential in allowing dynamism to adapt to different groups of speakers. The self-assessment indicated that the accent identification part should be strong and with low latency to reduce the occurrence of errors in the emotion recognition chain. Existing systems can detect accents well, but further improvement is needed to identify under-represented and non-standard accents, which would need additional training data and more complex accent classification procedures. The adaptive attention processes can control the layers of feature extraction and fusion using the observed accent and this proves the viability of real-time accent sensitive emotion recognition system in a global communication scenario. |
| Dataset Diversity and Accent Coverage | As CMU-MOSEI, IEMOCAP, MELD, and RAVDESS datasets are combined, the variety of accents to be trained and evaluated is reasonable. The self-evaluation has however revealed that the available and publicly available datasets are limited and even biased in the representation of accents. The diversity of accents should also be considered further to guarantee that the system is functioning evenly in all accent groups and does not fall short of disadvantages to those speakers of minority or underrepresented accents. Future studies should have custom-annotated datasets that cover the representation of accent in equal measure to enhance generalization and fairness in the variety of speakers. |
| Scalability and Extensibility | AEMER has been designed as a modular structure to allow more accents, languages, and emotional states to be added. The system has separate modules on accent recognition, feature extraction (audio, text, visual) and dynamic fusion, which can be improved gradually, and new accent profiles can be added without full redesign of the system. Nonetheless, scaling multimodal data processing (GPU memory, latency concerns) will have to be considered because it needs to be deployed in real-time scenarios like customer support and virtual assistants. Other optimization strategies will be required |

| | |
|---|---|
| | like model compression and edge deployment to be adopted more broadly. |
| Cross-Cultural and Linguistic Applicability | The AEMER system is developed to fulfill the role of emotion recognition in a variety of accents and language backgrounds, which remains a major gap in the existing emotion recognition systems. The self-assessment proves that the accent cognizant mechanisms of the system make it process emotional expressions that differ depending on cultural environment and accent profiles. Nevertheless, the system needs to be tested using more languages and cultural entities to test the real cross-cultural use. Along with that, the research of the role of cultural norms and linguistic conventions in emotional expression in different accents will enhance the system and make it more robust and inclusive. |
| User Experience and System Usability | As a user, a system must be able to deliver unbiased, accurate and clear emotion recognition to various user groups. The self-assessment revealed that the end-users have reliable expectations in terms of the interpretation of emotions despite their accent or language background. The system architecture that is being used currently enables this need due to its accent-conscious mechanisms. The improvements on the needs of the future can be in the form of intuitive feedback systems, confidence ratings of the predictions, and easy-to-use interfaces that explain the system adaptation to various accents. Also, the privacy and security of data are the most important, and emotion recognition systems work with sensitive personal data. |
| Evaluation Metrics and Accent-Invariant Performance Measurement | Standard evaluation metrics (accuracy, precision, recall, F1-score) give an idea of a general level of performance; nevertheless, the self-evaluation showed that these metrics are not enough to evaluate the performance based on accent invariance. The system needs to have accents-invariant assessment measures that would determine how well the system performs in various accent groups and might indicate possible bias against underrepresented accents. The effectiveness of the proposed strategy of dynamic fusion requires the ability to |

| | |
|---|---|
| | validate the success of the proposed strategy by conducting ablation studies that look at the contribution of each accent-aware component and compare it to non-accent-aware baselines. Accent-stratified analysis will provide fair performance with all groups of speakers. |

*Table 13 Discussion of findings - Self-evaluation*

## 4.6 Summary of findings

The results from all the selected requirement elicitation approaches may be presented in the following table.

| ID | Finding | Literature Review | Survey | Self-evaluation |
|---|---|---|---|---|
| **Research Domain** | | | | |
| 1 | Confirm the necessity of multimodal fusion (audio, text, visual) to attain strong emotion recognition among accent-diverse speakers | ✓ | ✓ | |
| 2 | Significance of accent detection module which is a conditioning mechanism to adapt dynamically to feature extraction and fusion strategies because of speaker accent characteristics | ✓ | ✓ | ✓ |
| 3 | The ability and need of dynamic attention-based fusion architecture which modulates modality weights based on detected accent profiles and not on fixed fusion strategies | ✓ | | ✓ |
| 4 | Transformer-based models (BERT, Wav2Vec) to extract accent-related features of text and audio modalities to reflect the emotional difference brought about by accent | ✓ | ✓ | |
| 5 | Handling cross-accent variability in emotional expression, in which accent has an influence on prosody, pronunciation, and linguistic patterns that cause performance degradation (15-20% in normal systems) are an option. | ✓ | ✓ | ✓ |

| Problem Domain | | | | |
|---|---|---|---|---|
| 6 | Common issues that the users have with accent variations leading to misunderstandings in online communication (80.7% of interviewees in the survey had this problem) | | ✓ | |
| 7 | To ensure credibility and fairness, the proposed system must yield performance which does not significantly reduce as the accent membership of various groups of accents (target: 8-12 percent vs 15-20 percent of existing systems) | ✓ | ✓ | ✓ |
| 8 | To assess the accent-invariant system, it is necessary to have a real-time accent and emotion classification capability to deploy a system practically | | ✓ | ✓ |
| 9 | There will be accent-stratified measures of evaluation (in addition to standard accuracy, preciseness, recall, F1) to obtain fairness and bias between accent groups when testing a model | ✓ | | ✓ |
| 10 | Introduction of an easy-to-use interface, with decipherable AI behaviors (attention visualizations, modality contribution explanations) to establish trust, particularly in the sensitive analysis of emotional data | | ✓ | |
| 11 | Possibility to apply accent-sensitive emotion recognition to multilingual and cross-cultural situations, including the underrepresented accents and the performance being considered as equal between the different speaker groups | ✓ | | ✓ |

*Table 14 Summary of findings*

## 4.7 Context diagram

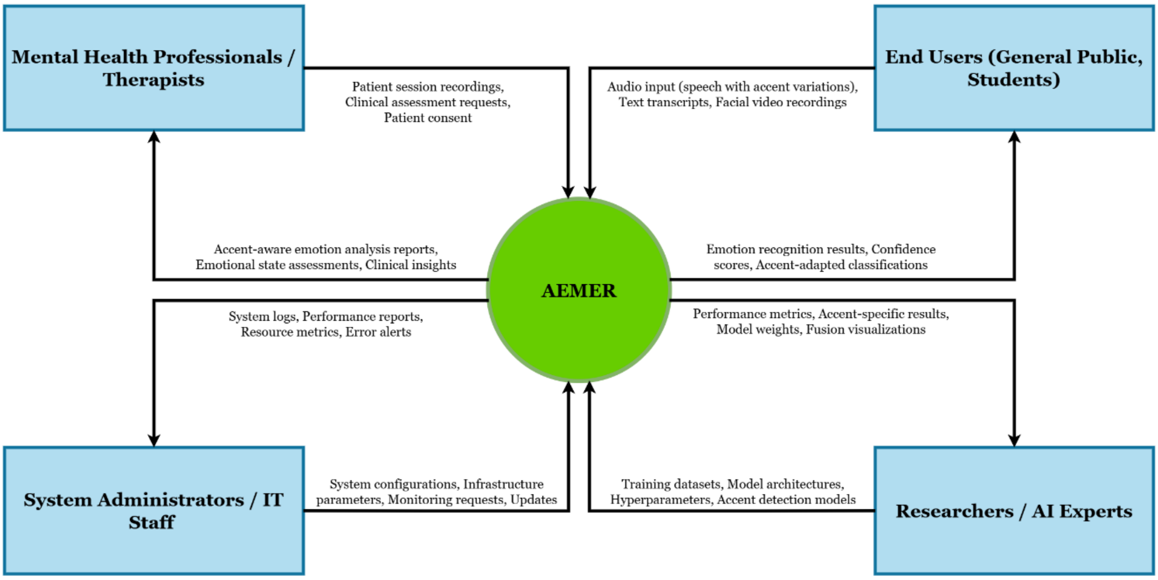

Mental Health Professionals / Therapists
Patient session recordings, Clinical assessment requests, Patient consent
Audio input (speech with accent variations), Text transcripts, Facial video recordings
End Users (General Public, Students)
Accent-aware emotion analysis reports, Emotional state assessments, Clinical insights
AEMER
Emotion recognition results, Confidence scores, Accent-adapted classifications
System logs, Performance reports, Resource metrics, Error alerts
Performance metrics, Accent-specific results, Model weights, Fusion visualizations
System Administrators / IT Staff
System configurations, Infrastructure parameters, Monitoring requests, Updates
Training datasets, Model architectures, Hyperparameters, Accent detection models
Researchers / AI Experts

