



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ανίχνευση κινδύνου μέσω αναγνώρισης συναισθήματος από
φωνή και κείμενο με χρήση μηχανικής μάθησης

ΜΑΡΙΑ ΡΑΦΑΗΛΙΑ ΤΑΚΑΚΗ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος

Επίκουρος Καθηγητής

Λαμία έτος 2026



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΕΣΣΑΛΙΑΣ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

Ανίχνευση κινδύνου μέσω αναγνώρισης συναισθήματος από φωνή και κείμενο με χρήση μηχανικής μάθησης

ΜΑΡΙΑ ΡΑΦΑΗΛΙΑ ΤΑΚΑΚΗ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΥΠΕΥΘΥΝΟΣ

Κολομβάτσος Κωνσταντίνος

Επίκουρος Καθηγητής

Λαμία έτος 2026



UNIVERSITY OF
THESSALY

SCHOOL OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE & TELECOMMUNICATIONS

Danger Detection Through Emotion Recognition from Speech and Text Using Machine Learning

MARIA RAFAHLIA TAKAKI

FINAL THESIS ADVISOR

Kolomvatsos Konstantinos

Lamia year 2026

«Με ατομική μου ευθύνη και γνωρίζοντας τις κυρώσεις ⁽¹⁾, που προβλέπονται από της διατάξεις της παρ. 6 του άρθρου 22 του Ν. 1599/1986, δηλώνω ότι:

1. Δεν παραθέτω κομμάτια βιβλίων ή άρθρων ή εργασιών άλλων αυτολεξεί **χωρίς να τα περικλείω σε εισαγωγικά** και χωρίς να αναφέρω το συγγραφέα, τη χρονολογία, τη σελίδα. Η αυτολεξεί παράθεση χωρίς εισαγωγικά χωρίς αναφορά στην πηγή, είναι λογοκλοπή. Πέραν της αυτολεξεί παράθεσης, λογοκλοπή θεωρείται και η παράφραση εδαφίων από έργα άλλων, συμπεριλαμβανομένων και έργων συμφοιτητών μου, καθώς και η παράθεση στοιχείων που άλλοι συνέλεξαν ή επεξεργάστηκαν, χωρίς αναφορά στην πηγή. Αναφέρω πάντοτε με πληρότητα την πηγή κάτω από τον πίνακα ή σχέδιο, όπως στα παραθέματα.

2. Δέχομαι ότι η αυτολεξεί **παράθεση χωρίς εισαγωγικά**, ακόμα κι αν συνοδεύεται από αναφορά στην πηγή σε κάποιο άλλο σημείο του κειμένου ή στο τέλος του, είναι αντιγραφή. Η αναφορά στην πηγή στο τέλος π.χ. μιας παραγράφου ή μιας σελίδας, δεν δικαιολογεί συρραφή εδαφίων έργου άλλου συγγραφέα, έστω και παραφρασμένων, και παρουσίασή τους ως δική μου εργασία.

3. Δέχομαι ότι υπάρχει επίσης περιορισμός στο μέγεθος και στη συχνότητα των παραθεμάτων που μπορώ να εντάξω στην εργασία μου εντός εισαγωγικών. Κάθε μεγάλο παράθεμα (π.χ. σε πίνακα ή πλαίσιο, κλπ), προϋποθέτει ειδικές ρυθμίσεις, και όταν δημοσιεύεται προϋποθέτει την άδεια του συγγραφέα ή του εκδότη. Το ίδιο και οι πίνακες και τα σχέδια

4. Δέχομαι όλες τις συνέπειες σε περίπτωση λογοκλοπής ή αντιγραφής.

Ημερομηνία:/...../20.....

Ο – Η Δηλ.

(1) «Όποιος εν γνώσει του δηλώνει ψευδή γεγονότα ή αρνείται ή αποκρύπτει τα αληθινά με έγγραφη υπεύθυνη δήλωση

του άρθρου 8 παρ. 4 Ν. 1599/1986 τιμωρείται με φυλάκιση τουλάχιστον τριών μηνών. Εάν ο υπαίτιος αυτών των πράξεων

σκοπενε να προσπορίσει στον εαυτόν του ή σε άλλον περιουσιακό όφελος βλάπτοντας τρίτον ή σκοπενε να βλάψει άλλον, τιμωρείται με κάθειρξη μέχρι 10 ε

ΠΕΡΙΛΗΨΗ

Στην παρούσα πτυχιακή εργασία εξετάζεται η χρήση φωνής και κειμένου για την αναγνώριση συναισθήματος, με στόχο η προκύπτουσα συναισθηματική πληροφορία να χρησιμοποιηθεί ως ένδειξη πιθανού κινδύνου. Το σύστημα που αναπτύχθηκε αποτελείται από δύο ξεχωριστά μοντέλα αναγνώρισης συναισθήματος. Το πρώτο επεξεργάζεται το ηχητικό σήμα, συνδυάζοντας το προεκπαιδευμένο μοντέλο WavLM με χειροκίνητα εξαγόμενα ακουστικά χαρακτηριστικά. Το δεύτερο αναλύει το κείμενο που προκύπτει από την ομιλία, χρησιμοποιώντας ένα fine-tuned μοντέλο RoBERTa. Για τη μετατροπή της ομιλίας σε κείμενο χρησιμοποιείται το Whisper, ενώ η ελληνική ομιλία μεταφράζεται επιπλέον στα αγγλικά πριν αναλυθεί από το μοντέλο κειμένου. Οι προβλέψεις των δύο μοντέλων συνδυάζονται μέσω ενός δυναμικού μηχανισμού συγχώνευσης, ο οποίος προσαρμόζει τη συνεισφορά κάθε modality ανάλογα με την αξιοπιστία της πρόβλεψής του. Το τελικό συναισθηματικό αποτέλεσμα χρησιμοποιείται στη συνέχεια για τον υπολογισμό μιας βαθμολογίας κινδύνου, με βάση προκαθορισμένα βάρη ανά συναίσθημα, η οποία οδηγεί σε ένδειξη DANGER ή SAFE. Το μοντέλο ήχου πέτυχε accuracy 73,74% και macro-F1 73,80% στο test set (RAVDESS και CREMA-D), ενώ το μοντέλο κειμένου πέτυχε accuracy 71,50% και macro-F1 59,01% στο test set του GoEmotions. Επειδή δεν ήταν διαθέσιμα επισημειωμένα δεδομένα με πραγματικές περιπτώσεις κινδύνου, το υβριδικό στάδιο και το στάδιο ανίχνευσης κινδύνου δεν αξιολογούνται ποσοτικά αλλά μέσω αντιπροσωπευτικών παραδειγμάτων. Η τελική ένδειξη κινδύνου πρέπει επομένως να ερμηνεύεται ως μια ένδειξη βασισμένη σε συναισθηματική πληροφορία και όχι ως απόδειξη ότι όντως συμβαίνει κάτι επικίνδυνο.

ABSTRACT

This thesis examines the use of speech and text for emotion recognition, with the goal of using the resulting emotional information as an indication of possible danger. The developed system consists of two separate emotion recognition models. The first one processes the audio signal, combining the pre-trained WavLM model with handcrafted acoustic features. The second one analyzes the text produced from the speech, using a fine-tuned RoBERTa model. Whisper is used to convert speech into text, and Greek speech is additionally translated into English before being processed by the text model. The predictions of the two models are combined through a dynamic fusion mechanism, which adjusts the contribution of each modality based on how reliable its prediction is. The final emotional result is then used to calculate a risk score, based on predefined weights for each emotion, which leads to a DANGER or SAFE indication. The audio model achieved an accuracy of 73.74% and a macro-F1 of 73.80% on the test set (RAVDESS and CREMA-D), while the text model achieved an accuracy of 71.50% and a macro-F1 of 59.01% on the GoEmotions test set. Because no labelled data with real danger cases were available, the hybrid stage and the danger detection stage are not evaluated quantitatively but through representative examples. The final danger indication should therefore be understood as an indication based on emotional information, and not as proof that a dangerous situation is actually taking place.

Contents

ΠΕΡΙΛΗΨΗ.....	2
ABSTRACT	3
CHAPTER 1	7
INTRODUCTION	7
1.1 Background & Motivation	7
1.2 Problem Statement	7
1.3 Aim & Objectives.....	8
1.4 Proposed Approach	9
1.5 Main Contributions	11
1.6 Thesis Organization	11
CHAPTER 2	13
BACKGROUND & RELATED WORK	13
2.1 Emotion & Emotion Representation	13
2.2 Speech Emotion Recognition	14
2.2.1 Acoustic Features	15
2.2.2 Machine Learning Approaches	15
2.2.3 Deep & Self-Supervised Speech Representations	16
2.3 Text Emotion Recognition	17
2.3.1 Emotion Recognition from Text.....	17
2.3.2 Transformer-Based Models	17
2.3.3 RoBERTa.....	18
2.4 Multimodal Emotion Recognition	18
2.4.1 Complementary Modalities	19
2.4.2 Multimodal Fusion Strategies	19
2.5 Automatic Speech Recognition & Multilingual Processing	20
2.6 Emotion Recognition in Safety & Emergency Contexts	21
2.7 Related work.....	22
2.8 Research Gap & Connection to This Thesis	23
CHAPTER 3	26
DATASETS & DATA PREPARATION	26
3.1 Overview.....	26
3.2 Speech Emotion Datasets.....	26
3.2.1 RAVDESS.....	26
3.2.2 CREMA-D.....	28
3.2.3 Custom Speech Recordings	29
3.3 Text Emotion Dataset.....	31
3.3.1 GoEmotions	31
CHAPTER 4	34

METHODOLOGY	34
4.1 System Overview	34
4.2 Audio Data Processing & Feature Representation	34
4.3 Text Data Preparation	35
4.4 Model Training & Evaluation Strategy.....	36
4.5 Speech Transcription & Translation.....	37
4.6 Audio-Text Fusion Methodology.....	37
4.7 Emotion-Based Danger Detector.....	38
CHAPTER 5	39
MODEL TRAINING & SYSTEM IMPLEMENTATION	39
5.1 Overview.....	39
5.2 Software Environment and Libraries.....	39
5.3 Audio Emotion Model Training.....	40
5.4 Text Emotion Model Training.....	41
5.5 Speech Transcription and Translation	42
5.6 Hybrid System Implementation.....	43
5.7 Emotion-Based Danger Detection	44
5.8 Live Microphone Inference	45
CHAPTER 6	46
EXPERIMENTAL SETUP & RESULTS	46
6.1 Experimental Setup.....	46
6.2 Evaluation Metrics.....	46
6.3 Experiment 1: Audio Emotion Recognition.....	47
6.4 Experiment 2: Text Emotion Recognition.....	48
6.5 Experiment 3: Hybrid Emotion Recognition.....	50
6.6 Experiment 4: Emotion-Based Danger Detection.....	52
CHAPTER 7	55
DISCUSSION, INTERPRETATION & FUTURE WORK	55
7.1 Results Interpretation	55
7.2 Strengths of the Proposed System	56
7.3 Limitations	56
7.3.1 Audio model.....	56
7.3.2 Text model.....	57
7.3.3 Transcription and translation	57
7.3.4 Hybrid Fusion.....	58
7.3.5 Emotion-Based Danger Detection	58
7.4 Future work.....	58
CHAPTER 8	60
CONCLUSION	60
8.1 Conclusion.....	60

REFERENCES.....	61
-----------------	----

CHAPTER 1

INTRODUCTION

1.1 Background & Motivation

Emotion plays an important part in human communication and social interaction. People can express how they feel in different ways, including through the words they use and through their voice. For this reason, emotion recognition has become an area of interest in artificial intelligence, since it can provide computer systems with information about a person's emotional state. [1], [2] Different approaches can be used to describe emotions, with one well-known categorical approach including happiness, sadness, anger, fear, disgust, and surprise [3].

Speech is one of the main sources that can be used to recognize emotions [4]. The human voice contains information not only about what is being said, but also about the way it is expressed. Characteristics such as pitch, intensity, and speaking rhythm can change depending on the emotional state of the speaker [2]. Speech Emotion Recognition uses information from the speech signal in order to identify the emotion expressed by a person. This field has been studied in human-computer interaction and can be applied in areas such as dialogue systems, call centers, vehicle systems, and healthcare [4].

Emotional information can also be expressed through language. Text emotion recognition focuses on identifying emotions from the words and linguistic content of a text. Several labelled datasets have been developed for studying emotion recognition from text. One example is GoEmotions, a large manually annotated dataset of English text containing a wide range of emotional categories [1].

Speech and text can therefore provide different information about the same spoken input. Speech can describe how something is said, while text provides information about what is being said. Relying only on one of these sources may leave part of the available emotional information unused. For this reason, multimodal emotion recognition combines information from different modalities so that they can complement each other and provide a more complete representation of the emotional state [2].