*Figure 5 Context diagram (self-composed)*

## 4.8 Use case diagram

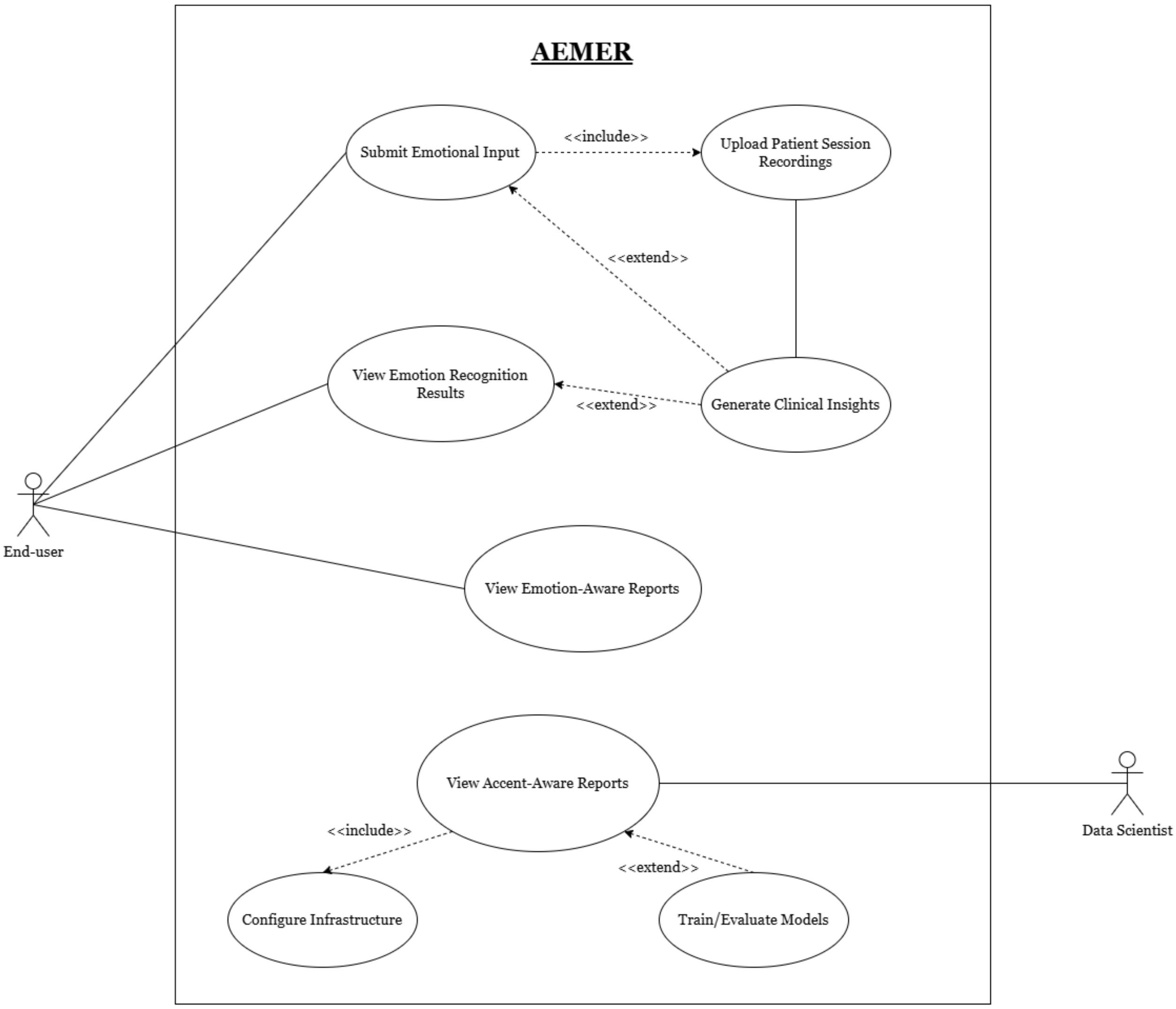


*Figure 6 Use-case diagram (Self-composed)*

## 4.9 Use case descriptions

| | | |
|---|---|---|
| Use case | Submit Emotional Input | |
| ID | UC01 | |
| Description | Submit emotional input (audio, text, video) for emotion recognition in AEMER. | |
| Actor | User | |
| Supporting Actor | None | |
| Stakeholder | Healthcare providers, therapists, educators, analysts | |
| Pre-condition | Users must log in, system is operational, valid input file(s) available | |
| Extended use-cases | Generate Clinical Insights | |
| Included use-cases | Upload Patient Session Recordings | |
| Main flow | **Actor** | **System** |
| | 1. Select input type<br>2. Upload or provide input<br>3. Proceed to results | 4. Validate input<br>5. Detect accent<br>6. Recognize emotion<br>7. Show results |
| Alternate flows | AF1: Real-time input via mic/camera.<br>AF2: Submit one modality only.<br>AF3: Upload multiple files (batch). | |
| Exceptional flows | EF1: Invalid/corrupt file—show error.<br>EF2: Accent not detected—use default.<br>EF3: Feature extraction fails—use available modality/display error. | |
| Post-condition | Processed input stored; results available to user; user notified when done. | |

*Table 15 Submit emotional input: use-case description (UC01)*

## 4.10 Requirements

There are two primary ways to communicate requirements: functional requirements and non-functional requirements.

### 4.10.1 Functional requirements

| ID | Requirements | Priority (MoSCoW) |
|---|---|---|
| FR01 | The system must detect speaker accents in real-time from audio input | Must Have |
| FR02 | The system must extract emotional features from audio, text, and visual modalities | Must Have |
| FR03 | The system must classify emotions across multiple categories (happy, sad, angry, neutral, fear, disgust, surprise) | Must Have |
| FR04 | The system must dynamically weight modality contributions based on detected accent characteristics | Must Have |
| FR05 | The system should provide confidence scores for emotion predictions | Should Have |
| FR06 | The system should support processing of multiple accent types (American, British, Canadian, South Asian) | Should Have |
| FR07 | The system should generate detailed logs of modality fusion weights for each prediction | Should Have |
| FR08 | The system could provide visualization dashboards showing modality contribution breakdowns | Could Have |
| FR09 | The system could export emotion recognition results in multiple formats (JSON, CSV, XML) | Could Have |
| FR10 | The system will not support real-time video streaming from multiple sources simultaneously | Will Not Have |

*Table 16 Functional requirements*

### 4.10.2 Non-functional requirements

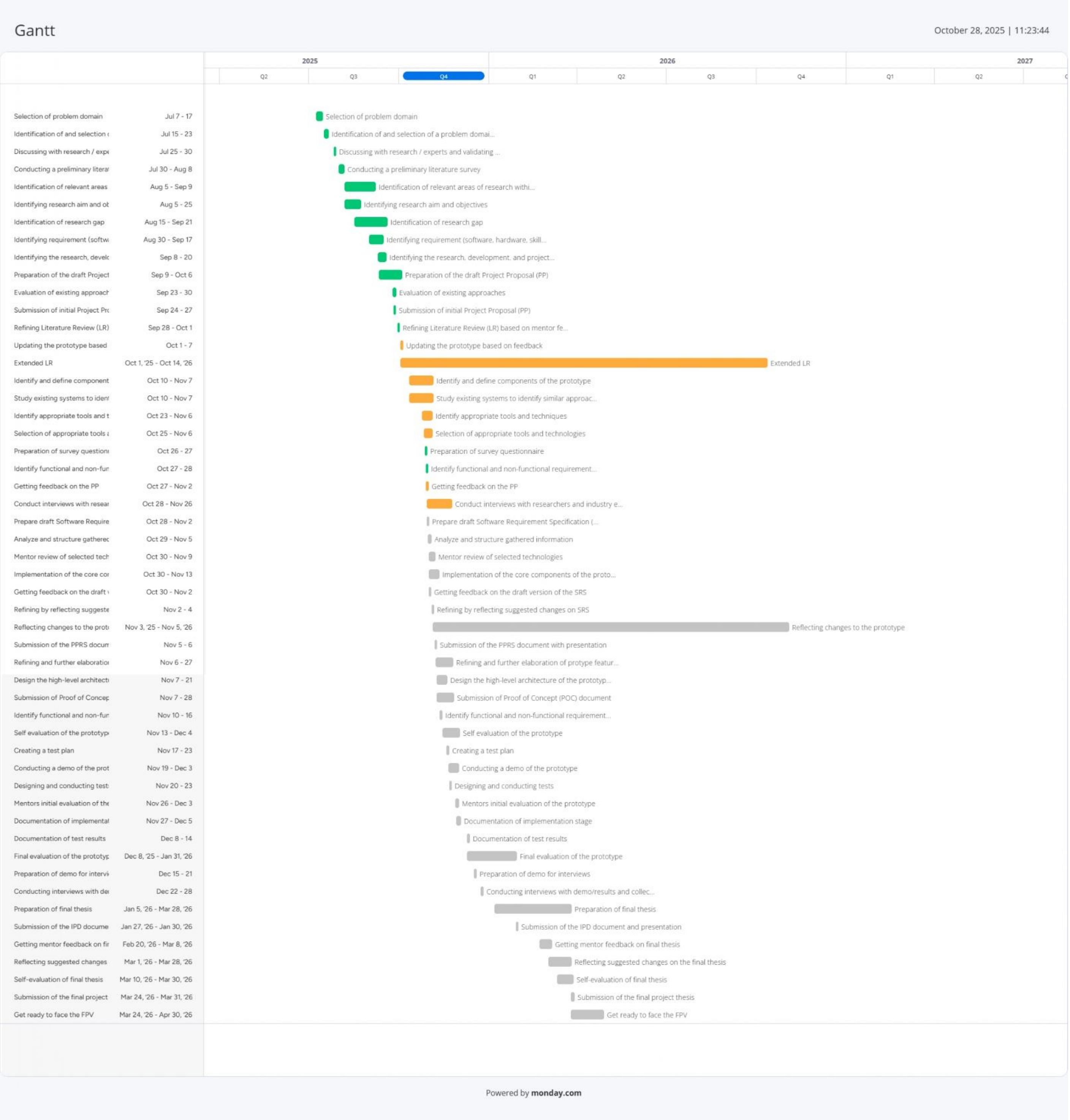
| ID | Non-Functional Requirement | Description | Priority (MoSCoW) |
|---|---|---|---|
| NFR01 | Performance | The system must maintain an average response time of less than 1 second for real-time emotion recognition | Must Have |
| NFR02 | Accuracy | The system must achieve baseline emotion recognition accuracy of at least 78% on CMU-MOSEI dataset | Must Have |
| NFR03 | Scalability | The system should handle concurrent processing of up to 10 audio-visual streams without performance degradation | Should Have |
| NFR04 | Security | The system should encrypt all processed multimodal data during transmission and storage | Could Have |
| NFR05 | Usability | The system could provide clear, interpretable emotion predictions with minimal technical expertise required | Could Have |

*Table 17 Non-functional requirements*

## 4.11 Chapter summary

This chapter described how the systematic identification of stakeholders and their connection to the AEMER system took place based on stakeholder analysis models and rich picture diagrams. Various elicitation methods that comprised literature review, domain specialists survey and interviewing, prototyping and self-assessment were used to collect the requirements. Both functional requirements like accent detection, multimodal feature extraction, and adaptive fusion as well as non-functional requirements that take care of generalizability and efficiency were found. The MoSCoW principle was used to define the priorities of requirement with the use of which the development was focused on the critical capabilities. Primary actors and interactions between system and actors were mapped using case diagrams, and this provided a complete basis regarding the accent-aware emotion recognition structure.

# 5. TIME SCHEDULE



Gantt

October 28, 2025 | 11:23:44

| | |
|---|---|
| Selection of problem domain | Jul 7 - 17 |
| Identification of and selection | Jul 15 - 23 |
| Discussing with research / expe | Jul 25 - 30 |
| Conducting a preliminary litera | Jul 30 - Aug 8 |
| Identification of relevant areas | Aug 5 - Sep 9 |
| Identifying research aim and ob | Aug 5 - 25 |
| Identification of research gap | Aug 15 - Sep 21 |
| Identifying requirement (softw | Aug 30 - Sep 17 |
| Identifying the research, develc | Sep 8 - 20 |
| Preparation of the draft Project | Sep 9 - Oct 6 |
| Evaluation of existing approach | Sep 23 - 30 |
| Submission of initial Project Pro | Sep 24 - 27 |
| Refining Literature Review (LR) | Sep 28 - Oct 1 |
| Updating the prototype based | Oct 1 - 7 |
| Extended LR | Oct 1, '25 - Oct 14, '26 |
| Identify and define component | Oct 10 - Nov 7 |
| Study existing systems to ident | Oct 10 - Nov 7 |
| Identify appropriate tools and t | Oct 23 - Nov 6 |
| Selection of appropriate tools a | Oct 25 - Nov 6 |
| Preparation of survey question | Oct 26 - 27 |
| Identify functional and non-fur | Oct 27 - 28 |
| Getting feedback on the PP | Oct 27 - Nov 2 |
| Conduct interviews with resear | Oct 28 - Nov 26 |
| Prepare draft Software Require | Oct 28 - Nov 2 |
| Analyze and structure gathered | Oct 29 - Nov 5 |
| Mentor review of selected tech | Oct 30 - Nov 9 |
| Implementation of the core cor | Oct 30 - Nov 13 |
| Getting feedback on the draft | Oct 30 - Nov 2 |
| Refining by reflecting suggeste | Nov 2 - 4 |
| Reflecting changes to the prot | Nov 3, '25 - Nov 5, '26 |
| Submission of the PPRS docum | Nov 5 - 6 |
| Refining and further elaboratio | Nov 6 - 27 |
| Design the high-level architect | Nov 7 - 21 |
| Submission of Proof of Concep | Nov 7 - 28 |
| Identify functional and non-fur | Nov 10 - 16 |
| Self evaluation of the prototyp | Nov 13 - Dec 4 |
| Creating a test plan | Nov 17 - 23 |
| Conducting a demo of the prot | Nov 19 - Dec 3 |
| Designing and conducting test | Nov 20 - 23 |
| Mentors initial evaluation of the | Nov 26 - Dec 3 |
| Documentation of implementat | Nov 27 - Dec 5 |
| Documentation of test results | Dec 8 - 14 |
| Final evaluation of the prototyp | Dec 8, '25 - Jan 31, '26 |
| Preparation of demo for intervi | Dec 15 - 21 |
| Conducting interviews with de | Dec 22 - 28 |
| Preparation of final thesis | Jan 5, '26 - Mar 28, '26 |
| Submission of the IPD docume | Jan 27, '26 - Jan 30, '26 |
| Getting mentor feedback on fir | Feb 20, '26 - Mar 8, '26 |
| Reflecting suggested changes | Mar 1, '26 - Mar 28, '26 |
| Self-evaluation of final thesis | Mar 10, '26 - Mar 30, '26 |
| Submission of the final project | Mar 24, '26 - Mar 31, '26 |
| Get ready to face the FPV | Mar 24, '26 - Apr 30, '26 |

2025
2026
2027
Q2
Q3
Q4
Q1
Q2
Q3
Q4
Q1
Q2
Selection of problem domain
Identification of and selection of a problem domai...
Discussing with research / experts and validating ...
Conducting a preliminary literature survey
Identification of relevant areas of research withi...
Identifying research aim and objectives
Identification of research gap
Identifying requirement (software, hardware, skill...
Identifying the research, development, and project...
Preparation of the draft Project Proposal (PP)
Evaluation of existing approaches
Submission of initial Project Proposal (PP)
Refining Literature Review (LR) based on mentor fe...
Updating the prototype based on feedback
Extended LR
Identify and define components of the prototype
Study existing systems to identify similar approac...
Identify appropriate tools and techniques
Selection of appropriate tools and technologies
Preparation of survey questionnaire
Identify functional and non-functional requirement...
Getting feedback on the PP
Conduct interviews with researchers and industry e...
Prepare draft Software Requirement Specification (...
Analyze and structure gathered information
Mentor review of selected technologies
Implementation of the core components of the proto...
Getting feedback on the draft version of the SRS
Refining by reflecting suggested changes on SRS
Reflecting changes to the prototype
Submission of the PPRS document with presentation
Refining and further elaboration of protype featur...
Design the high-level architecture of the prototyp...
Submission of Proof of Concept (POC) document
Identify functional and non-functional requirement...
Self evaluation of the prototype
Creating a test plan
Conducting a demo of the prototype
Designing and conducting tests
Mentors initial evaluation of the prototype
Documentation of implementation stage
Documentation of test results
Final evaluation of the prototype
Preparation of demo for interviews
Conducting interviews with demo/results and collec...
Preparation of final thesis
Submission of the IPD document and presentation
Getting mentor feedback on final thesis
Reflecting suggested changes on the final thesis
Self-evaluation of final thesis
Submission of the final project thesis
Get ready to face the FPV

Powered by **monday.com**

*Figure 7 Gantt chart*

# REFERENCES

Jia, Muhan, and Zijian Sun. “A Survey of Multi-Modal Emotion Recognition Based on Deep Learning.” *Highlights in Science, Engineering and Technology* 119 (December 2024): 533–40. https://doi.org/10.54097/37zncv36.

Mollahosseini, Ali, Behzad Hasani, and Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild.” *IEEE Transactions on Affective Computing* 10, no. 1 (2019): 18–31. https://doi.org/10.1109/TAFFC.2017.2740923.

Fard, Ali Pourramezan, Mohammad Mehdi Hosseini, Timothy D. Sweeny, and Mohammad H. Mahoor. “AffectNet+: A Database for Enhancing Facial Expression Recognition with Soft-Labels.” arXiv:2410.22506. Preprint, arXiv, October 29, 2024. https://doi.org/10.48550/arXiv.2410.22506.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. n.d.

Van Genugten, Claire R., Melissa S. Y. Thong, Wouter Van Ballegooijen, et al. “Beyond the Current State of Just-in-Time Adaptive Interventions in Mental Health: A Qualitative Systematic Review.” *Frontiers in Digital Health* 7 (January 2025): 1460167. https://doi.org/10.3389/fdgth.2025.1460167.

Liang, Paul Pu, Ruslan Salakhutdinov, and Louis-Philippe Morency. *Computational Modeling of Human Multimodal Language: The MOSEI Dataset and Interpretable Dynamic Fusion*. n.d.

Antoniou, Nikolaos, Athanasios Katsamanis, Theodoros Giannakopoulos, and Shrikanth Narayanan. “Designing and Evaluating Speech Emotion Recognition Systems: A Reality Check Case Study with IEMOCAP.” *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, June 4, 2023, 1–5. https://doi.org/10.1109/ICASSP49357.2023.10096808.

Lerner, Jennifer S., Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. “Emotion and Decision Making.” *Annual Review of Psychology* 66, no. 1 (2015): 799–823. https://doi.org/10.1146/annurev-psych-010213-115043.

Lim, Tristan. “Emotion-Aware Decision Support System for Real-Time Financial Sentiment and Behavior-Based Trading Risk Advisory.” Preprint, SSRN, 2025. https://doi.org/10.2139/ssrn.5183852.

Tabassum, Nowshin, Tasfia Tabassum, Fardin Saad, Tahiya Sultana Safa, Hasan Mahmud, and Md. Kamrul Hasan. “Exploring the English Accent-Independent Features for Speech Emotion Recognition Using Filter and Wrapper-Based Methods for Feature Selection.” *INTERSPEECH 2023*, ISCA, August 20, 2023, 3217–21. https://doi.org/10.21437/Interspeech.2023-1888.

Henry, Lauren M, Morkeh Blay-Tofey, Clara E Haeffner, et al. “Just-In-Time Adaptive Interventions to Promote Behavioral Health: Protocol for a Systematic Review.” *JMIR Research Protocols* 14 (February 2025): e58917. https://doi.org/10.2196/58917.

Tripathi, Samarth, Sarthak Tripathi, and Homayoon Beigi. “Multi-Modal Emotion Recognition on IEMOCAP Dataset Using Deep Learning.” arXiv:1804.05788. Preprint, arXiv, November 6, 2019. https://doi.org/10.48550/arXiv.1804.05788.

Wu, Chengyan, Yiqiang Cai, Yang Liu, et al. “Multimodal Emotion Recognition in Conversations: A Survey of Methods, Trends, Challenges and Prospects.” arXiv:2505.20511. Preprint, arXiv, September 9, 2025. https://doi.org/10.48550/arXiv.2505.20511.

Ramaswamy, Manju Priya Arthanarisamy, and Suja Palaniswamy. “Multimodal Emotion Recognition: A Comprehensive Review, Trends, and Challenges.” *WIREs Data Mining and Knowledge Discovery* 14, no. 6 (2024): e1563. https://doi.org/10.1002/widm.1563.

Bagher Zadeh, AmirAli, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. “Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph.” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, 2018, 2236–46. https://doi.org/10.18653/v1/P18-1208.

Levine, Daniel. “Neuroscience of Emotion, Cognition, and Decision Making: A Review.” *Medical Research Archives* 10, no. 7 (2022). https://doi.org/10.18103/mra.v10i7.2869.