Emotion recognition can also be useful in applications related to emergency situations. Studies using real emergency call recordings have found that emotions such as fear and anger are frequently expressed by callers. Other emotional states such as stress, panic, pain, and annoyance may also appear during urgent calls. At the same time, recognizing emotions in real situations is more difficult than in controlled recordings, since real-life emotional expressions can vary considerably between speakers and may be more complex than acted emotions [5].

However, recognizing a strong emotion does not by itself provide a complete understanding of the situation in which it occurs. In the context of this thesis, emotional information is therefore treated as an indication that can contribute to the identification of possible danger rather than as direct proof that a dangerous event is taking place.

Based on this idea, this thesis investigates the use of both speech and text for emotion recognition. The two sources are analyzed separately, and their outputs are then combined in order to obtain a final representation of the emotional information contained in the spoken input. This information is subsequently used as the basis for an emotion-based indication of potential danger.

1.2 Problem Statement

Recognizing emotions from real speech is a challenging task because emotional expression can

vary considerably between speakers and recording conditions [5]. Systems that perform well with controlled or acted data may face additional difficulties when they are applied to real situations [4]. Differences between speakers, the quality of the recording, short utterances, and the presence of mixed emotions can all make the recognition process more difficult.

Another challenge is that speech and text represent different parts of the available emotional information. The audio signal provides information related to the way an utterance is expressed, while the textual content provides information related to its linguistic meaning. Using only one of these sources may therefore limit the information available to the emotion recognition system. Combining different modalities can help make use of complementary information contained in each source [2].

In the system examined in this thesis, the textual information is obtained from spoken input. This means that the text branch depends on an automatic transcription stage before emotion classification can be performed. As a result, the final textual input is also affected by the information produced during speech recognition. This creates an additional processing step compared with systems in which the text is provided directly.

A further difficulty is the relationship between emotion and danger. The recognition of emotions such as fear or anger can provide useful information in situations related to urgency or emergency [5], but an emotional state alone does not describe the complete context of an event. Therefore, the objective of this thesis is not to treat a recognized emotion as direct evidence of danger, but to use the final emotional information as an indication of a potentially dangerous situation.

The main problem addressed in this work is therefore how speech and text emotion recognition can be combined within a single processing pipeline and how the resulting emotional information can be used to produce an interpretable indication of potential danger. The system must first obtain emotional information independently from the audio and textual inputs and then combine these results before producing the final emotion-based danger indication.

1.3 Aim & Objectives

The main aim of this thesis is to develop and examine a system that can recognize emotions from both speech and text and use this information as an indication of potential danger. Instead of relying only on the acoustic characteristics of the voice or only on the meaning of the spoken words, the proposed system uses both types of information. In this way, the emotional state of the speaker is examined from two different sources before the final result is produced.

The emotion recognition task is based on seven emotion categories: neutral, happy, sad, anger, fear, disgust, and surprise. Two separate emotion recognition components are used for this purpose. The first one processes the audio signal and produces an emotion prediction based on information contained in the speaker's voice. The second one analyzes the textual content of the spoken input and produces a separate emotion prediction. Keeping the two components separate makes it possible to examine the information provided by each source before their results are combined.

Another objective of the thesis is to support spoken input in both English and Greek. Since the text emotion recognition component processes textual information, the spoken input first has to be converted into text. For English speech, the produced transcription can be processed directly by the text emotion recognition component. In the case of Greek speech, an additional translation stage is required so that an English version of the spoken content can be provided to the text model. In this way, the system can process speech from both languages while using the same text emotion recognition component.

After the audio and text components produce their predictions, their results are combined in order to obtain a final emotion result. The purpose of this combination is to take into account the

information provided by both sources rather than depending completely on only one of them. The final emotional information is then used in the last stage of the system, where an emotion-based risk value is produced. This value is used to provide an indication of whether the input may be related to a potentially dangerous situation.

- ◆ To develop an audio emotion recognition component capable of distinguishing between the seven selected emotion categories.
- ◆ To develop a text emotion recognition component that uses the same emotion categories.
- ◆ To process spoken input and automatically obtain the textual information required by the text component.
- ◆ To support both English and Greek spoken input, including the translation of Greek speech when necessary.
- ◆ To combine the emotion predictions obtained from the audio and text components into a final emotion result.
- ◆ To use the final emotional information to produce an indicative measure of potential danger.
- ◆ To evaluate the performance of the individual emotion recognition components and examine the behavior of the complete processing pipeline.

The overall purpose is not only to classify the emotion expressed in an input, but also to investigate how information from speech and text can work together within the same system. The thesis therefore examines the complete process, starting from spoken input, continuing with emotion recognition from the two sources, and ending with the production of an emotion-based indication of possible danger.

1.4 Proposed Approach

The approach followed in this thesis is based on the use of both speech and text for emotion recognition. The main idea is that a spoken input can provide emotional information in more than one way. Information can be obtained directly from the audio signal, but at the same time the words spoken by the person can also provide useful information about the expressed emotion. For this reason, the proposed system processes these two types of information separately at first and combines their results at a later stage. The complete process starts from a speech recording and ends with an emotion-based indication of possible danger.

The first part of the system is responsible for recognizing emotions directly from the audio signal. For the development of this part, recordings from the RAVDESS [6] and CREMA-D [7] datasets are used together with additional speech recordings collected specifically for this thesis. Both datasets contain recordings in which different emotions are expressed by multiple speakers, providing labelled speech examples that can be used for emotion recognition. The additional recordings were included in order to add samples created specifically for the purposes of the present work. All the audio data used by the system are finally organized according to the seven emotion categories considered in this thesis: neutral, happy, sad, anger, fear, disgust, and surprise.

The audio component uses two different types of information from each speech recording. The first one is based on acoustic characteristics that can be calculated directly from the audio signal. These characteristics describe different properties of the voice and provide information that may be useful for distinguishing between emotional states. The second type of information is obtained using WavLM [8], a pre-trained model designed to learn useful representations from speech. Instead of using only one of these approaches, the system combines the WavLM representations with the acoustic features extracted from the recording. The combined information is then given to the audio classifier, which produces a prediction for the emotional state expressed in the input. More detailed information about the exact acoustic features, the WavLM representation and the classification process is presented in later chapters.

The second part of the system focuses on the textual content of speech. For the development of the text emotion recognition component, the GoEmotions dataset is used [1]. GoEmotions contains a large collection of English text samples labelled with different emotions. However, the original emotion categories of the dataset are not exactly the same as the seven categories selected for this thesis. For this reason, the data are processed and only suitable samples are used so that the text component follows the same emotion classification scheme as the audio component. This is important because both components must eventually produce results that refer to the same set of emotions before their outputs can be combined.

After the preparation of the textual data, RoBERTa [9] is used as the main model for text emotion recognition. The pre-trained model is fine-tuned using the selected text samples so that it can classify an English sentence into one of the seven emotion categories. In this part of the system, the model does not receive information from the audio signal itself. It only examines the textual content that is provided to it. In this way, the audio and text components remain independent during emotion recognition and each one produces its own result before the fusion stage takes place.

An additional step is required because the complete system receives spoken input rather than text that has already been written by the user. The speech must therefore first be converted into textual form before it can be processed by RoBERTa. For this purpose, Whisper is used as the speech recognition component of the system [10]. Whisper can process multilingual speech, which makes it suitable for the English and Greek inputs considered in this thesis. When the input is in English, the produced transcription can be used directly by the text emotion recognition model. When Greek speech is provided, an additional translation process is needed because the text emotion model operates on English input. The Greek spoken content is therefore converted into an English textual form before being passed to the text classifier.

At this point, the system has two separate emotion results for the same spoken input. One result comes from the audio component and is based on the information contained in the voice, while the other comes from the text component and is based on the linguistic content obtained from the speech. These results are combined through a dynamic fusion mechanism. The purpose of the fusion stage is not simply to give the two components the same importance in every case. Instead, the contribution of each component can change depending on information related to the reliability of the prediction produced for the specific input.

This is useful because the two components do not necessarily behave in the same way for every recording. In some cases, the audio information may provide a clearer emotion result, while in other cases the textual information may be more useful. The fusion mechanism therefore examines information produced by the two components and adjusts their contribution according to predefined rules. The result of this process is a final distribution across the seven emotion categories. No additional multimodal classifier is trained specifically for this stage. Instead, the combination is carried out by the predefined fusion procedure used in the system.

The final part of the proposed approach uses this combined emotion information to produce an indication related to potential danger. The complete emotion distribution is taken into account rather than using only the emotion with the highest score. Different emotion categories can have a different influence on the calculated risk value according to predefined parameters of the system. After the risk value is calculated, it is compared with a predefined threshold in order to produce the final danger related output.

Overall, the proposed approach brings the different parts of the work together in a single processing pipeline. A speech recording is first examined directly through the audio emotion recognition component, while its spoken content is also converted into text and analyzed by the text emotion recognition component. The two emotion results are then combined, and the final emotional information is used to calculate the danger-related indication. In this way, the system makes use of both the way something is spoken and the content of what is being said before producing its final result.

1.5 Main Contributions

The main contribution of this thesis is the development of a complete emotion recognition pipeline that brings together information from both speech and text. Although the two types of information are processed independently, their results are connected within the same system, allowing the final decision to take into account both the characteristics of the speaker's voice and the content of the spoken message.

Another contribution of the work is the combination of different types of information within the audio emotion recognition stage. Pre-trained speech representations are used together with acoustic features extracted directly from the audio signal. This makes it possible to examine emotional information from the recording using both learned representations and characteristics calculated from the speech itself.

The thesis also connects emotion recognition with multilingual spoken input. Through the use of automatic transcription and translation when required, the complete pipeline can process both English and Greek speech while keeping the same text emotion recognition component. In this way, the text analysis stage can be included in a system that starts directly from spoken input rather than from text that is already available.

Finally, the work extends the emotion recognition pipeline with a final stage that uses the combined emotion information to produce an indicative measure of potential danger. This creates a connection between multimodal emotion recognition and the examination of potentially dangerous situations, while keeping the final result as an emotion-based indication rather than treating it as direct evidence that a dangerous event has occurred.

1.6 Thesis Organization

The rest of this thesis is organized into seven additional chapters, each covering a different part of the work.

Chapter 2 presents the theoretical background and related work concerning emotion recognition from speech and text. It introduces the main concepts used in speech emotion recognition, text emotion recognition, and multimodal emotion recognition. It also discusses the use of pre-trained models for speech and language processing, as well as previous work related to emotion recognition in situations where emotional information may be useful for identifying urgency or possible danger.

Chapter 3 describes the datasets used for the development and evaluation of the emotion recognition components. The speech datasets used for the audio part of the system are presented together with the additional recordings collected for this thesis. The textual dataset used for the development of the text emotion recognition component is also described. The chapter explains the main characteristics of the datasets and how the available data are prepared for the following stages of the work.

Chapter 4 presents the methodology followed for the development of the proposed system. It describes the processing of the audio and textual data, the extraction and preparation of the information used by the two emotion recognition components, and the general training and evaluation procedure. The methodology used for combining the audio and text predictions and for producing the final emotion-based risk indication is also presented.

Chapter 5 focuses on the implementation of the complete system. It explains the software tools and libraries used during development and describes the implementation of the audio and text emotion recognition components. The transcription and translation stages are also presented, followed by the dynamic fusion process, the danger-related decision stage, and the use of the complete system with spoken input.

Chapter 6 presents the experimental results obtained from the developed models and the

complete processing pipeline. The performance of the audio and text emotion recognition components is examined using quantitative evaluation metrics. In addition, representative examples are used to examine the behavior of the combined audio-text system and the final danger related output.

Chapter 7 discusses the results of the work and the main observations that can be made from them. The strengths and limitations of the different parts of the system are examined, including the audio and text models, the transcription and translation process, the combination of the two modalities, and the final emotion-based danger indication. Possible directions for future improvement and extension of the proposed system are also presented.

Finally, Chapter 8 summarizes the main conclusions of the thesis and provides an overall view of the work carried out and the results obtained.

CHAPTER 2

BACKGROUND & RELATED WORK

2.1 Emotion & Emotion Representation

Emotions are complex mental states that are connected with a person's subjective experience and with the way people react to situations and changes in their environment. Emotional expression can also appear through different channels, including the voice, facial expressions, gestures, and other non-verbal signals. Because emotions can be described in different ways, their representation is an important part of emotion recognition research. In studies that use artificial intelligence for emotion recognition, two main approaches are commonly considered: categorical and dimensional emotion models [11].