Sun, Yifan, Tian Lu, Xuanyi Wang, et al. “Physiological Feedback Technology for Real-Time Emotion Regulation: A Systematic Review.” *Frontiers in Psychology* 14 (May 2023): 1182667. https://doi.org/10.3389/fpsyg.2023.1182667.

Jerčić, Petar, and Veronica Sundstedt. “Practicing Emotion-Regulation through Biofeedback on the Decision-Making Performance in the Context of Serious Games: A Systematic Review.” *Entertainment Computing* 29 (March 2019): 75–86. https://doi.org/10.1016/j.entcom.2019.01.001.

Lima, Rodrigo, Alice Chirico, Andrea Gaggioli, Hugo Gamboa, and Sergi Bermúdez I Badia. “Real-Time Emotion Regulation in Virtual Reality: An Adaptive Experience Using Breathing Biofeedback.” Preprint, In Review, January 16, 2025. https://doi.org/10.21203/rs.3.rs-5804829/v1.

Kotta, Kundan Sai, Sai Nikhil Samineni, and Asst G Kavitha. “Speech Emotion through Voice & Accent.” *International Journal of Scientific Research* 9, no. 6 (2023).

Aruna Gladys, A., and V. Vetriselvi. “Survey on Multimodal Approaches to Emotion Recognition.” *Neurocomputing* 556 (November 2023): 126693. https://doi.org/10.1016/j.neucom.2023.126693.

Livingstone, Steven R., and Frank A. Russo. “The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A Dynamic, Multimodal Set of Facial and Vocal Expressions in North American English.” *PLOS ONE* 13, no. 5 (2018): e0196391. https://doi.org/10.1371/journal.pone.0196391.

# APPENDIX A

| | | |
|---|---|---|
| Use case | View Emotion Recognition Results | |
| ID | UC02 | |
| Description | View emotion recognition results with predicted emotions, confidence scores, and accent information. | |
| Actor | User | |
| Supporting Actor | None | |
| Stakeholder | Healthcare providers, therapists, analysts | |
| Pre-condition | Submit Emotional Input use case completed, results stored in system | |
| Extended use-cases | 1. Generate Clinical Insights.<br>2. View Emotion-Aware Reports. | |
| Included use-cases | 1. Download recognition results.<br>2. View modality contribution weights. | |
| Main flow | **Actor** | **System** |
| | 1. Request results view.<br>2. Select result to view.<br>3. Download/export results. | 4. Fetch stored results.<br>5. Display predictions with scores.<br>6. Show accent & modality info.<br>7. Enable export options. |
| Alternate flows | AF1: Filter results by date range.<br>AF2: Sort by emotion or confidence.<br>AF3: View results in table/chart format. | |
| Exceptional flows | EF1: Results not found—display error.<br>EF2: Data retrieval fails—show retry option. | |
| Post-condition | Results displayed to user; export files generated; viewing logged in system. | |

*Table 18 View Emotion Recognition Results: use-case description (UC02)*

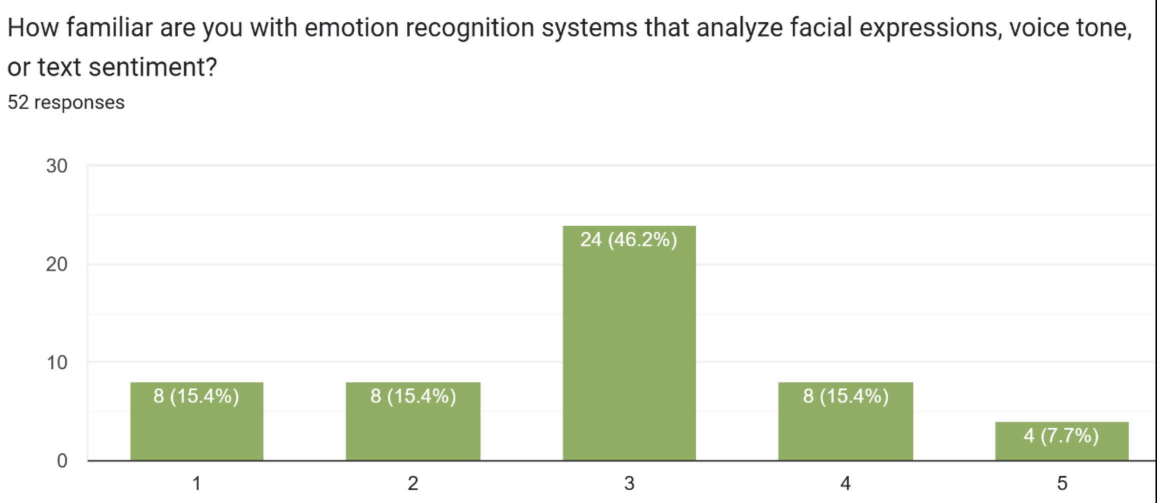
<table>
<tr><td>Use case</td><td colspan="2">View Emotion-Aware Reports</td></tr>
<tr><td>ID</td><td colspan="2">UC03</td></tr>
<tr><td>Description</td><td colspan="2">Access and view comprehensive emotion reports with trends, statistics, and visualizations.</td></tr>
<tr><td>Actor</td><td colspan="2">User</td></tr>
<tr><td>Supporting Actor</td><td colspan="2">None</td></tr>
<tr><td>Stakeholder</td><td colspan="2">Healthcare providers, educators, analysts, researchers</td></tr>
<tr><td>Pre-condition</td><td colspan="2">Clinical insights generated, reports stored in database</td></tr>
<tr><td>Extended use-cases</td><td colspan="2">1. View Accent-Aware Reports.<br>2. Export reports.</td></tr>
<tr><td>Included use-cases</td><td colspan="2">None</td></tr>
<tr><td rowspan="2">Main flow</td><td><b>Actor</b></td><td><b>System</b></td></tr>
<tr><td>1. Select report type.<br>2. Choose date range.<br>3. View report.</td><td>4. Retrieve report data.<br>5. Render visualizations.<br>6. Display metrics &amp; charts.<br>7. Enable export.</td></tr>
<tr><td>Alternate flows</td><td colspan="2">AF1: Generate custom reports.<br>AF2: Schedule automated reports.<br>AF3: Share reports with colleagues.</td></tr>
<tr><td>Exceptional flows</td><td colspan="2">EF1: Report generation failed—display error.<br>EF2: Missing data—show available alternatives.</td></tr>
<tr><td>Post-condition</td><td colspan="2">Reports displayed/exported; access logged in system.</td></tr>
</table>

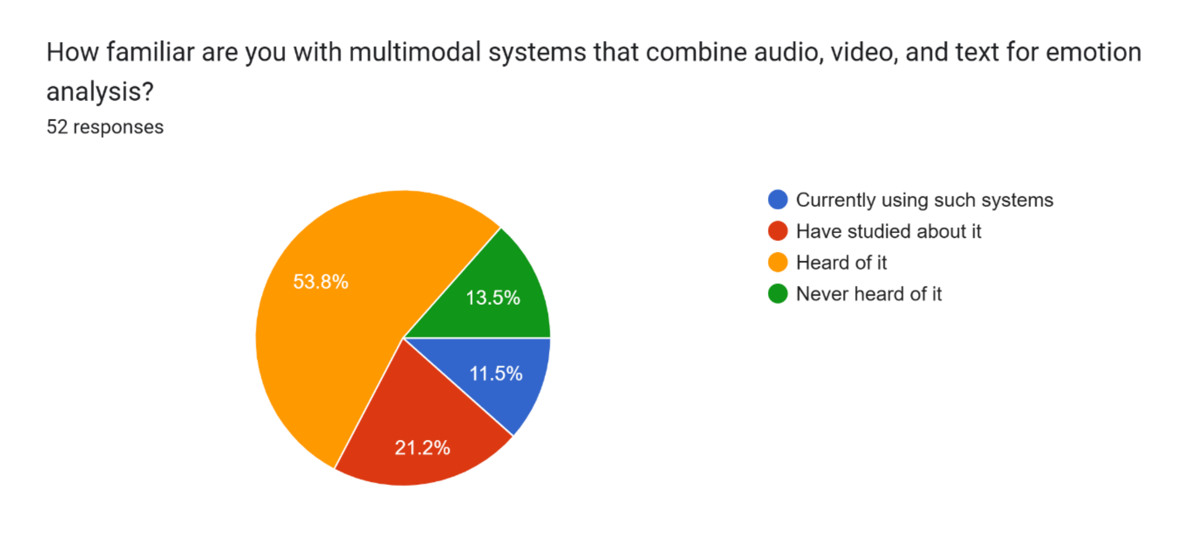
*Table 19 View Emotion-Aware Reports: use-case description (UC03)*