The categorical approach represents emotions as separate emotional states. Instead of describing an emotion through continuous values, a specific category is used to indicate the emotional state that is being expressed. A widely known group of emotions associated with this approach includes happiness, anger, sadness, surprise, disgust, and fear [3], [11]. These emotions have been examined extensively in research concerning basic emotions and emotional expressions.

An important part of the basic emotion view is the idea that some emotional responses have a biological basis. Research on basic emotions has examined evidence related to universality in emotional expression and physiology, supporting the existence of biologically based aspects of emotion. At the same time, this view does not exclude the influence of culture and social learning. Cultural and social factors can affect the control of emotional expressions, the representation of emotional experiences, the evaluation of emotional situations, and the way people respond to their own emotions. Evidence from facial expression research has also supported a group that includes happiness, surprise, fear, sadness, anger, and disgust. However, the exact number of emotions that should be considered basic has not been treated as completely settled, and further research has been considered necessary for determining whether additional emotions should be included [3].

Facial expressions are not the only type of evidence considered within the basic emotion approach. Physiological responses have also been examined, with different patterns of autonomic nervous system activity reported for emotions including anger, fear, sadness, and disgust. Additional research reproduced several of these observations and provided further evidence concerning emotion-specific physiological activity. The basic emotion approach therefore considers evidence from both emotional expressions and physiological responses [3].

Although categorical models describe emotions as separate states, another way of representing affect is to place emotional states within a continuous space. This is the basis of dimensional approaches. Dimensional emotion models are described using dimensions such as valence, arousal, and dominance. Valence is connected with the positive or negative quality of an emotional experience, while arousal describes its level of activation. The review also identifies the two-dimensional valence-arousal representation as a commonly used dimensional approach in emotion recognition research [11].

One of the most influential dimensional descriptions of affect is the circumplex model. Affective states are presented as systematically related rather than completely independent from one another and are arranged within a two-dimensional bipolar space. In this representation, the main dimensions are pleasure-displeasure and degree of arousal. The circumplex model arranges affective concepts around a circular structure, including pleasure, excitement, arousal, distress, displeasure, depression, sleepiness, and relaxation. These concepts are positioned at different locations around the circle, with pleasure-displeasure forming one axis and arousal the other [12].

This representation also provides a way of describing a wider range of emotional terms. Many affect terms do not simply fall directly on the pleasure or arousal axes but instead appear in different areas around the affective space. An emotional term can therefore be represented through a combination of pleasure and arousal rather than through only one of these dimensions. The dimensions are bipolar, and opposite affective states can appear in different areas of the same space. The axes can also be rotated while the same general circular organization is maintained. For example, the pleasure and arousal dimensions can be described through combinations such as excitement-depression and distress-contentment. These alternatives provide complementary ways of describing the same general structure of affect. The circumplex representation also allows affective states of different intensities to occupy different positions within the space. States of moderate intensity can be located closer to the middle, while the origin of the space can correspond to an adaptation level or a neutral feeling [12].

Categorical and dimensional representations describe affect in different ways. The categorical approach separates emotional states into distinct labels, while the dimensional approach represents them through continuous emotional dimensions. However, these two approaches are not always completely separated in emotion recognition research. Values from valence and arousal can be divided into discrete targets for classification, while categorical emotions can also be mapped onto a valence-arousal space. A systematic review examined 51 studies with respect to their affective state modeling. Among them, 44 used categorical emotion models, 15 used dimensional models, and eight included both approaches. The two-dimensional valence-arousal model was the most common dimensional representation among the examined studies [11].

Categorical modeling was the preferred approach among the studies examined in the systematic review. This trend was partly associated with the widespread use of datasets such as IEMOCAP and MELD, which contain categorical emotion annotations, as well as with the use of similar annotation systems for comparison with previous studies. From a practical point of view, categorical modeling can use a single model to distinguish between several emotional states, while dimensional modeling is usually handled separately for each affective dimension. Differences also exist in the annotation process, since categorical labels are easier to explain to human annotators, whereas dimensional labeling requires a more detailed perception of the emotional state being observed or experienced. The review also examined agreement between different annotation schemes and reported differences between self-annotations, partner annotations, and ratings from external annotators. Based on these findings, annotation quality and inter-rater agreement were identified as important issues that should be considered in the development and evaluation of emotion recognition systems [11].

Based on the emotion representations discussed above, this thesis follows a categorical approach to emotion recognition. Both the speech and text components are designed to produce predictions using the same seven emotion categories: neutral, happy, sad, anger, fear, disgust, and surprise. Six of these categories correspond to emotions commonly considered within the basic emotion approach [3], while neutral is included as an additional category in the classification scheme used in this work. Using the same categorical representation for both components also provides a common output structure, which allows their emotion predictions to be combined later in the proposed system.

2.2 Speech Emotion Recognition

Speech is an important source of emotional information because the way a person speaks can provide information beyond the linguistic content of an utterance. Speech emotion recognition focuses on identifying emotional states from speech and has become an important area in human-computer interaction, with applications in spoken dialogue systems, call centers, vehicle systems, and medical applications. A traditional speech emotion recognition process can generally be divided into two main stages: feature extraction and classification. In the first stage, different characteristics of the speech signal are extracted, while in the second stage this information is

provided to a classifier in order to determine the emotional state. Different speech features and classification methods have been used for this purpose, while the development of deep learning has also introduced approaches that can learn useful features automatically from speech. At the same time, applying emotion recognition systems outside controlled conditions introduces additional difficulties that still need to be addressed [4].

Current speech emotion recognition approaches can therefore use different ways of obtaining information from the audio signal. Some methods first calculate hand-crafted features and then provide them to a machine learning classifier. Other approaches use deep learning models that learn useful information directly from speech, while speech signals can also be transformed into representations such as spectrograms before being processed by deep networks. Combinations of hand-crafted and deep features have also been examined, allowing information obtained through different feature extraction methods to be used together [11].

2.2.1 Acoustic Features

The information contained in speech can be described through different acoustic characteristics. Hand-crafted features are selected and calculated based on previous knowledge of the speech signal and are commonly divided into prosodic, voice quality, and spectral features. Prosodic features describe characteristics such as pitch, tone, rhythm, and stress, with common examples including fundamental frequency, energy, and speech rate. Fundamental frequency, usually represented as F0, can vary according to the emotional state of the speaker. Higher and wider F0 ranges have been associated with emotional states such as anger, while sadness can be related to lower and narrower values. Differences between emotional states can also appear in the intensity of speech and in the rate at which words are spoken [2].

Voice quality features provide another way of describing the speech signal and include formants, harmonic-to-noise ratio, jitter, and shimmer. Formants are related to characteristics of the vocal tract, while the harmonic-to-noise ratio describes the relation between harmonic and noise components in speech. Jitter represents rapid variations in fundamental frequency, whereas shimmer is related to changes in amplitude between neighboring vocal cycles. Spectral features, on the other hand, describe information from the frequency representation of the signal, since emotional expression can affect how spectral energy is distributed across different frequency regions. Examples include Linear Prediction Coefficients, Linear Prediction Cepstral Coefficients, and Mel-Frequency Cepstral Coefficients. MFCCs are widely used in speech processing and represent information from the power spectrum using the nonlinear mel scale, which approximates characteristics of human auditory perception [2].

Hand-crafted acoustic features can therefore provide different types of information about emotional speech, but their usefulness may vary between speakers and speech conditions. Their extraction also depends on the selection of suitable characteristics based on previous knowledge. Because it can be difficult for manually selected features to represent all the emotional information contained in speech, recent research has also moved toward methods that can learn speech representations automatically. These approaches are discussed later in the section on deep and self-supervised speech representations [2].

2.2.2 Machine Learning Approaches

Traditional speech emotion recognition systems commonly combine acoustic feature extraction with a classification algorithm. Once the selected characteristics have been calculated from the speech signal, the resulting feature representation is used as the input of a classifier. Different machine learning methods have been examined for this task. Support Vector Machines are among the classifiers that have been used for speech emotion recognition. Other approaches discussed in the literature include Gaussian Mixture Models and Hidden Markov Models. K-Nearest Neighbor and decision tree methods have also been applied to emotion recognition. These approaches operate on information that has already been extracted from the audio signal,

meaning that the quality and selection of the input features form an important part of the recognition process [4].

As speech emotion recognition developed, deep learning methods also became increasingly important. In contrast with approaches that depend completely on manually designed features, deep learning models can learn internal representations from the input data. Deep Neural Networks use multiple layers between their inputs and outputs, while convolutional and recurrent architectures have also been applied to different processing tasks. Recurrent Neural Networks and Long Short-Term Memory networks are particularly relevant to sequential information and have been used for speech-based classification [4].

The use of deep learning also changed the relationship between feature extraction and classification. In traditional approaches these stages can be treated separately, whereas a deep network can include feature learning as part of the overall training process. A systematic review of speech emotion recognition methods describes a shift from traditional machine learning approaches that require pre-extracted inputs toward deep learning methods capable of learning features from speech. However, hand-crafted features have not disappeared from recent research, and combinations of manually extracted and deep features are also used. This development is important because the different approaches do not necessarily need to be treated as completely separate alternatives. A system can use expert-designed acoustic information while also using representations obtained through deep learning. In this way, information derived directly from known acoustic characteristics can be combined with information learned automatically from larger speech representations [11].

2.2.3 Deep & Self-Supervised Speech Representations

One limitation of relying only on hand-crafted features is that their design requires previous knowledge and that a manually defined feature set may not represent all the emotional information contained in speech. Deep learning provides another approach by allowing models to learn speech representations automatically from the input data. Self-supervised learning extends this idea by allowing models to be pre-trained on large amounts of speech without requiring manually assigned labels for every training example. The representations learned during pre-training can later be used for different downstream speech tasks. Models such as wav2vec, wav2vec 2.0, and HuBERT are examples of self-supervised approaches developed for speech representation learning [2].

WavLM is another self-supervised pre-trained model developed for a wide range of speech processing tasks. Speech contains different types of information, including spoken content, speaker identity, and paralinguistic information, making the development of a general representation a challenging task. During pre-training, WavLM combines masked speech prediction with denoising and uses a Transformer architecture with gated relative position bias to better capture the sequence ordering of the input. These characteristics were designed to support different downstream tasks rather than focusing only on automatic speech recognition [8].

WavLM was evaluated on the SUPERB benchmark, which covers different aspects of speech, including content, speaker information, semantics, paralinguistics, and generation. Emotion Recognition is included among the tasks of this benchmark, together with applications such as Automatic Speech Recognition, Speaker Identification, and Speech Translation. Therefore, the evaluation of WavLM also includes its use for extracting information relevant to emotion recognition [8].

Self-supervised WavLM representations have also been examined specifically for speech emotion recognition. One recent approach uses WavLM to extract deep acoustic representations from speech and then processes these representations using a bidirectional recurrent network and multi-head attention. The recurrent component is used to capture temporal dependencies in the speech sequence, while the attention mechanism gives greater importance to speech segments

that contain more relevant emotional information. The approach was evaluated on the RAVDESS and EmoDB emotion datasets, providing an example of how WavLM representations can be adapted specifically to speech emotion recognition [13].

The use of self-supervised models does not exclude hand-crafted acoustic features. Recent speech emotion recognition research also includes approaches that combine hand-crafted and deep features within the same system [11]. This idea is directly related to the audio approach followed in this thesis, where acoustic features are combined with representations obtained from WavLM before emotion classification. As a result, the audio component uses information obtained from both manually defined acoustic characteristics and self-supervised speech representations.

2.3 Text Emotion Recognition

Text provides another source of information for identifying emotional states. Unlike speech emotion recognition, which analyzes characteristics of the audio signal, text emotion recognition focuses on the linguistic content and the emotional information expressed through language. Different methods have been developed for extracting useful information from text, ranging from traditional representations such as bag-of-words to neural and contextual representations learned from large amounts of textual data [2].

The availability of annotated datasets has also supported the development of language-based emotion classification. One example is GoEmotions, which was created as a large human-annotated dataset for fine-grained emotion recognition. It contains approximately 58,000 English Reddit comments labelled with one or more of 27 emotion categories or Neutral. The dataset was designed to provide a broader emotion taxonomy than many earlier resources and to support the development and evaluation of emotion classification models [1].

2.3.1 Emotion Recognition from Text

Early approaches to textual feature extraction often used representations such as bag-of-words, which describe text through the occurrence of words but do not preserve word order or grammatical relationships, leading to a loss of contextual information. Word embeddings were introduced as another way of representing text by mapping words to numerical vectors, with related words obtaining similar representations. Approaches such as Word2Vec and GloVe learned these representations from large amounts of unlabeled text but generally assigned a single vector to each word regardless of the context in which it appeared. This limitation led to the development of contextual approaches, where the representation of a word can change according to the surrounding language [2].