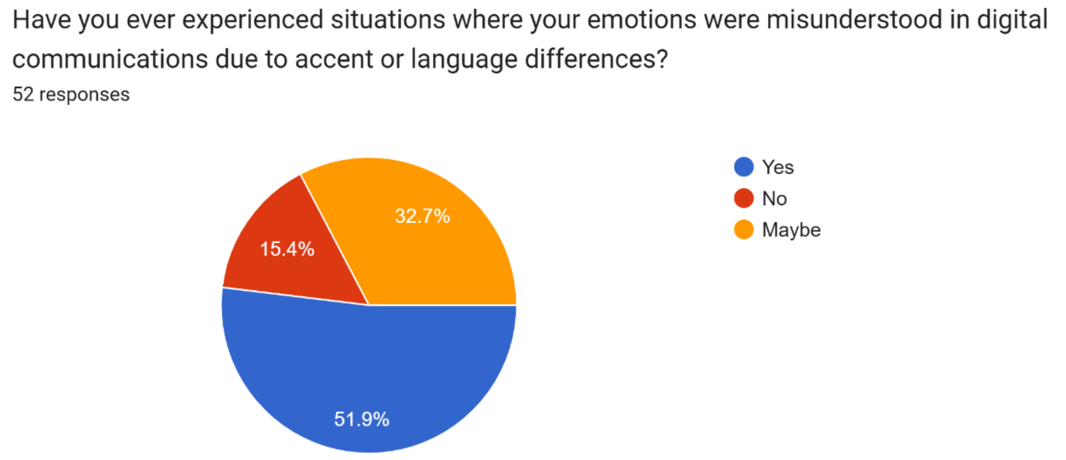
| | | |
|---|---|---|
| Use case | View Accent-Aware Reports | |
| ID | UC04 | |
| Description | View reports analyzing emotion recognition performance across different accent groups. | |
| Actor | Data Scientist | |
| Supporting Actor | System Administrator | |
| Stakeholder | Researchers, ML engineers, quality assurance teams | |
| Pre-condition | Emotion recognition results with accent labels, system analyzed | |
| Extended use-cases | 1. Train/Evaluate Models. | |
| Included use-cases | 1. Configure Infrastructure.<br>2. Analyze accent metrics. | |
| Main flow | **Actor** | **System** |
| | 1. Request accent report.<br>2. Select accent groups.<br>3. Review analysis. | 4. Filter by accent.<br>5. Calculate accent-specific metrics.<br>6. Compare across accents.<br>7. Display accent report. |
| Alternate flows | AF1: Analyze individual accent subgroups.<br>AF2: Generate accent-invariant performance metrics.<br>AF3: Compare model versions. | |
| Exceptional flows | EF1: Insufficient accent diversity—show warning.<br>EF2: Metrics calculation error—retry. | |
| Post-condition | Reports generated; accent performance analyzed; data saved for model improvement. | |

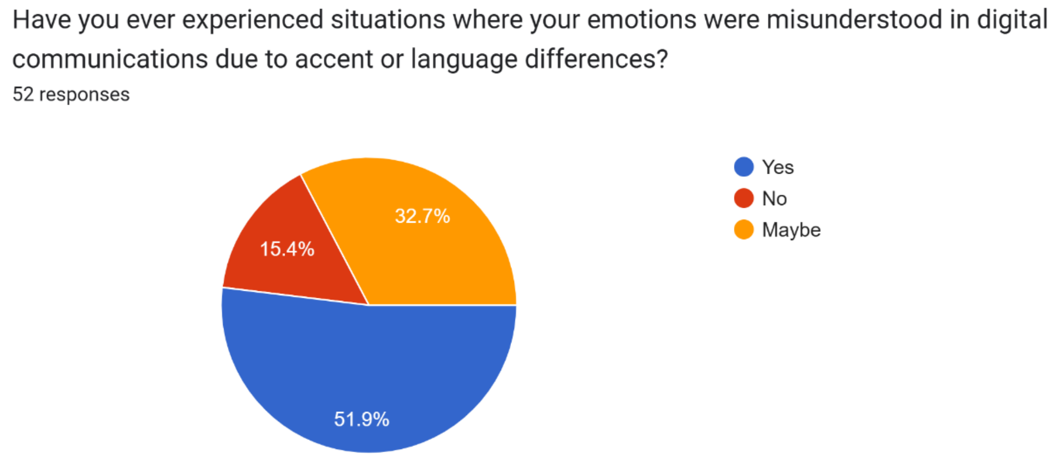
*Table 20 View Accent-Aware Reports: use-case description (UC04)*

# APPENDIX B

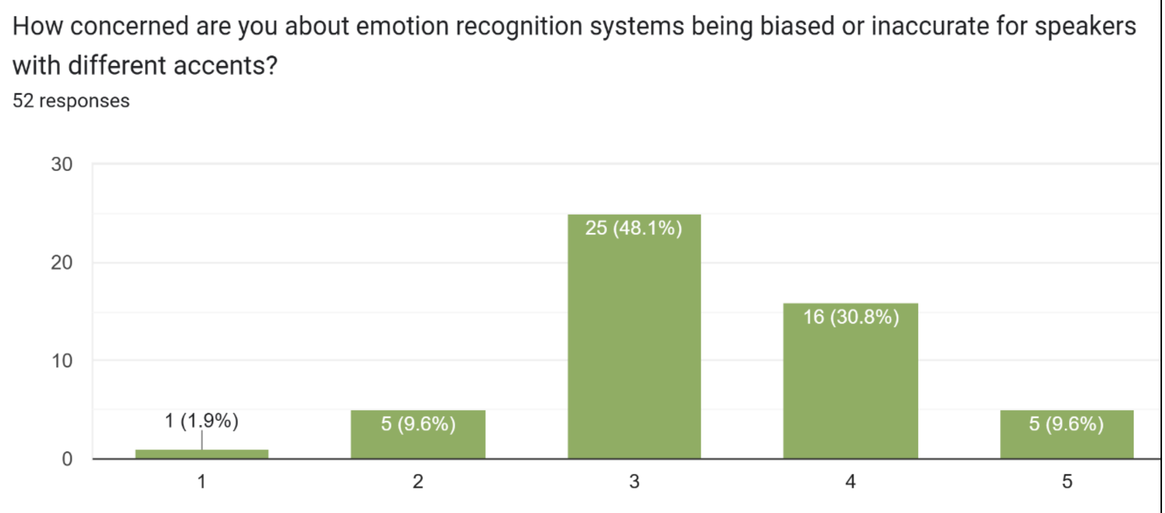
<table>
<tr><td><b>Question</b></td><td>How familiar are you with emotion recognition systems that analyze facial expressions, voice tone, or text sentiment?</td></tr>
<tr><td><b>Aim of question</b></td><td>To determine the level of knowledge of the respondents on the technology of emotion recognition that assists in customizing the interface and complexity of documentation of the system.</td></tr>
<tr><td colspan="2">
How familiar are you with emotion recognition systems that analyze facial expressions, voice tone, or text sentiment?<br>52 responses
</td></tr>
<tr><td><b>Finding &amp; Conclusion</b></td><td>The mean familiarity score (2.69/5.0) and concentration level (44.2) that is moderate with the peak at level 3 (28.2) reveals that most users are aware of emotion recognition systems on a basic level with little deep knowledge. This observation implies that a design should be more intuitive, easy to use and provide onboarding resources and proper instructions of how the system works to ensure that users with different levels of technical expertise can get used to it.</td></tr>
</table>

| | |
|---|---|
| **Question** | How familiar are you with multimodal systems that combine audio, video, and text for emotion analysis? |
| **Aim of question** | It is important to identify the necessity of the components of the system to be educated in multimodal approaches to understand their perception of this system by users. |
|  How familiar are you with multimodal systems that combine audio, video, and text for emotion analysis?<br>52 responses<br> | |
| **Finding & Conclusion** | Although 57.7 percent of those who participated in the survey have heard about multimodal systems, only 30.8 percent (16 respondents) of them have some experience in terms of studying or using it. This knowledge gap focuses on the fact that it is essential to give strict explanations of the multimodal fusion approach of AEMER and its advantages so that the users should know how the combination of audio, visual, and textual cues leads to the accuracy of emotion recognition. |

<table>
<tr><td><b>Question</b></td><td>Have you ever experienced situations where your emotions were misunderstood in digital communications due to accent or language differences?</td></tr>
<tr><td><b>Aim of question</b></td><td>To confirm the central problem statement of AEMER through the measurement of the actual experiences of an emotional misunderstanding that happens because of accent.</td></tr>
<tr><td colspan="2"><br>Have you ever experienced situations where your emotions were misunderstood in digital communications due to accent or language differences?<br>52 responses<br> </td></tr>
<tr><td><b>Finding &amp; Conclusion</b></td><td>Quite important: 51.9 percent of the respondents stated that they encountered emotional misunderstandings concerning accents, and 28.8 percent stated that they may have encountered them. Such a high rate of 80.7 percent is a strong indication of the underlying driving force behind the accent-conscious architecture of AEMER. The results affirm that the variability of various accents has a significant effect on the effectiveness of emotional communication during online communication and that there is a dire necessity of developing specific solutions that will be able to detect various emotions accurately given the diversity of linguistic backgrounds.</td></tr>
</table>

| | |
|---|---|
| **Question** | Have you ever experienced situations where your emotions were misunderstood in digital communications due to accent or language differences? |
| **Aim of question** | To confirm the central problem statement of AEMER through the measurement of the actual experiences of an emotional misunderstanding that happens because of accent. |
| <br>Have you ever experienced situations where your emotions were misunderstood in digital communications due to accent or language differences?<br>52 responses<br> | |
| **Finding & Conclusion** | Quite important: 51.9 percent of the respondents stated that they encountered emotional misunderstandings concerning accents, and 28.8 percent stated that they may have encountered them. Such a high rate of 80.7 percent is a strong indication of the underlying driving force behind the accent-conscious architecture of AEMER. The results affirm that the variability of various accents has a significant effect on the effectiveness of emotional communication during online communication and that there is a dire necessity of developing specific solutions that will be able to detect various emotions accurately given the diversity of linguistic backgrounds. |

| | |
|---|---|
| **Question** | How concerned are you about emotion recognition systems being biased or inaccurate for speakers with different accents? |
| **Aim of question** | To assess user consciousness and anxiety about possible bias in emotion recognition systems by the algorithm. |



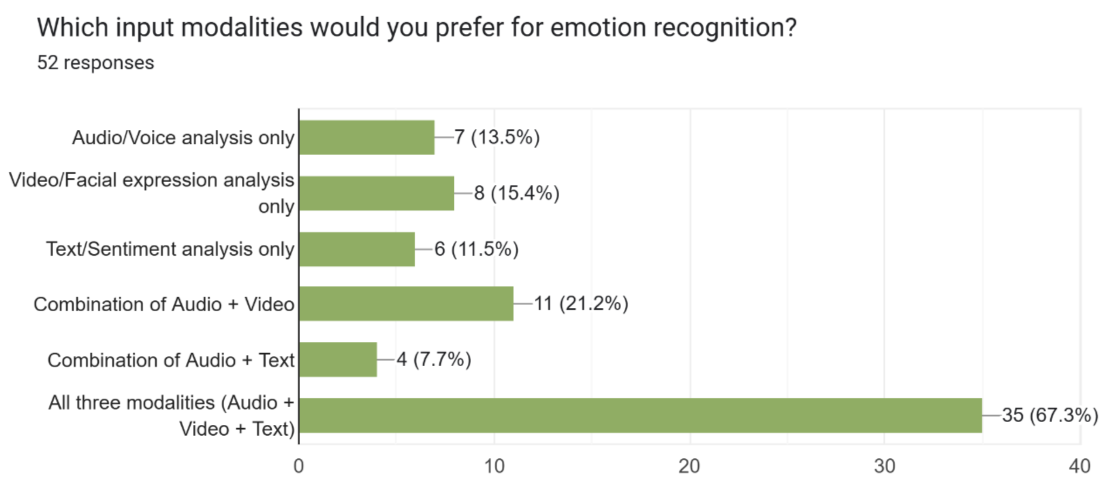
How concerned are you about emotion recognition systems being biased or inaccurate for speakers with different accents?