Text emotion recognition also depends on the emotional categories used to annotate the available data. GoEmotions was developed using a fine-grained taxonomy containing 27 emotion categories together with Neutral. The study also examined whether these categories could be grouped into broader emotion representations and reported that the relations among the emotions allow higher-level groupings when required by a downstream task. The dataset was also used to evaluate emotion classification models, including a fine-tuned BERT model [1].

2.3.2 Transformer-Based Models

Contextual language models represent words according to the surrounding text rather than assigning the same representation to a word in every situation. This makes it possible for the representation of a word to change depending on the context in which it appears [2].

The Transformer architecture was introduced as an approach to sequence processing based on attention mechanisms rather than recurrent or convolutional networks. A central part of the architecture is self-attention, which allows different positions of the same sequence to be related when constructing its representation. The Transformer follows an encoder-decoder structure in

which both parts contain stacked attention and feed-forward layers. Multi-head attention extends this process by allowing the model to attend to information from different representation subspaces and different positions at the same time [14].

BERT is a pre-trained language representation model designed to learn bidirectional contextual representations from unlabeled text. Instead of processing the context only from one direction, it considers both the left and right context during pre-training. One of its main pre-training objectives is masked language modeling, in which some input tokens are hidden and the model is trained to predict their original values. After pre-training, the model can be adapted to different natural language processing tasks through fine-tuning, with relatively small changes to the task-specific architecture [15].

Pre-trained transformer models have also been applied to emotion classification. In the GoEmotions study, a fine-tuned BERT model was used as a baseline for fine-grained emotion classification and achieved better results than the BiLSTM model examined in the same work [1].

RoBERTa was later developed by re-examining several choices used during BERT pre-training while maintaining the general BERT architecture. These modifications are discussed in the following subsection [9].

2.3.3 RoBERTa

RoBERTa, which stands for Robustly Optimized BERT Pretraining Approach, was developed by modifying several choices used during BERT pre-training. The original RoBERTa study examined the effect of training for longer periods, using larger batches and more data, removing the next sentence prediction objective, using longer sequences, and changing the masking pattern dynamically during training. These modifications were combined into the RoBERTa pre-training procedure [9].

One of the changes introduced in RoBERTa is dynamic masking, where the masked positions can change during training instead of remaining fixed. The model also removes the next sentence prediction objective used in the original BERT procedure and is trained using larger mini-batches and more pre-training data. RoBERTa was evaluated on several natural language processing benchmarks, including GLUE, RACE, and SQuAD, where the modified pre-training procedure produced improved results. These findings showed that changes in the way BERT is pre-trained can have an important effect on downstream performance [9].

More recent research has also examined the use of RoBERTa specifically for fine-grained emotion recognition. One approach extends the standard RoBERTa model with an emotion-specific attention mechanism designed to give greater importance to emotionally relevant parts of the text. A TF-IDF-based gating mechanism is also used to reduce the influence of less informative tokens. This approach was developed to address difficulties such as subtle emotional expressions, class imbalance, and noisy or informal textual input, and the reported results showed improved emotion classification performance compared with the baseline RoBERTa model [16].

In this thesis, RoBERTa is used as the basis of the text emotion recognition component. The pre-trained model is adapted to the emotion categories used in the proposed system so that textual input can be classified according to the same categorical emotion representation used by the audio component. This common output structure later allows the predictions obtained from speech and text to be combined within the multimodal stage.

2.4 Multimodal Emotion Recognition

Multimodal emotion recognition combines information from more than one source in order to identify emotional states. Speech, text, and visual information are among the modalities commonly considered in this type of system. The motivation for combining them is that a single modality may provide only part of the available emotional information, while different sources

can provide information that supports or complements one another. As a result, multimodal systems aim to use information from several modalities in order to obtain a more complete representation of emotional expression [2], [17].

2.4.1 Complementary Modalities

Different modalities describe different aspects of emotional expression. Speech can provide information related to the way an utterance is produced, while text provides information related to its semantic content. Audio can include characteristics such as speech rate, tone, and volume, while textual information can provide richer semantic information. At the same time, audio does not provide contextual semantic information in the same way as text, while textual information may be ambiguous and can also be affected by the quality of text recognition. For this reason, combining speech and text can provide information that is not equally available from either modality when it is considered alone [18].

The combination of several modalities can also provide a richer and more complete representation of emotional information. However, the contribution of each modality to the final recognition result is not necessarily the same. Different sources may provide more useful information in different situations, which means that multimodal emotion recognition involves not only combining several modalities but also considering their relative contribution to the final prediction [17].

Relying too strongly on a single modality can also introduce limitations. Greater dependence on textual information may overlook emotional cues contained in vocal tone, while relying mainly on vocal or visual information may fail to capture information expressed through linguistic content. For this reason, some multimodal approaches adjust the contribution of the available modalities according to the information provided in each case [17].

2.4.2 Multimodal Fusion Strategies

The information obtained from different modalities can be combined at different stages of an emotion recognition system. Fusion methods are commonly distinguished according to the point at which multimodal information is integrated. These approaches include feature-level or early fusion, decision-level or late fusion, hybrid fusion, and methods that perform fusion within intermediate layers of deep models [2], [18].

In early fusion, also referred to as feature-level fusion, information from the individual modalities is combined after feature extraction and before the final classification stage. The resulting multimodal representation is then processed by a common classification model. This approach allows relationships between modality features to be introduced early in the recognition process. However, differences in time scales, temporal structures, and the characteristics of the individual modalities can make this type of fusion more difficult [2].

In late fusion, also known as decision-level fusion, the individual modalities are processed separately before their results are combined. Each modality can therefore use its own feature representation and classification model, while the predictions produced by the separate components are later integrated into a final result. Possible combination methods include averaging, weighted sums, majority voting, or additional decision models. This approach is more flexible with respect to the individual processing of each modality and does not require the same temporal synchronization of feature representations as early fusion [2].

Decision-level fusion has also been applied specifically to speech and text emotion recognition. In this case, audio and textual information are processed separately and their outputs are combined after the individual recognition stages. The distinction between the two modalities is important because audio provides information related to how an utterance is spoken, while text provides information related to its semantic content [18].

Another approach is hybrid fusion, which combines characteristics of early and late fusion. Some information can first be integrated at the feature level, while additional results are later combined at the decision level. Fusion can also take place inside intermediate layers of deep neural networks, allowing relationships between modalities to be learned during the processing of their representations [2].

Apart from the stage at which fusion is performed, the contribution assigned to each modality is another important consideration. Fixed weighting gives the modalities the same relative influence across different inputs, while adaptive approaches allow their contribution to change according to the available information. Recent multimodal emotion recognition approaches have therefore examined dynamic weighting mechanisms that adjust the influence of the individual modalities instead of keeping it constant in every case [17].

The approach used in this thesis follows a decision-level structure because the speech and text components first perform emotion recognition independently and their outputs are combined afterwards. The contribution of the two components is not fixed in every case. Instead, the proposed dynamic fusion mechanism adjusts their relative contribution using predefined information related to the reliability of the individual predictions. In this way, the audio and text recognition stages remain separate while both contribute to the final emotion result.

2.5 Automatic Speech Recognition & Multilingual Processing

Automatic Speech Recognition provides a way to convert spoken input into a textual representation that can later be processed by language-based systems. Whisper was developed as a multilingual and multitask speech processing model and was trained on approximately 680,000 hours of audio paired with text. Its training data cover different speakers, recording environments, and languages, while part of the data is also used for speech-to-English translation. The model follows an encoder-decoder Transformer architecture, where the audio signal is converted into a log-Mel spectrogram and processed by the encoder, while the decoder generates the corresponding text output. Within the same sequence-to-sequence framework, the model can perform several tasks, including speech transcription, speech translation, language identification, voice activity detection, and time-aligned transcription [10].

Multilingual processing is an important part of this design. Approximately 117,000 hours of the training data contain audio in 96 languages other than English, while additional data are included for translation from other languages into English. The required task is specified through special tokens provided to the decoder. These tokens can indicate the spoken language and whether the model should perform transcription or translation. The multilingual capabilities of the model were evaluated for speech recognition, speech-to-English translation, and language identification, with results showing that performance can vary across languages. Greek is also included among the languages reported in the multilingual evaluations [10].

When automatic transcription is used as input to an emotion recognition system, the quality of the ASR output becomes important for the subsequent emotion classification process. Variability and errors in automatically generated transcripts can affect emotion recognition performance because the textual representation may differ from the original spoken content. Recent multimodal research has therefore examined emotion recognition systems that combine vocal information with ASR-generated transcripts while specifically considering the presence of transcription noise. One such approach uses noise-aware training to make the text representation more robust to common ASR errors before combining information from speech and text [19].

Beyond transcription errors, multilingual emotion recognition introduces another difficulty when textual information is translated from one language into another. Emotional content may not always be preserved in exactly the same way after translation, and emotion recognition performance can therefore vary between languages. Research comparing original English texts with versions translated into Spanish, German, and French found that emotions could still be

identified across the examined languages, although recognition performance differed between them. Research comparing original English texts with versions translated into Spanish, German, and French showed that emotion identification remained possible across the examined languages, although performance varied between them. These results indicate that translation and linguistic differences are relevant factors in cross-lingual emotion recognition [20].

In this thesis, automatic speech recognition connects the original spoken input with the text emotion recognition component. Speech is first transcribed before textual emotion classification is performed. English speech can be processed using the produced transcript, while Greek speech requires additional processing so that the textual input is available in the language expected by the text emotion model. The complete transcription and translation procedure used in the proposed system is described later in the methodology and implementation chapters.

2.6 Emotion Recognition in Safety & Emergency Contexts

Emotion recognition can be relevant in situations where people communicate under urgent or stressful conditions. Emergency call centers are one example, since callers may express different emotional states while describing an event and requesting assistance. In recordings from real emergency calls, states such as stress, pain, fear, panic, annoyance, and anger have been observed, while fear and anger were among the most frequent emotional categories expressed by callers [5]. Speech produced under stress can also present changes in acoustic characteristics. Studies examining stressed speech have reported changes in fundamental frequency, voice intensity, speech rate, pauses, and hesitations, with particularly strong changes in some characteristics occurring in dangerous or emergency situations [21].

The conditions in which this type of speech is recorded can create additional difficulties for automatic emotion recognition. Real emergency conversations can involve many different speakers, short speech segments, uneven distributions of emotions, mixed emotional expressions, and recordings of lower quality [5]. Background noise is another important factor, since it can reduce speech emotion recognition performance when systems are used outside controlled environments. This is especially relevant in critical applications such as emergency response systems, where both accurate emotion recognition and understandable speech can be important [22].

Emotion recognition in these environments also differs from recognition based on acted or controlled emotional recordings. Acted datasets normally contain emotional expressions produced under controlled conditions, whereas spontaneous datasets represent more natural interactions and can contain greater variability and more subtle emotional expressions. Both types of data can be useful, but spontaneous recordings provide information that is particularly relevant when emotion recognition systems are considered for real-world use [11]. Research on speech under stress follows a similar direction, with datasets containing simulated as well as real stress conditions, including medical emergencies, aviation scenarios, and other situations involving increased pressure [21].

The difficulty of moving from controlled data to real emergency recordings has also been observed experimentally. When the same speech emotion recognition architecture was evaluated on a research corpus and on recordings from a real medical emergency call center, lower recognition performance was reported for the real-life recordings [5]. This difference shows that performance obtained under controlled conditions cannot automatically be expected to remain unchanged when a system is applied to more variable and demanding real-world environments.

Here, emotional information is used as an indication that can contribute to the assessment of a potentially dangerous situation. The presence of emotions such as fear or anger is not treated as direct evidence that danger exists, since emotional reactions can also appear in situations that are not dangerous. Instead, the emotional information produced by the system is used as the basis for an indicative emotion-based risk assessment.

2.7 Related work

Research on automatic emotion recognition has followed several different directions, including speech-based recognition, multimodal emotion recognition, robustness to real-world conditions, and the use of emotional or linguistic information in safety-related applications. In speech emotion recognition, recent work has increasingly examined the use of pre-trained speech representations instead of relying only on manually selected acoustic features. One recent approach uses WavLM to obtain deep acoustic representations, followed by a bidirectional recurrent network for modeling temporal information and a multi-head attention mechanism for focusing on emotionally important parts of the speech signal. The approach was evaluated on the RAVDESS and EmoDB datasets and reported strong emotion recognition performance [13]. Another transformer-based approach focuses on combining dynamic and static information from speech. It uses locally adaptive multi-head attention to identify useful information at different temporal regions and global attention to capture information across the complete speech sequence [23]. These approaches show how pre-trained representations and attention mechanisms are being used to extract richer emotional information from speech.

An important issue in speech emotion recognition is whether a system can maintain its performance when it is applied to speakers or datasets that were not used during training. A recent benchmarking study examined transformer-based speech emotion recognition under speaker-independent and cross-corpus conditions. When models were trained on one emotional speech dataset and tested on another, most experimental cases achieved accuracy below 40%. However, when several datasets were combined during training, performance on unseen data improved, with one experiment reaching 58.85% accuracy on a dataset that had not been included during training [24]. These results highlight the difficulty of transferring speech emotion recognition models between different datasets and also indicate that greater diversity in the training data can improve generalization.

Other research has focused on combining speech with textual information. Speech and text provide different types of emotional information, since the audio signal contains characteristics related to vocal expression while text provides information about linguistic meaning. Multimodal approaches have therefore been developed to process these sources together [18]. The contribution of each modality, however, does not need to remain constant for every input. A recent transformer-based multimodal approach uses adaptive weighting so that the influence of the available modalities can change according to the information contained in each case. The approach was evaluated on the IEMOCAP and MELD datasets and showed that dynamically allocating the contribution of the modalities can improve multimodal emotion recognition [17]. This is particularly relevant when one source provides clearer emotional information than another for a specific input.

The use of speech and text together introduces an additional difficulty when the textual information is produced automatically from speech. Automatic Speech Recognition can introduce errors into the transcript, which may then affect the emotion prediction of the text component. Recent work has examined this problem by combining vocal information with noisy ASR-generated transcripts. In this approach, the text component is exposed to different forms of transcription noise during training and is later combined with information obtained directly from the voice. Experiments comparing different sources of text showed that systems using accurate ground-truth transcripts generally had an advantage over ordinary ASR-based approaches, although the proposed noise-aware method was able to achieve strong results even when automatically generated transcripts were used [19]. This work demonstrates that transcription quality is an important consideration in multimodal emotion recognition when textual information is obtained automatically from spoken input.

The transition from controlled emotional datasets to real-world recordings creates further challenges for emotion recognition systems. Research using recordings from a real medical emergency call center compared the same speech emotion recognition architecture on a

commonly used research corpus and on real emergency conversations. The reported unweighted accuracy was 63% on IEMOCAP and 45.6% on the real emergency-call corpus when four emotional classes were considered. The emergency recordings contained a large number of speakers and included greater variability, mixed emotional expressions, short speech segments, and recordings of lower quality [5]. These results show that performance obtained on controlled datasets may decrease when the same type of system is applied to more complex real-world speech.

Background noise is another factor that can affect speech emotion recognition in real environments. Recent work has examined the use of self-supervised speech representations together with speech enhancement in order to improve emotion recognition under noisy conditions. The proposed framework was designed to support both emotion recognition and enhanced speech and reported improved performance when evaluated under previously unseen noise conditions [22]. Research on speech under stress provides a related direction, where conventional acoustic methods, transfer learning, and self-supervised models have been evaluated on datasets containing simulated and real stress conditions [21]. These studies show that real-world emotion-related speech processing involves additional challenges beyond the classification of controlled emotional recordings.

Research related to safety has also examined dangerous linguistic content directly rather than first identifying an emotional category. One recent study evaluated large language models for detecting dangerous statements in authentic text-based clinical conversations. In that work, dangerous speech was defined using expressions related to self-harm, suicidal tendencies, severe self-neglect, or harm to others, and the task was treated as a binary classification between dangerous and non-dangerous content. The evaluated models showed considerable differences in performance, with reported accuracy values ranging from 0.66 to 0.92. The study also identified limited contextual understanding and difficulty in providing more detailed risk levels beyond binary classification as important challenges [25]. This approach differs from emotion recognition because the final decision is based directly on the meaning of the textual content rather than on an intermediate emotional classification.

Overall, previous research shows progress in several areas related to the problem examined in this thesis. Self-supervised speech representations and attention-based models have been used to improve the extraction of emotional information from audio [13], [23], while multimodal approaches have examined how speech and text can provide complementary information and how their contribution can be adjusted dynamically [17]. Other studies have considered problems related to transcription errors [19], speaker and dataset differences [24], background noise [22], stress-related speech [21], and the more difficult conditions found in real emergency recordings [5]. At the same time, safety-related research has also examined the direct identification of dangerous linguistic content [25]. These studies provide important foundations for the present work, while also showing that emotion recognition, multimodal processing, automatic transcription, real-world robustness, and danger-related analysis represent closely connected but different research problems.

2.8 Research Gap & Connection to This Thesis

The studies discussed in the previous sections show that considerable progress has been made in emotion recognition from speech and text. Recent speech emotion recognition systems use deep and self-supervised representations, while multimodal approaches combine information from different sources and investigate different ways of controlling their contribution to the final prediction [13], [17]. At the same time, other studies focus on specific difficulties such as speaker-independent recognition, background noise, automatic transcription errors, and the use of emotion recognition in real-world environments [5], [19], [22], [24]. These studies address important parts of the problem, but they also show that emotion recognition, transcription, multimodal fusion, and safety-related interpretation are often examined through different

research tasks and experimental settings.

One important issue concerns the representation of emotional information from speech. Pre-trained speech models can provide rich representations learned from large amounts of audio, while manually extracted acoustic features can describe specific properties of the speech signal. Recent research has examined both deep representations and combinations of different types of speech information [11], [13]. In this thesis, the audio emotion recognition component follows this direction by combining acoustic features extracted directly from the speech signal with representations obtained from WavLM. The purpose is not to introduce a new speech representation model, but to use information obtained from both predefined acoustic characteristics and a pre-trained speech model within the same emotion classification component.

Another issue appears when the text component receives its input from spoken language rather than from text that is already available. Automatic transcription errors can affect the information provided to an emotion recognition model [19], while multilingual emotion recognition can also be influenced by translation and differences between languages [20]. In the proposed system, spoken input is therefore connected directly with the text emotion recognition stage through automatic transcription. English speech can be processed using the produced transcript, while Greek speech requires an additional translation stage before it is analyzed by the English text emotion model. This allows both languages to be processed by the same overall pipeline, while also introducing the limitations associated with transcription and translation.

Multimodal fusion creates a further design problem because the two modalities may not provide equally reliable information for every input. Previous research has examined adaptive weighting approaches in which the importance of individual modalities changes according to the available information [17]. In this thesis, the speech and text branches are trained independently, and their predictions are combined at decision level. The fusion process does not use a separately trained multimodal classifier. Instead, a dynamic fusion mechanism adjusts the relative contribution of the two branches using predefined information related to the reliability of their predictions. This keeps the two emotion recognition components independent while allowing their influence on the final result to change for each input.

The connection between emotion recognition and danger-related applications also requires careful interpretation. Research using real emergency conversations shows that emotions such as fear and anger can occur frequently in urgent situations, but real recordings also contain substantial variability and are more difficult to process than controlled emotional data [5]. Other safety-related research approaches the problem differently by analyzing textual content directly and classifying it as dangerous or non-dangerous according to indications such as self-harm, suicide, severe self-neglect, or harm to others [25]. In that case, danger is treated as a direct text classification problem rather than as an emotion recognition problem.

The approach followed in this thesis is different from direct dangerous-speech classification. The system does not contain a separate classifier trained on labelled DANGER and SAFE examples. Instead, the final hybrid emotion distribution is used to calculate an indicative emotion-based risk value using predefined emotion weights and a predefined decision threshold. The purpose of this stage is therefore to examine how the emotional information obtained from speech and text can be used as an indication related to potential danger, rather than to claim that emotion alone can confirm whether a real dangerous event is taking place. This distinction is also consistent with the implementation of the thesis, where the fusion is based on predefined reliability rules and the risk stage operates on the resulting hybrid emotion distribution rather than on a separately trained danger model.

Overall, the focus of this thesis is not to introduce a new foundational speech or language model. Instead, it investigates how existing speech and text emotion recognition components can be integrated into a complete processing pipeline that starts from spoken input. The proposed approach combines speech emotion recognition, automatic transcription, text emotion

recognition, multilingual processing, dynamic decision-level fusion, and a final emotion-based risk indication within the same system. In this way, the work connects several research areas discussed throughout this chapter while keeping the final danger-related result as an indicative assessment based on emotional information. The following chapter presents the datasets used for developing and evaluating the individual emotion recognition components.

CHAPTER 3

DATASETS & DATA PREPARATION

3.1 Overview

This chapter presents the datasets used for the development and evaluation of the speech and text emotion recognition components. Since the proposed system processes emotional information from two different modalities, separate datasets are required for the audio and text parts of the system.

For speech emotion recognition, two publicly available emotional speech datasets are used together with additional recordings collected specifically for this thesis. The public datasets are RAVDESS and CREMA-D, both of which contain labelled emotional speech produced by multiple speakers [6], [7]. The additional recordings were created in order to include further speech samples that follow the emotion categories considered in the proposed system.

For text emotion recognition, the GoEmotions dataset is used. GoEmotions contains a large collection of English text samples annotated with a fine-grained set of emotional categories [1]. Since the original datasets don't use exactly the same emotion labels, the available categories are organized into the common seven-class representation used throughout this thesis: neutral, happy, sad, anger, fear, disgust, and surprise.

The following sections describe the main characteristics of each dataset, the emotion categories used from them, and the way the available data are organized for the later training and evaluation stages.

3.2 Speech Emotion Datasets

The audio emotion recognition component requires speech recordings with known emotional labels. For this purpose, two publicly available datasets, RAVDESS and CREMA-D, were used together with a custom dataset created for this thesis. The three sources differ in the number of speakers, the available emotion categories, and the way the recordings were created. Using more than one dataset provides a wider variety of speakers and emotional expressions than relying on a single source.

Since the datasets do not originally contain exactly the same emotion categories, their labels were adapted to a common set of seven emotions used throughout this work: neutral, happy, sad, anger, fear, disgust, and surprise. This common representation makes it possible to combine recordings from different sources during the development of the audio model.

The following subsections present the main characteristics of RAVDESS, CREMA-D, and the custom recordings, together with the way each dataset is used in this thesis.

3.2.1 RAVDESS

RAVDESS is a multimodal emotional database containing speech, song, and facial expressions. The complete dataset contains 7,356 recordings produced by 24 English-speaking professional actors, with an equal distribution of 12 male and 12 female speakers. Each recording is available in different forms, including audio-video, video-only, and audio-only versions. The speech part of the dataset is highly controlled. The actors speak two short English sentences while expressing different emotions. The original speech categories are neutral, calm, happy, sad, angry, fearful, disgust, and surprise. With the exception of neutral, the emotions are expressed at two intensity levels, normal and strong, and each expression is repeated twice [6].

This structure is useful for emotion recognition because the same sentences are spoken by different people and with different emotional expressions. As a result, the linguistic content changes very little, while characteristics such as the speaker, emotion, and emotional intensity change. This makes RAVDESS a useful, controlled dataset for studying emotional information in the voice.

The complete database contains 4,320 speech recordings across the three available modalities. Therefore, the audio-only speech part contains 1,440 WAV recordings. The original audio-only files are provided at a sampling rate of 48 kHz [6]. Only these speech audio recordings are relevant to the audio model developed in this thesis, since song and visual information are not used.

Another useful characteristic of RAVDESS is its filename structure. Each filename contains numerical fields describing the modality, channel, emotion, intensity, statement, repetition, and actor. The emotion and speaker identity can therefore be obtained directly from the filename [6]. The actor number also indicates gender, with odd numbers representing male actors and even numbers representing female actors.

Code	Original RAVDESS emotion
01	Neutral
02	Calm
03	Happy
04	Sad
05	Angry
06	Fearful
07	Disgust
08	Surprise

In this research, these original labels are converted to the seven common categories. In particular, both neutral and calm are mapped to neutral. The remaining categories are converted directly to happy, sad, anger, fear, disgust, and surprise.

RAVDESS also includes a validation process in which participants were asked to recognize the expressed emotions and evaluate aspects such as emotional intensity and genuineness. A total of 247 participants took part in the main validation procedure, while another 72 were involved in the test-retest reliability evaluation. This makes it possible to compare the intended emotion of each recording with the way it was perceived by human listeners. However, the dataset also has some limitations. The emotions were acted and recorded under controlled conditions, which means that they may differ from emotional expressions that occur naturally in everyday speech. In addition, the speech recordings are based on only two different sentences, so the variety of linguistic content is limited [6].

In this thesis, speaker identity is preserved when organizing the RAVDESS data. Actors 01–16 are used during the initial model configuration process, actors 17–20 are used for validation, and actors 21–24 are kept exclusively for final testing. After the model configuration is selected, actors 01–20 are used for final training. In this way, the final RAVDESS test recordings come from speakers that were not used during model development.