52 responses

| | |
|---|---|
| **Finding & Conclusion** | The mean score of 3.31/5.0 with 82.7% scoring 3 or higher in the moderate-to-high level of concern indicates that the users are quite aware that there are problems with emotion recognition systems that may be biased. This observation allows supporting the significance of the accent-conscious design of AEMER and justifies the necessity of clear bias reduction practices, mixed training data, and comprehensible artificial intelligence mechanisms to gain user trust. |

| | |
|---|---|
| **Question** | Which input modalities would you prefer for emotion recognition? |

| | |
|---|---|
| **Aim of question** | To determine the preferences of the users on the type of data to be inputted in the system, to guide the multimodal architecture design of the system. |

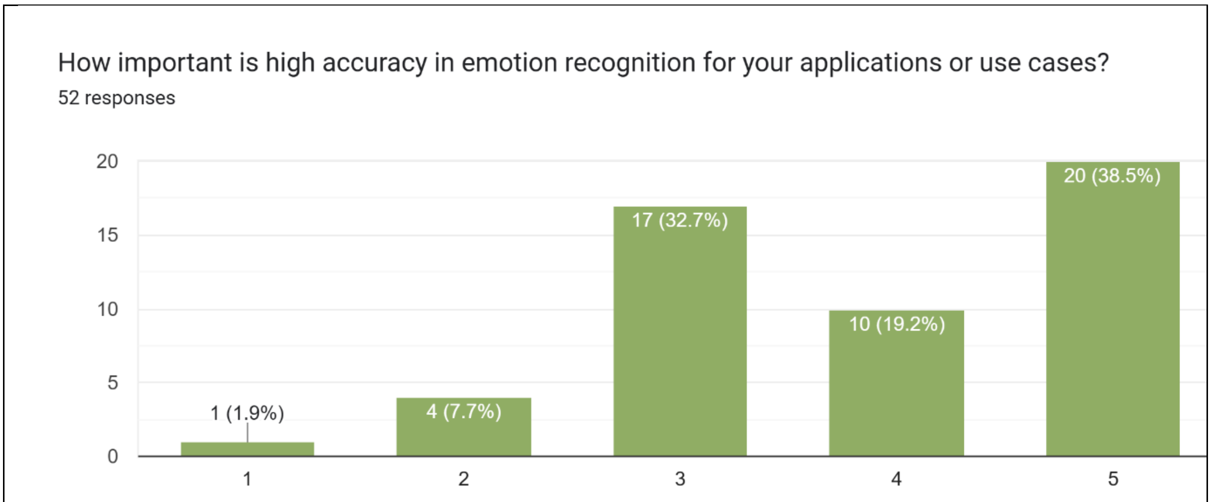


Which input modalities would you prefer for emotion recognition?

52 responses

| | |
|---|---|
| **Finding & Conclusion** | The overwhelmingly high 69.2% preference of all three modalities confirms multimodal fusion used by AEMER. It is evident that users appreciate the importance of using a combination of various sources of information in order to achieve strong emotion recognition. The high tendency of multimodal integration affirms that users want to make use of holistic emotion analysis that makes use of complementary signals through voice, face expressions as well as textual material in tandem. |

| | |
|---|---|
| **Question** | How important is high accuracy in emotion recognition for your applications or use cases? |
| **Aim of question** | To measure the importance of recognition accuracy as compared with other system characteristics in the priorities of the users. |

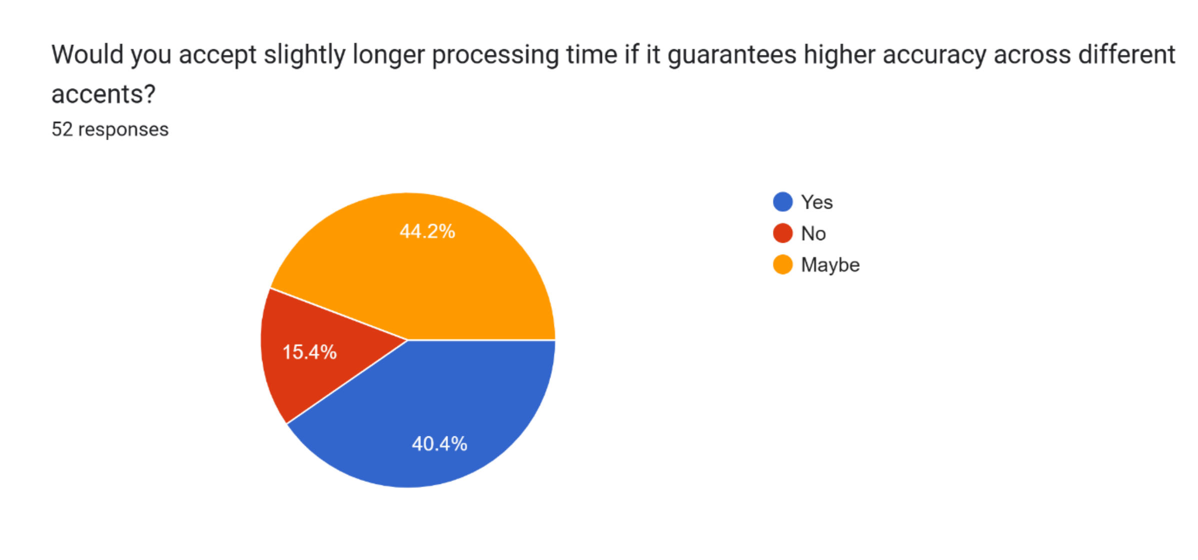


How important is high accuracy in emotion recognition for your applications or use cases?

52 responses

| | |
|---|---|
| **Finding & Conclusion** | High-perceived accuracy is also evidently in the first place, as 51.9% of them considered it very important (scores 4-5) and the mean of 3.77/5.0. This observation explains why AEMER has focused on complex attention systems, dynamic fusion approaches, and accent conscious adaptation to give maximum recognition accuracy. Users require emotion detection that is reliable and especially on professional applications where it may have serious ramification. |

| | |
|---|---|
| **Question** | Would you accept slightly longer processing time if it guarantees higher accuracy across different accents? |
| **Aim of question** | To determine whether the users were ready to sacrifice processing speed to achieve cross-accent accuracy. |

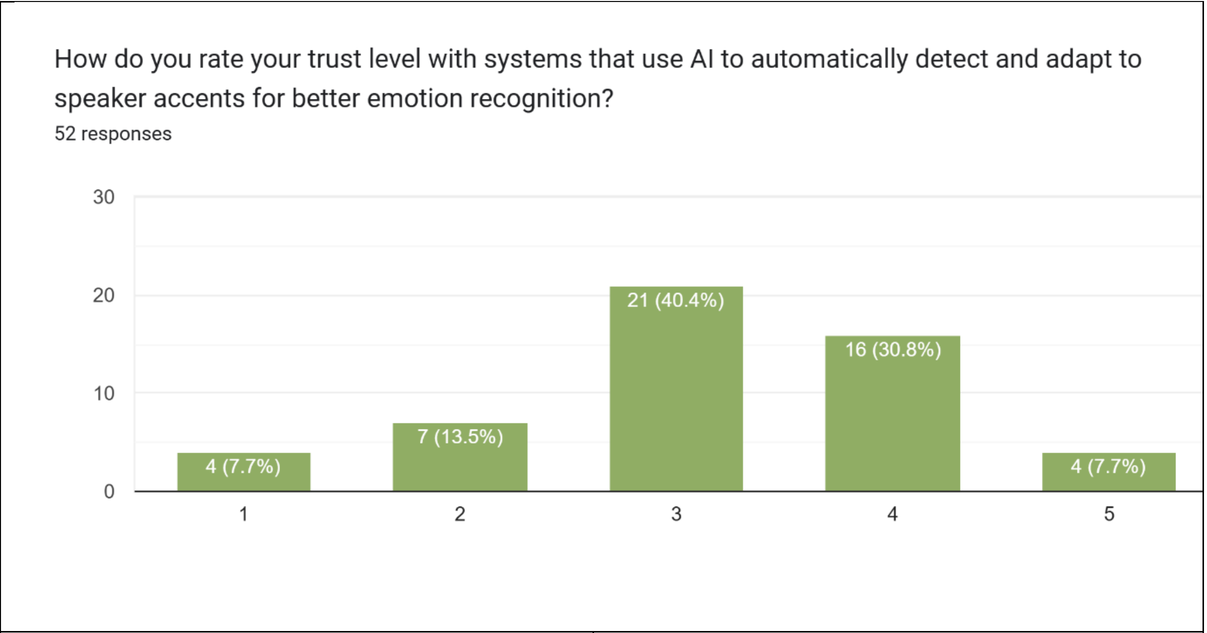


Would you accept slightly longer processing time if it guarantees higher accuracy across different accents?