Dataset information	RAVDESS
Language	English
Speakers	24 professional actors
Gender distribution	12 male and 12 female speakers
Type of recordings	Acted emotional speech
Total number of recordings	7,356 recordings in the complete database
Audio-only speech samples	1,440 recordings
Spoken sentences	2 different sentences
Emotion categories	8 original categories
Emotional intensity	Normal and strong for all emotions except neutral
Audio file format	WAV
Sampling rate	48 kHz
Data used in this thesis	Audio-only speech recordings
Label adaptation	Calm is included in the neutral category

Table 3.1: Main information about the RAVDESS dataset

3.2.2 CREMA-D

CREMA-D is another multimodal dataset designed for the study of emotional expression. It contains 7,442 clips produced by 91 actors, including 48 male and 43 female speakers. The actors range from 20 to 74 years of age and represent different demographic backgrounds [7].

Compared with RAVDESS, CREMA-D contains a much larger number of speakers. This is useful for speech emotion recognition because the model is exposed to greater variation between individual voices. The recordings were produced by professional actors under the guidance of theatre directors. The actors were asked to express a target emotion while speaking predefined sentences, and the directors helped them produce the required emotional expression [7].

The dataset contains six emotions: happy, sad, anger, fear, disgust, and neutral. Surprise is not included in CREMA-D [7]. Because the classification system developed in this thesis contains seven categories, CREMA-D contributes samples to six of them, while the surprise class is provided by the other speech datasets.

The actors speak 12 different sentences, whose linguistic content was selected to be emotionally neutral. This means that the emotion is mainly communicated through the way the sentence is spoken rather than through strongly emotional words. The use of 12 sentences also provides greater linguistic variety than RAVDESS. Also, this dataset includes variation in emotional intensity. One of the sentences was recorded at low, medium, and high intensity, while the remaining sentences were produced without a predefined intensity level [7]. This provides both stronger and more subtle examples of emotional speech.

One important part of CREMA-D is the way the recordings were evaluated by human listeners.

The clips were presented in audio-only, visual-only, and audio-visual form, and a total of 2,443 people took part in the crowd-sourced evaluation, providing ratings for both the perceived emotion and its intensity. Based on these responses, the recordings were divided into three groups. In matching clips, the emotion recognized by the listeners agreed with the emotion intended by the actor. In non-matching clips, the listeners generally agreed on an emotion, but it was different from the intended one. Finally, ambiguous clips were those for which there was not enough agreement on a single emotional category [7].

The emotion information in the CREMA-D filenames is converted to the common labels of this thesis. The codes ANG, DIS, FEA, HAP, NEU, and SAD correspond to anger, disgust, fear, happy, neutral, and sad respectively. Speaker identity is also obtained from the filename so that the data can be separated according to speakers.

For model development, CREMA-D is divided at speaker level. Approximately 15% of its speakers are used for validation and another 15% for final testing, while the remaining speakers are used for training. The same speaker is not allowed to appear in more than one of these groups. This prevents recordings from the same person from appearing in both training and test data.

Dataset information	CREMA-D
Language	English
Speakers	91 actors
Gender distribution	48 male and 43 female speakers
Age range	20–74 years
Type of recordings	Acted emotional speech
Total number of recordings	7,442 clips
Spoken sentences	12 different sentences
Emotion categories	Happy, sad, anger, fear, disgust, and neutral
Emotional intensity	Low, medium, and high for one sentence; not predefined for the remaining sentences
Human evaluation	Ratings from 2,443 participants
Evaluation formats	Audio-only, visual-only, and audio-visual
Perceptual grouping	Matching, non-matching, and ambiguous clips
Data used in this thesis	Audio recordings
Validation data	15% of the speakers
Test data	15% of the speakers

Table 3.2: Main information about the CREMA-D dataset

3.2.3 Custom Speech Recordings

In addition to RAVDESS and CREMA-D, a custom speech dataset was created specifically for this

thesis. The purpose of this dataset was to provide additional speech samples that follow the same seven emotion categories used by the audio emotion recognition model.

The custom dataset contains 252 original audio recordings from three speakers. Each speaker recorded the same 12 phrases while expressing seven different emotions: neutral, happy, sad, anger, fear, disgust, and surprise. Therefore, each speaker produced 84 recordings, corresponding to 12 phrases for each of the seven emotions. With three speakers, the complete dataset contains 252 recordings.

Dataset information	Custom speech dataset
Speakers	3
Spoken phrases	12 different phrases
Emotion categories	Neutral, happy, sad, anger, fear, disgust, and surprise
Recordings per speaker	84
Total number of recordings	252
Audio file format	WAV
Dataset organization	Speaker and emotion information included in the filenames
Data used in this thesis	Training data for the audio emotion model
Augmented recordings	Used during training
Main purpose	Additional emotional speech data for model development

Table 3.3: Main information about the custom speech dataset

The same phrases were used for all speakers and emotional categories so that the spoken content remained consistent while the expressed emotion changed. The selected phrases are short everyday statements and do not contain an obvious emotional meaning. This makes it possible to focus more on the way the phrase is spoken rather than relying only on emotionally strong words. The phrases used in the custom dataset are presented in Table 3.4.

Phrase	Original Greek text	English translation
P1	Σήμερα το πρωί βγήκα από το σπίτι.	This morning I left the house.
P2	Το βιβλίο βρίσκεται πάνω στο τραπέζι.	The book is on the table.
P3	Αύριο θα πάω στο πανεπιστήμιο.	Tomorrow I will go to the university.
P4	Το αυτοκίνητο είναι παρκαρισμένο έξω.	The car is parked outside.
P5	Έκλεισα το παράθυρο πριν φύγω.	I closed the window before I left.
P6	Το τηλέφωνο βρίσκεται μέσα στην τσάντα.	The phone is inside the bag.
P7	Το μάθημα αρχίζει στις εννέα.	The class starts at nine.

P8	Περπάτησα μέχρι την πλατεία.	I walked to the square.
P9	Άφησα το κλειδί δίπλα στην πόρτα.	I left the key next to the door.
P10	Το λεωφορείο έφτασε πριν λίγο.	The bus arrived a short while ago.
P11	Η συνάντηση θα γίνει το απόγευμα.	The meeting will take place in the afternoon.
P12	Στο δωμάτιο υπάρχει ένα μικρό τραπέζι.	There is a small table in the room.

Table 3.4: Phrases used in the custom speech dataset

The recordings were organized using filenames that include information about the speaker and the expressed emotion. Speaker identifiers such as spk01, spk02, and spk03 are used to distinguish the three speakers. The emotion label is also included in the filename so that the recordings can be automatically assigned to the corresponding category during dataset loading.

The custom recordings are used only as training data for the audio emotion recognition model. They are not included in the final public test set, which is based on unseen speakers from RAVDESS and CREMA-D. This keeps the custom recordings as an additional source of training examples without using the same data for the final quantitative evaluation.

Augmented versions of the custom recordings are also used during training in order to introduce additional variation to the available speech samples. These recordings are treated as modified versions of the original data and are not considered separate test samples.

The main limitation of the custom dataset is its relatively small size and the limited number of speakers. Three speakers cannot represent the full range of voices, accents, speaking styles, and recording conditions that may appear in real-world speech. For this reason, the custom dataset is used to support the larger public datasets rather than as the main source for evaluating the audio model.

3.3 Text Emotion Dataset

The text emotion recognition component requires labelled textual data instead of speech recordings. For this purpose, the GoEmotions dataset was selected. It provides a large collection of English text samples annotated with emotional categories and was created specifically for fine-grained emotion recognition [1]. Each sample consists of written text together with one or more emotional labels. For this thesis, the dataset is processed so that the text model follows the same seven emotion categories used by the audio model: neutral, happy, sad, anger, fear, disgust, and surprise.

3.3.1 GoEmotions

GoEmotions is a large human-annotated dataset created from English comments collected from Reddit. It contains approximately 58,000 comments labelled according to 27 different emotion categories together with Neutral. The comments were selected from popular English-language subreddits. The original Reddit data covered a long period of online discussions, while deleted and non-English comments were removed during the dataset creation process. Additional filtering was also performed to reduce unsuitable content and control the length and distribution of the selected comments [1]. The final samples were limited to relatively short comments, making them suitable for sentence-level emotion classification.

An important characteristic of GoEmotions is that the emotion labels were assigned through human annotation, and a single comment can contain one or more emotional labels. This allows the dataset to represent cases where more than one emotion is expressed in the same text, such

as surprise together with amusement. The original dataset includes 27 emotion categories and Neutral, covering both positive and negative emotions as well as more specific emotional reactions such as joy, sadness, anger, fear, disgust, surprise, admiration, approval, disappointment, gratitude, nervousness, and remorse [1].

For this thesis, only the GoEmotions categories that directly correspond to the seven emotion classes used by the proposed system are selected. These categories are anger, fear, joy, sadness, disgust, surprise, and neutral. The remaining GoEmotions emotion categories are not used for the development of the text classifier. Since the terminology of the dataset is slightly different from the final label names used in this thesis, joy is renamed as happy and sadness as sad, while the other selected categories remain unchanged.

GoEmotions label	Emotion used
Neutral	Neutral
Joy	Happy
Sadness	Sad
Anger	Anger
Fear	Fear
Disgust	Disgust
Surprise	Surprise

Table 3.5 Mapping used in the text dataset

Since the text classifier developed in this thesis performs single-label classification, samples containing more than one of the seven selected emotions are not used. A sample is retained only when exactly one of these seven emotional labels is active. This creates a clear target category for every text given to the classifier. Before the dataset is divided into training, validation, and test sets, samples with missing or empty text are removed. Exact duplicate texts are also removed so that the same sentence is not repeatedly included in the data that are later separated into the different subsets.

The resulting dataset is then divided using a stratified split. Approximately 70% of the samples are used for training, 10% for validation, and 20% for testing. Stratification is used so that the relative distribution of the seven emotion categories remains approximately similar in each subset. The test set is kept separate from training and model selection and is used only for the final evaluation of the text model.

Another characteristic of the processed dataset is that the number of available examples differs between emotional categories. For this reason, balancing is applied only to the training set during the later model-development stage. Larger classes can be limited to a maximum number of training examples, while smaller classes can be increased through random oversampling. The validation and test samples are not modified in this way, so they remain separate evaluation sets.

Dataset information	GoEmotions
Language	English

Type of data	Reddit comments
Dataset size	Approximately 58,000 comments
Annotation	Human-annotated
Original emotion categories	27 emotions and Neutral
Original task	Fine-grained emotion recognition
Label structure	One or more emotions can be assigned to a comment
Categories used in this thesis	Neutral, joy, sadness, anger, fear, disgust, and surprise
Label adaptation	Joy → happy, sadness → sad
Classification used in this thesis	Single-label
Training data	Approximately 70%
Validation data	Approximately 10%
Test data	Approximately 20%
Main purpose in this thesis	Development of the text emotion model

Table 3.6: Main information about the GoEmotions dataset

CHAPTER 4

METHODOLOGY

4.1 System Overview

The proposed system is designed to recognize emotion from both speech and textual information and then use the final emotional representation to provide an indication of possible emotional risk. The complete methodology consists of two separate emotion recognition components, one based on the audio signal, and one based on the textual content of the spoken input. Both components use the same seven emotion categories neutral, happy, sad, anger, fear, disgust, and surprise.

The processing procedure begins with an audio recording. The same recording follows two different paths. In the first path, the speech signal is processed directly by the audio emotion recognition component. In the second path, the spoken content is converted into text and then analyzed by the text emotion recognition component. As a result, the system produces two separate emotion predictions for the same input, one based on characteristics of the voice and another based on the verbal content.

The audio component uses a combination of information obtained from a pre-trained speech representation model and acoustic characteristics extracted directly from the waveform. These two groups of features are combined and provided to an emotion classifier. The text component follows a different procedure, where the transcribed text is processed by a language model that has been adapted to the same seven-class emotion recognition problem.

Since the text emotion model operates on English text, the language of the original speech is also considered. English speech can be processed directly after transcription. When Greek speech is detected, an additional translation stage is used to produce English text before textual emotion recognition is performed.

After the audio and text components produce their predictions, the two outputs are combined through a dynamic fusion mechanism. The contribution of each modality is not fixed for every input. Instead, the method considers both the general performance of the two trained models and information related to the quality of their predictions for the current sample. The final result of this stage is a single distribution across the seven emotion categories. No additional hybrid classifier is trained at this stage.

Finally, the hybrid emotion distribution is provided to an indicative emotion-based danger detector. The detector considers the contribution of the different emotional categories and uses a predefined decision threshold to produce a SAFE or DANGER indication. This stage is based only on the emotional information produced by the system and does not represent a separately trained danger classifier.

4.2 Audio Data Processing & Feature Representation

Before feature extraction, all audio recordings pass through the same preprocessing procedure. Each recording is converted to mono and resampled to 16 kHz. A common format is used so that recordings coming from different datasets are processed in the same way before they are given to the audio model. The duration is limited to a maximum of six seconds, which also gives a common upper limit to the amount of audio processed for each sample. Silence at the beginning and end of the recording is removed, while empty or almost silent recordings are rejected because they cannot provide a useful speech signal for the following feature extraction stage. The same preprocessing settings are also kept during inference so that new recordings are processed in the same way as the recordings used during training.

For the custom speech recordings, an additional data augmentation step is used. The aim of this step is to create slightly different versions of the available recordings and include them as additional training samples. Four types of modification are applied: gaussian noise, small changes in volume, mild reverberation, and slight time stretching. These changes introduce variation in the audio while the original emotion label remains unchanged. One augmented version is created for each selected custom recording. The changes are kept relatively small, for example, the time-stretching rate is 0.98 or 1.02, while the volume is multiplied by 0.85 or 1.15.

After preprocessing, two different groups of features are extracted from every recording. The first group is obtained from the pre-trained WavLM model. In this work, WavLM is used only as a feature extractor and is not fine-tuned, so its pre-trained parameters remain unchanged. For feature extraction, the WavLM representation is obtained using layer 9, and this setting is kept the same during training and inference.

WavLM produces a sequence of representations across the duration of each recording rather than a single representation for the complete audio signal. To convert this sequence into a form that can be used by the classifier, mean and standard deviation pooling are applied across the valid temporal positions. In this way, information from the whole recording is summarized into a single feature representation. An attention mask is also used during this process so that padding values added during processing do not affect the calculated features.

In parallel, a set of acoustic features is extracted directly from the audio signal, including MFCCs and their temporal differences, chroma, zero-crossing rate, RMS energy, spectral centroid, spectral bandwidth, spectral contrast, spectral roll-off, and spectral flatness. For several of these features, statistical values such as the mean and standard deviation are calculated across the recording. These acoustic measurements are then combined with the representation produced by WavLM, so the final input contains information from both the pre-trained speech model and the original audio signal. Before this combined representation is given to the LinearSVC classifier, it is standardized using a scaler fitted on the training data. The same scaler is later applied during inference so that new recordings are processed in the same way as the data used during training.

4.3 Text Data Preparation

Before the text data are used to train the emotion recognition model, they go through a preparation stage. First, samples with missing or empty text are removed because they do not provide valid input for the text classifier. Exact duplicate texts are also removed so that the same sentence is not repeatedly included in the dataset that will later be divided into training, validation, and test sets.

The text model is designed to predict one of the seven emotion categories used in this thesis. For this reason, only samples that correspond to a single selected emotion are kept. Samples that contain more than one of the selected emotions are not used, since the classifier is trained as a single-label classification model and therefore each input needs one clear target emotion.

After this filtering process, the remaining data are divided into training, validation, and test sets. Approximately 70% of the samples are used for training, 10% for validation, and 20% for testing. A stratified split is used so that the proportion of the different emotion classes remains approximately similar in each subset. The validation set is used during model development, while the test set is kept separate and is used only after training and model selection have been completed.

Because the number of available samples is not the same for every emotion, balancing is applied only to the training set. Classes with many examples can be limited, while classes with fewer examples can be increased through oversampling. Class weights are also calculated from the training data so that errors in less represented emotions can have a greater influence during training. These changes are not applied to the validation or test sets, so their original distributions remain unchanged for evaluation.

Finally, the text samples are converted into tokens using the tokenizer of the pre-trained RoBERTa model. Texts that are longer than the maximum accepted length are truncated, while padding is added dynamically when the samples are organized into batches. This converts the original text into the input format required by the model before training

4.4 Model Training & Evaluation Strategy

The audio and text models are trained separately because they process different types of information. The audio model uses features extracted from the speech signal, while the text model uses the written content. However, both models are trained to predict the same seven emotion categories. This common output is necessary because their predictions are later combined during the audio-text fusion stage.

For the audio model, the public speech datasets are divided according to speaker identity rather than individual recordings. This means that recordings from the same speaker are not allowed to appear in both the training and final test data. This separation is used so that the final evaluation is performed on voices that the model has not seen during training. In RAVDESS, actors 01–16 are initially used during the configuration search, actors 17–20 are used for validation, and actors 21–24 are kept only for the final test. After the configuration is selected, actors 01–20 are used for the final training of the model. CREMA-D follows the same speaker-independent idea, with separate groups of speakers used for training, validation, and final testing.

During the development of the audio model, different configurations of the LinearSVC classifier and different contributions of the available training datasets are tested. Each configuration is evaluated using the validation data from RAVDESS and CREMA-D, and the configuration with the best overall validation performance is selected. Using separate validation data for this process allows the model configuration to be chosen without using the final test speakers.

After the configuration has been selected, the final audio training set is created using the available training recordings from RAVDESS and CREMA-D together with the original and augmented custom recordings. The custom recordings are used only as additional training data and are not included in the final test set. A StandardScaler is fitted using the final training data, and the LinearSVC classifier is then trained using the selected configuration. The final evaluation is performed on the RAVDESS and CREMA-D speakers that were kept separate from both training and model selection. In this way, the reported results represent the performance of the model on unseen speakers.

The text model follows a different training procedure. The pre-trained RoBERTa model is fine-tuned using the prepared training set, while the validation set is used to monitor its performance during training. Early stopping is applied so that training can stop when the validation performance no longer improves, instead of continuing unnecessarily. Macro-F1 is used to select the best model checkpoint because it considers the performance of all emotion classes without allowing the larger classes to dominate the model selection.

After training and model selection have been completed, the text model is evaluated on the separate test set. This test set is not used for training, early stopping, or the selection of the best checkpoint, so it remains independent from the model development process. This provides a clearer indication of how the trained model performs on text samples that were not used to adjust it.

For both the audio and text models, evaluation is not based only on overall accuracy. Accuracy, Balanced Accuracy, Macro-F1, Weighted-F1, class-level results, and confusion matrices are also considered. These measures provide a more complete view of the model performance because the emotion categories do not contain the same number of samples. For example, Macro-F1 gives the same importance to every emotion class, while Weighted-F1 also considers the number of samples belonging to each class. The confusion matrix is used to examine which emotions are correctly recognized and which ones are more often confused with each other.

4.5 Speech Transcription & Translation

The complete system receives an audio recording as input, but the text emotion model requires written text. For this reason, the spoken content must first be converted into text. Whisper is used for this stage, connecting the original speech recording with the text emotion recognition part of the system.

Since the system is designed to process both Greek and English speech, the language of each recording is detected before transcription. Whisper provides language probabilities, and the system compares the probabilities for Greek and English to select between these two languages. The selected language is then used during transcription. A corresponding initial prompt is also provided to Whisper in order to guide the transcription process according to the detected language.

When the detected language is English, the resulting transcript can be given directly to the text emotion model. When Greek is detected, an additional translation stage is required because the text emotion model was trained using English text. Two possible English versions are produced. The first is created by translating the Greek transcript into English, while the second is produced by asking Whisper to translate the original audio directly into English. Using two different translation paths makes it possible to compare their results instead of depending on only one translation.

If both English translations are available, their similarity is also examined. This provides an indication of how closely the two translation methods agree. Duplicate translations are removed so that the same text is not processed more than once. Each available text candidate is then passed through the text emotion model, and the candidate that gives the clearest emotion prediction is selected. The selection considers how strong the highest prediction is, how clearly it is separated from the next emotion, and the overall uncertainty of the prediction.

Finally, the selected text is given a reliability value before it is used in the audio-text fusion stage. The reliability can be reduced when the transcript is very short or repetitive, when translation is required, when the two translation methods disagree, or when the text emotion prediction is uncertain. This allows the following fusion stage to take into account that the textual information may not always be equally reliable for every recording.

4.6 Audio-Text Fusion Methodology

After the audio and text models produce their emotion predictions, their outputs are combined to obtain one final emotion result. A new hybrid model is not trained for this stage. Instead, a dynamic fusion mechanism based on predefined rules is used. The contribution of audio and text is therefore not fixed for every recording but can change depending on the information available for each input.

The process first considers the general performance of the two trained models. Their evaluation results, based on Accuracy and Macro-F1, are used to provide an initial indication of how much each model should contribute to the fusion. These initial weights are then adjusted according to the current prediction. In this way, the fusion does not depend only on the overall performance of each model but also considers how clear its prediction is for the specific recording being processed.

To estimate the quality of each prediction, the system considers the strongest predicted emotion, how clearly it is separated from the next possible emotion, and the uncertainty of the complete emotion distribution. For the text branch, the reliability of the transcription and translation process is also considered. This is important because the textual result may become less reliable when the transcript is very short, when translation is required, or when the available translations show low agreement.

When the audio and text models predict different emotions, their prediction quality is compared and the modality with the clearer result can receive a greater contribution. At the same time, the weights are limited so that one modality does not completely remove the other from the fusion process. When translated text is used, the contribution of the text branch is further limited, especially when the available translations show very low agreement. These adjustments are used because transcription and translation can introduce additional uncertainty into the textual information.

Once the final audio and text weights have been determined, the two emotion distributions are combined using logarithmic fusion. If both models agree on the same strongest emotion, that emotion receives a small additional reinforcement. If a valid text result is not available, the system uses only the audio prediction instead of attempting to combine unreliable or missing textual information. The final output of this stage is one emotion distribution across the seven emotion categories, which is then passed to the emotion-based danger detector.

4.7 Emotion-Based Danger Detector

After the audio and text predictions have been combined, the final hybrid emotion distribution is given to an emotion-based danger detector. The detector uses the information from all seven emotion categories rather than considering only the emotion with the highest value. In this way, more than one emotion can influence the final decision when several emotions have a noticeable presence in the hybrid result.

The emotions do not contribute equally to the detector. Fear has the highest contribution, followed by anger, while sadness and disgust have smaller contributions. Neutral, happy, and surprise do not increase the danger indication. These contributions are predefined design parameters and are applied to the complete hybrid emotion distribution. This allows the detector to consider the overall emotional state produced by the system instead of depending only on a single predicted emotion.

The result produced from these emotional contributions is then compared with a predefined decision threshold of 0.40. When the result reaches or exceeds this threshold, the detector produces a DANGER indication. Otherwise, the output is SAFE. The threshold is kept fixed and is not learned during the training of either the audio or text model.

The danger detector is not a separately trained binary classifier. It operates only on the final hybrid emotion distribution and uses the predefined emotion contributions and decision threshold to produce the final indication. For this reason, its output represents an emotion-based indication of possible danger and should not be interpreted as proof that a real dangerous situation is occurring.

CHAPTER 5

MODEL TRAINING & SYSTEM IMPLEMENTATION

5.1 Overview

This chapter describes the implementation and training of the different parts of the proposed system. The system includes an audio emotion model, a text emotion model, a speech transcription and translation stage, a fusion process, and the final emotion-based danger detector.

The audio and text models are trained separately because they use different types of input. The audio model analyzes the speech signal, while the text model analyzes the words of the same recording. Both models predict the same seven emotion categories so that their results can later be combined.

After the two models produce their predictions, Whisper is used to convert speech into text. When the speech is in Greek, the text is translated into English before it is processed by the text model. The audio and text predictions are then combined through the dynamic fusion process. Finally, the hybrid emotion result is passed to the danger detector, which produces the final SAFE or DANGER indication. The complete system can be used with both stored audio files and recordings from a microphone.

5.2 Software Environment and Libraries

The proposed system was implemented in Python. Different libraries were used for audio processing, model training, text processing, and evaluation.

PyTorch and the Transformers library were used for the pre-trained WavLM and RoBERTa models. Librosa was used for audio preprocessing and acoustic feature extraction. Scikit-learn was used for the LinearSVC classifier, feature standardization, and the calculation of evaluation metrics.

Whisper was used for speech transcription, while NumPy was used for numerical operations. Matplotlib was used to create plots and confusion matrices for the evaluation of the models.

The main libraries and tools used in the implementation are presented in Table 5.1.

Library / Tool	Main use
Python	Main implementation language
PyTorch	Deep learning operations
Transformers	WavLM and RoBERTa models
Librosa	Audio processing and feature extraction
Scikit-learn	LinearSVC, preprocessing and evaluation
Whisper	Speech transcription
NumPy	Numerical operations

Matplotlib	Plots and confusion matrices
------------	------------------------------

Table 5.1: Main libraries and tools used in the implementation

5.3 Audio Emotion Model Training

The audio emotion model was trained using recordings from RAVDESS, CREMA-D, and the custom speech dataset. The preprocessing and feature extraction procedure described in Section 4.2 was followed for all recordings. In the implementation, the audio was converted to mono, resampled to 16 kHz, and limited to a maximum duration of six seconds.