52 responses

| | |
|---|---|
| **Finding & Conclusion** | There is high acceptance among users to accept longer processing times (36.5%) as well as conditionally open (Maybe) which means that the majority of users accept slower and more accurate processing. This observation justifies AEMER to employ such computationally intensive aspects as dynamic attention networks and accent-detecting machines. The 50 percent Maybe answer however indicates that optimization should be an option in order to make a trade-off between accuracy and moderate latency. |

| | |
|---|---|
| **Question** | How do you rate your trust level with systems that use AI to automatically detect and adapt to speaker accents for better emotion recognition? |
| **Aim of question** | To understand the user confidence in the AI-powered accent adaptation systems. |

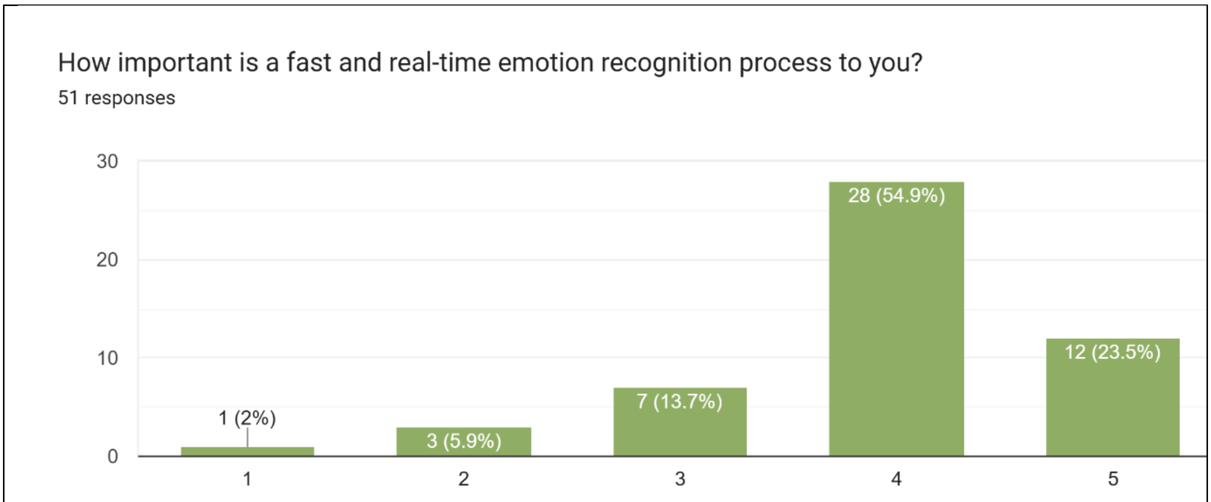


How do you rate your trust level with systems that use AI to automatically detect and adapt to speaker accents for better emotion recognition?

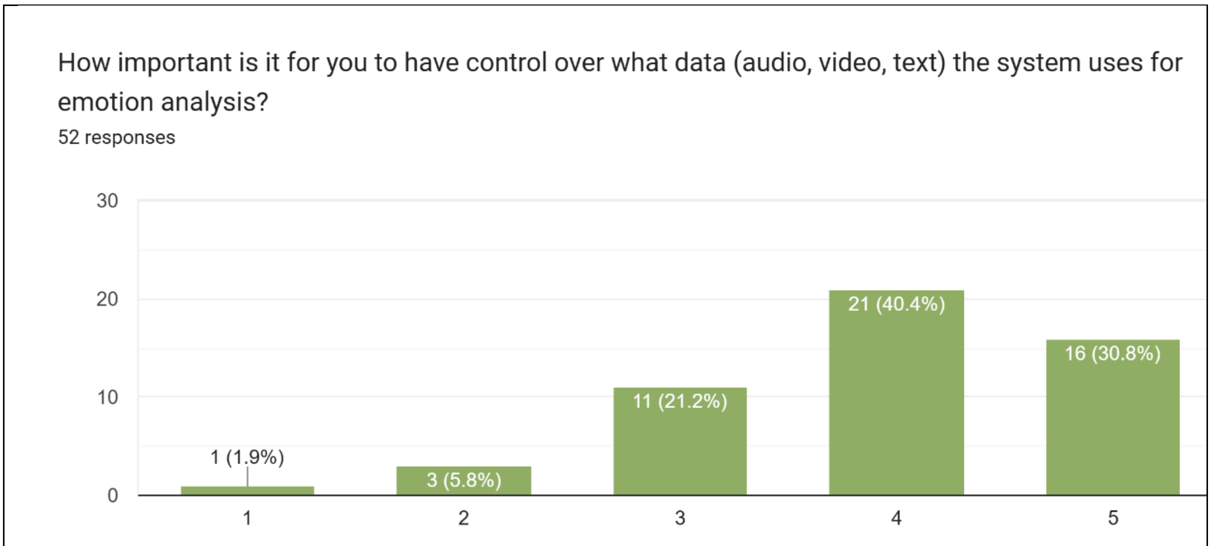
52 responses

| | |
|---|---|
| **Finding & Conclusion** | Intermediate levels of trust (mean 3.29/5.0) and 46.2% in the neutral category show that there is moderate doubt of AI accent adaptation. Although 38.4 percent has positive trust (score 4-5), the large neutral response indicated that users require confirmation of system reliability by transparent functionality, confidence ratings, and explainable decision-making to develop confidence in the accent-aware solutions by AEMER. |

| | |
|---|---|
| **Question** | How important is a fast and real-time emotion recognition process to you? |
| **Aim of question** | To establish the priority of real-time processing capabilities of user applications. |

| | |
|---|---|
|  | |
| **Finding & Conclusion** | They put a lot of importance on real-time processing where 69.2% consider it important (4-5) with average of 3.98/5.0. The conclusion here implies that inference pipelines, model design, and possible edge deployment strategies should be optimized to make sure that AEMER can be deployed with low latency to serve applications such as video conferencing, live customer service, and live emotional feedback systems. |

| | |
|---|---|
| **Question** | How important is it for you to have control over what data (audio, video, text) the system uses for emotion analysis? |
| **Aim of question** | To determine how important data privacy and control mechanisms are to the user. |

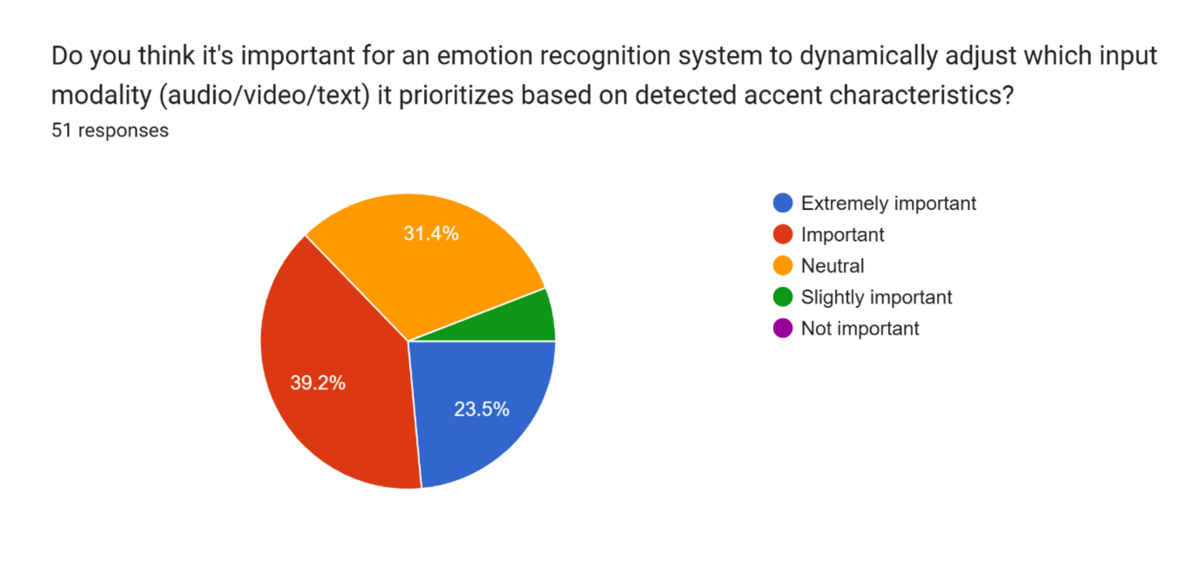


How important is it for you to have control over what data (audio, video, text) the system uses for emotion analysis?

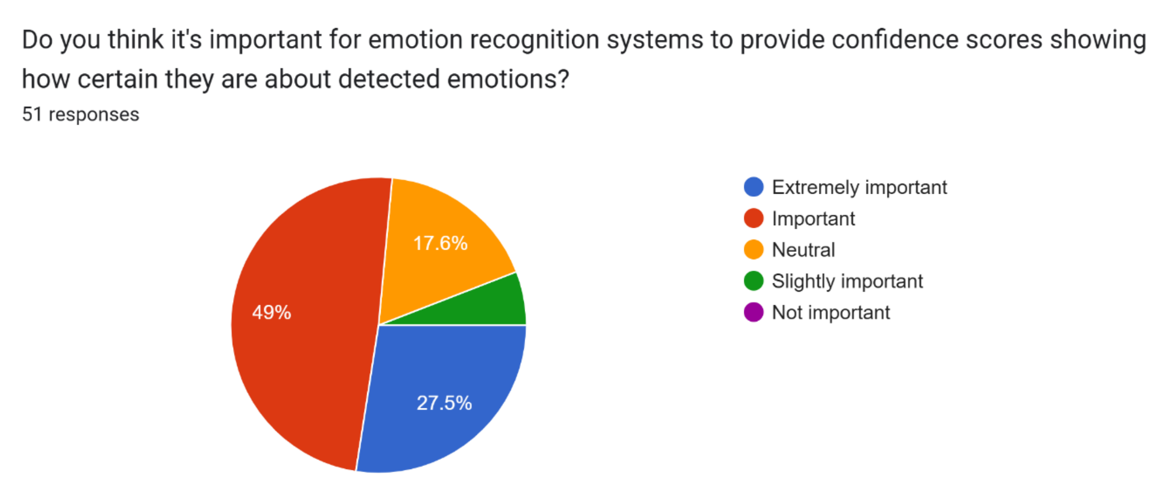
52 responses

| | |
|---|---|
| **Finding & Conclusion** | The issue of data control is a significant one, and 73.1% (scores 4-5) rated it important, and the mean of 3.96/5.0 is high. This observation prompts a need to establish finer privacy controls that enable customers to enable/disable certain modalities, a publicly accessible policy of data use, and explicit consent procedures. To win the trust and be ethical, AEMER should offer users agency of their emotional data. |

| | |
|---|---|
| **Question** | Do you think it's important for an emotion recognition system to dynamically adjust which input modality (audio/video/text) it prioritizes based on detected accent characteristics? |
| **Aim of question** | To analyze how the core innovation of accent-aware dynamic fusion by AEMER is appreciated by the users. |

<table>
<tr><td colspan="2">Do you think it's important for an emotion recognition system to dynamically adjust which input modality (audio/video/text) it prioritizes based on detected accent characteristics?<br>51 responses </td></tr>
<tr><td><b>Finding &amp; Conclusion</b></td><td>There is a high level of support of dynamic adjustment, where 69.2 percent of the respondents rated it as important or extremely important. This confirmation of the dynamic attention-based fusion mechanism of AEMER discloses the fact that users are aware of the worth of adaptive systems that weigh modalities by the aspects of the context such as accent features. This feature considers user requirements of high-context emotion recognition.</td></tr>
</table>

| | |
|---|---|
| **Question** | Do you think it's important for emotion recognition systems to provide confidence scores showing how certain they are about detected emotions? |
| **Aim of question** | To evaluate the need to have transparency and explainability features. |

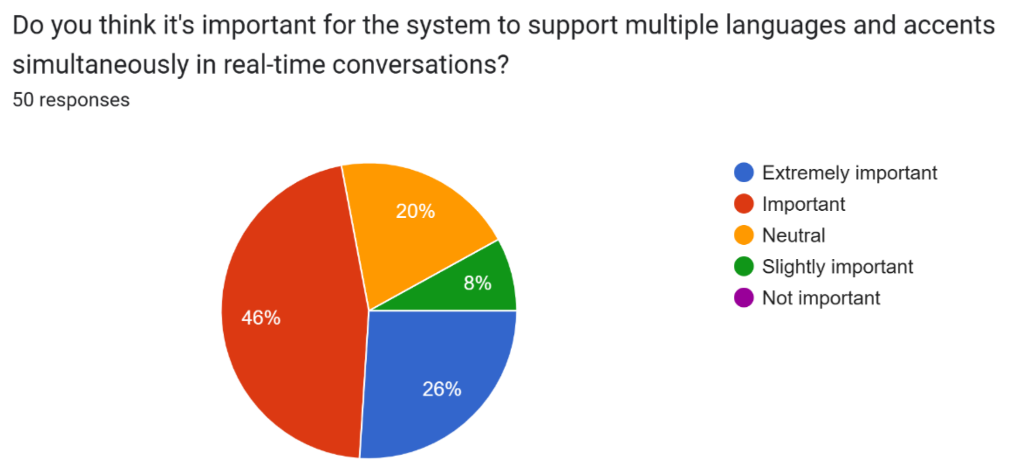


Do you think it's important for emotion recognition systems to provide confidence scores showing how certain they are about detected emotions?

51 responses

| | |
|---|---|
| **Finding & Conclusion** | The need to have transparency mechanisms is evident as 73.1% regard the confidence scores as important or very important. This result suggests the use of probabilistic outputs, quantification of uncertainty, and explicit confidence measures in the interface of AEMER. Confidence scores allow the user to understand system outputs in a proper manner and make effective decisions especially when it comes to high stakes applications. |

| | |
|---|---|
| **Question** | Do you think it's important for the system to support multiple languages and accents simultaneously in real-time conversations? |
| **Aim of question** | To identify the necessity of multi-accent processing abilities at the same time. |

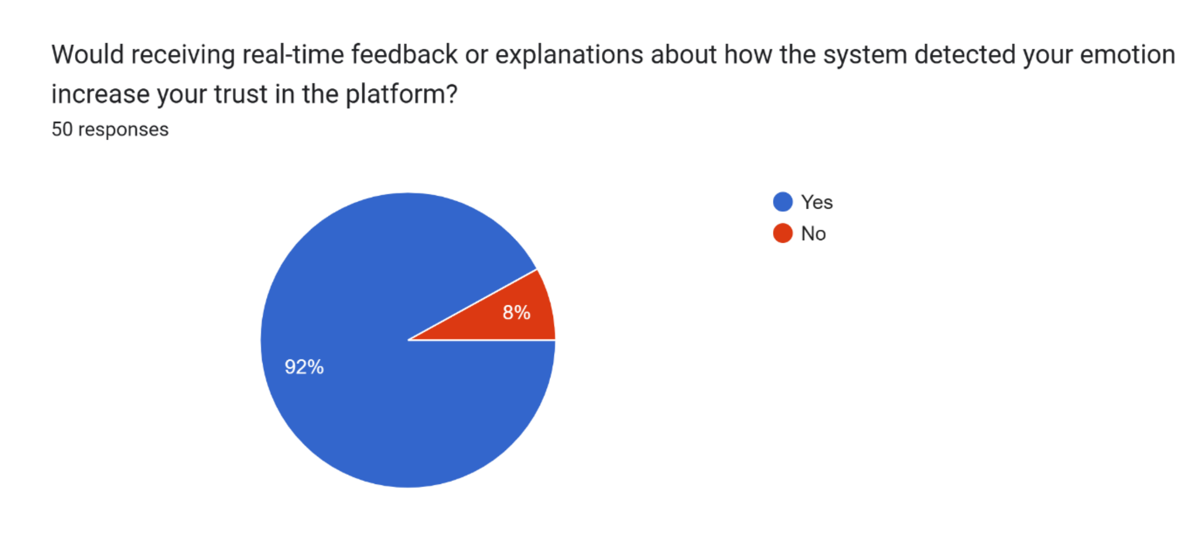


Do you think it's important for the system to support multiple languages and accents simultaneously in real-time conversations?

50 responses

| | |
|---|---|
| **Finding & Conclusion** | The support of multiple languages and accents are important or very important to 69.3% of the respondents (important or extremely important), which speaks of the global, multilingual character of contemporary digital communication. This result confirms the architecture of AEMER to deal with different accents and implies a further expansion of the system in terms of the multi-lingual possibilities and the necessity to serve the interests of many users in the multilingual environments. |

| | |
|---|---|
| **Question** | Would receiving real-time feedback or explanations about how the system detected your emotion increase your trust in the platform? |
| **Aim of question** | To determine the significance of explainable AI characteristics in creating a sense of trust in users. |

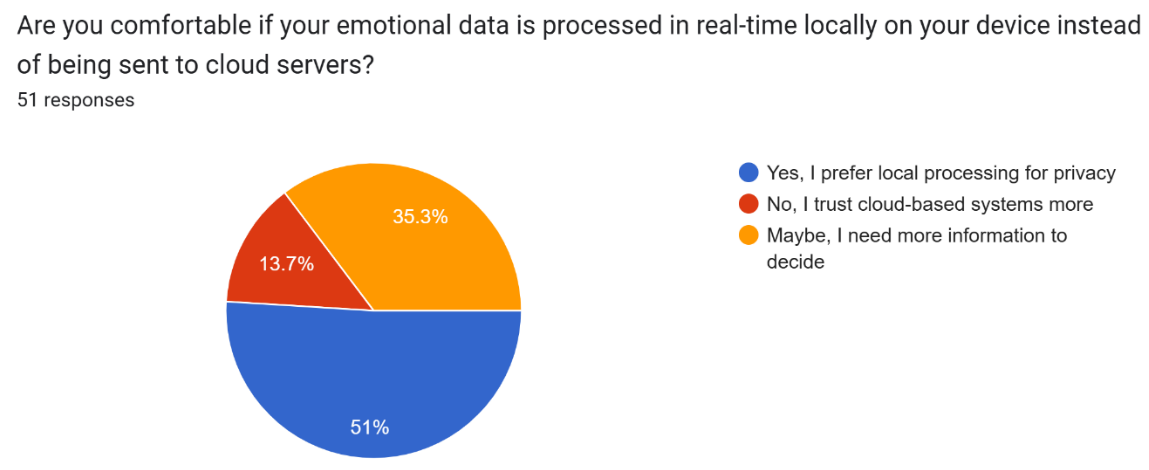


Would receiving real-time feedback or explanations about how the system detected your emotion increase your trust in the platform?

50 responses

| | |
|---|---|
| **Finding & Conclusion** | The overwhelming majority (88.5%) in relation to real-time feedback confirms that explainability becomes the most important factor in the user trust and system adoption. This discovery requires the introduction of interpretable attention visualization, modality contribution visualizations and explanations of detection reasoning. Clear functionality to aid the end-user to learn about system choices is a prerequisite to the acceptance of AEMER, especially as the topic of emotional data analysis is a sensitive one. |

| | |
|---|---|
| **Question** | Are you comfortable if your emotional data is processed in real-time locally on your device instead of being sent to cloud servers? |
| **Aim of question** | To determine the preference on privacy regarding location of data processing. |



Are you comfortable if your emotional data is processed in real-time locally on your device instead of being sent to cloud servers?

51 responses

35.3%
13.7%
51%
Yes, I prefer local processing for privacy
No, I trust cloud-based systems more
Maybe, I need more information to decide

| | |
|---|---|
| **Finding & Conclusion** | The high prevalence of local processing (46.2) and another 38.5% willing to do so shows that the issue of privacy is a key consideration. This result justifies the creation of edge-deployable implementation of AEMER that processes emotional information at the device, reducing the amount of data transfer and improving confidentiality. The provision of local and cloud-based deployment with trade-off descriptions would meet the needs of various types of users and address the issue of privacy. |