For the WavLM representation, the pre-trained WavLM-base-plus model was used with its parameters frozen. The representation from layer 9 was selected and mean and standard deviation pooling were applied, resulting in 1,536 features for each recording. In parallel, 211 acoustic features were extracted using Librosa. These included MFCC-based features, chroma, zero-crossing rate, RMS energy, and different spectral characteristics. The two feature groups were concatenated into a final 1,747-dimensional feature vector.

WavLM-base-plus is a pre-trained model used to extract useful information from speech. In this work, it does not directly predict one of the seven emotions. Instead, it produces speech representations that are later combined with the acoustic features extracted from the same recording. During training, the WavLM parameters remain frozen, so the model itself is not updated and is used only as a feature extractor. WavLM was selected because it was developed for different speech-processing tasks and has also been evaluated for emotion recognition, making it a suitable choice for obtaining speech representations for this work [8]. Its pre-trained representations also allow useful information to be extracted from the recordings without having to train a speech representation model from the beginning. The final emotion classification is then performed by the audio classifier.

Before classification, the final feature vectors were standardized using StandardScaler. The scaler was fitted on the training data and the same fitted scaler was then used for validation, testing, and inference. The standardized features were then provided to a LinearSVC classifier. LinearSVC is the final classifier of the audio emotion model and uses the combined WavLM and acoustic features to classify each recording into one of the seven emotion categories. It was selected because the audio information had already been converted into a fixed-length feature vector that could be provided directly to a separate classifier. Its main parameter, C, could also be tested using the validation data before the final model was trained. This allowed the feature extraction and emotion classification stages to remain separate, while the classifier configuration was selected using the validation procedure described below.

During model development, different values of the LinearSVC parameter C and different source weights for CREMA-D and the custom dataset were tested. Six values were examined for C (0.001, 0.003, 0.005, 0.010, 0.020, and 0.030), together with four CREMA-D weights (0.60, 0.75, 0.90, and 1.00) and four custom-data weights (1.50, 2.00, 2.50, and 3.00). This resulted in 96 different configurations. Each configuration was evaluated on the speaker-independent validation data from RAVDESS and CREMA-D. The configurations were compared using a selection criterion based on 55% of the mean validation accuracy and 45% of the harmonic mean of the Macro-F1 results. The final selected configuration used C = 0.001, a CREMA-D weight of 0.9, and a custom-data weight of 2.5. These values were therefore selected from the validation search rather than being chosen directly. Class-balanced sample weights were also calculated separately from the training data and combined with the dataset source weights in order to take class imbalance into account.

The configuration search was performed without using the final test speakers. For RAVDESS, actors 01–16 were used during configuration search, actors 17–20 for validation, and actors 21–24 were reserved for final testing. After the configuration was selected, actors 01–20 were used

for final training. CREMA-D was also divided according to speaker identity, with approximately 15% of the speakers used for validation and 15% reserved for final testing. The custom recordings, including their augmented versions, were used only during training and were not included in the final public test set. The augmented custom recordings received 25% of the source weight assigned to the original custom recordings.

5.4 Text Emotion Model Training

The text emotion model was trained using the GoEmotions dataset. Following the preparation procedure described in Section 4.3, only the seven source categories required by the system were selected: anger, fear, joy, sadness, disgust, surprise, and neutral. The labels joy and sadness were renamed as happy and sad, while the remaining selected labels were kept unchanged. Samples containing more than one of these selected emotions were excluded because the model was trained for single-label classification.

Before the dataset was divided, samples with missing or empty text and exact duplicate texts were removed. The remaining samples were separated using stratified splitting, with approximately 70% used for training, 10% for validation, and 20% for testing. A random seed of 42 was used for the split. The validation and test sets were kept unchanged after this stage, while balancing was applied only to the training set.

To reduce the effect of class imbalance during training, classes containing more than 10,000 samples were limited to this number, while classes containing fewer than 1,000 samples were increased through random oversampling. The oversampling procedure reused existing training samples rather than creating new synthetic text. In addition, class weights were calculated from the final training set according to the frequency of each emotion class. Less represented classes received larger weights, while more frequent classes received smaller weights. A weight power of 0.35 was then applied to control how strongly this correction affected the training process. Using a value below 1 reduces the difference between the calculated class weights, so the imbalance is corrected without giving excessively large importance to the least represented classes. The resulting weights were used in the loss function during model training. Oversampling and class weighting therefore have different roles. Oversampling changes how often samples from less represented classes are presented to the model, while class weights change how strongly their errors contribute to the training loss. Neither operation was applied to the validation or test sets.

After the balancing procedure, the final training set contained 17,070 samples, while the unchanged validation and test sets contained 2,463 and 4,926 samples, respectively. These sets were then used for model training, model selection, and final evaluation.

RoBERTa-base is a pre-trained Transformer model used for processing textual information. In this work, it is used as the main model of the text emotion recognition branch and is fine-tuned to classify the input text into the seven emotion categories. Unlike WavLM in the audio branch, the RoBERTa parameters are updated during training so that the model can adapt its pre-trained language representations to the emotion classification task. RoBERTa was selected because it can represent words according to their surrounding context and can be fine-tuned for text classification tasks [9]. This is useful for emotion recognition, where the meaning of a word or sentence can depend on the context in which it appears.

Training was configured for a maximum of nine epochs with a learning rate of 1.3×10^{-5} . The relatively small learning rate was used because RoBERTa is already a pre-trained model and its parameters are updated during fine-tuning rather than learned from the beginning. This allows the model to adapt gradually to the emotion classification task. A training batch size of 4 was used, together with gradient accumulation over four steps. As a result, the gradients from several smaller batches are accumulated before the model parameters are updated. The evaluation batch size was set to 8. During validation, the model parameters were not updated, and the validation

data were used only to monitor model performance.

The AdamW optimizer was used during fine-tuning together with a linear learning-rate schedule. Additional training settings included a weight decay of 0.01, a warm-up ratio of 0.08, and label smoothing of 0.015. Weight decay was used as a regularization setting during optimization, while the warm-up ratio means that the learning rate increases gradually during the first 8% of the training steps before continuing with the linear schedule. Label smoothing slightly reduces the certainty of the target labels during the calculation of the loss instead of treating the correct class as an absolute target. These values were predefined training settings of the implemented model and were kept fixed during the final training process.

The model was evaluated on the validation set at the end of each epoch, with Macro-F1 used to select the best checkpoint. Early stopping was applied with a patience of three evaluation rounds, meaning that training stopped when the validation Macro-F1 did not improve for three consecutive evaluations. Although the maximum number of epochs was set to nine, the final training run stopped after five epochs. After training stopped, the checkpoint with the highest validation Macro-F1 was automatically restored, and this model was then used for the final evaluation on the independent test set.

5.5 Speech Transcription and Translation

The transcription stage was implemented using the Whisper small model [10]. Whisper is a pre-trained model that can process speech in different languages and perform language detection, transcription, and translation. It was selected because the system needs to process both Greek and English recordings and these tasks can be handled within the same model. The small version was used as the selected Whisper configuration for the current implementation. Since the proposed system supports Greek and English speech, Whisper was first used to determine which of these two languages was more likely for each recording.

For transcription, a temperature of 0.0 and a beam size of 5 were used. The temperature controls the amount of randomness during the generation of the transcript. Setting it to 0.0 makes the decoding deterministic, so Whisper selects the most likely transcription instead of introducing random variation between executions. A beam size of 5 means that Whisper keeps up to five possible decoding sequences under consideration during the transcription process and compares them before selecting the final result. This allows the transcription to be selected from several possible sequences rather than following only the first available prediction. The detected language was also given directly to Whisper together with a short initial prompt. The Greek prompt was “Η ηχογράφηση είναι στα ελληνικά.”, while the English prompt was “The recording is in English.” The prompt provides an additional indication of the expected language during transcription. These settings were predefined implementation choices and were kept fixed for all recordings.

When English speech was detected, the resulting transcript was passed directly to the text emotion model. For Greek speech, an English version was required because the text classifier was trained on English data. Two translation methods were used. The first translated the Greek transcript into English using Google Translate, while the second used Whisper to translate the original audio directly into English.

When both translation methods produced a result, their similarity was calculated to obtain a translation agreement value between 0 and 1. This comparison considered both the overall similarity of the two sentences and the words that they had in common. A value closer to 1 indicates that the two translation methods produced more similar English text.

Each available English candidate was also processed by the text emotion model before the final text was selected. For every candidate, a quality value was calculated using three parts of the emotion prediction. The highest predicted emotion probability had the largest contribution, with a weight of 50%, because it shows how strongly the model supports its most likely emotion. The

difference between the first and second highest probabilities contributed 35%, since a larger difference indicates that the model separates its first choice more clearly from the next possible emotion. The remaining 15% was based on the overall uncertainty of the complete emotion distribution, using the entropy of the prediction. Lower entropy corresponds to a more concentrated and clearer prediction. These three parts were combined to form the candidate quality value. Short texts and translated texts received a small reduction, and the candidate with the highest final value was selected as the input of the text branch.

Finally, the selected text was assigned a reliability value before the audio-text fusion stage. This value can be reduced when the text is very short or repetitive, when translation is required, when the two translations have low agreement, or when the emotion prediction of the text model is uncertain. The exact reliability values are used during the fusion process described in Section 5.6, where they determine how strongly the text prediction contributes to the final hybrid result.

5.6 Hybrid System Implementation

The hybrid stage combines the emotion distributions produced by the audio and text models. No additional multimodal classifier is trained during this stage. Instead, the contribution of the two already trained models is adjusted dynamically for every recording according to their overall performance and the quality of their current predictions.

Before fusion, the output of the LinearSVC audio classifier must be converted into a normalized emotion distribution. Since LinearSVC produces decision values rather than probabilities, a temperature value of 1.30 is applied before normalization. Increasing the temperature slightly reduces large differences between the decision values and produces a smoother distribution across the seven emotions. This is useful because the fusion stage requires a complete distribution rather than only the strongest audio prediction. The value 1.30 was used as a predefined implementation setting to apply a mild smoothing effect without making the audio distribution almost uniform.

The first weighting step is based on the overall evaluation performance of the two models. For each model, 40% of Accuracy and 60% of Macro-F1 are used. Accuracy is included because it represents the overall percentage of correct predictions, while Macro-F1 gives the same importance to every emotion class. Macro-F1 therefore receives the larger weight of 60%, since the emotion classes are not equally represented and the fusion should not depend mainly on performance in the larger classes. Accuracy still contributes 40%, so the general predictive performance of the model is also taken into account. The resulting audio and text values are normalized so that their initial weights add up to 1. These percentages are predefined design values and were not selected through a separate optimization process.

The general performance of a model is not enough to decide how much it should contribute to every individual recording. For this reason, the quality of the current prediction is also calculated. Three characteristics are used: the highest emotion value, the margin between the first and second highest emotions, and the entropy of the complete emotion distribution. Their contributions are 40%, 35%, and 25%, respectively. The highest emotion value receives the largest contribution, 40%, because it shows how strongly the model supports its main prediction. The margin contributes 35% because a large difference between the first and second emotion indicates that the first prediction is more clearly separated from its closest alternative. The remaining 25% is based on entropy and represents the uncertainty of the complete seven-emotion distribution. In this case, lower entropy increases the quality value because the prediction is more concentrated. The three values were given similar importance because all of them describe a different part of prediction certainty, while slightly more importance was assigned to the strongest emotion and its separation from the second emotion.

The base model weight and the prediction quality are then combined. A 30% baseline contribution is retained, while 70% of the adjustment depends on the quality of the current

prediction. The 30% part prevents a model with an uncertain prediction from immediately losing almost all of its contribution, since its general evaluation performance is still relevant. The larger 70% part allows the current recording to have a stronger influence on the final weight. In this way, the system is not completely fixed according to the average model performance but can react when one model produces a much clearer prediction for a particular recording. These values were chosen as predefined settings to give greater importance to the current sample while still keeping part of the original model contribution. For the text branch, the prediction quality is also multiplied by the reliability of the transcription and translation process before this calculation is performed.

When the strongest audio and text emotions are different, the system checks whether one prediction is clearly better than the other. A minimum quality difference of 0.10 is required before an additional adjustment is made. This threshold prevents very small differences between the two quality values from changing the weights unnecessarily. If the audio quality is at least 0.10 higher, its contribution is multiplied by 1.30. If the text quality is at least 0.10 higher, its contribution is multiplied by 1.20. These factors provide a moderate increase rather than replacing the other modality completely. A slightly larger factor is used for audio because the text branch may also contain uncertainty introduced by transcription or translation. The exact threshold and multiplication factors are predefined implementation parameters and were not obtained through a separate parameter search.

The margins of the two predictions are also compared when the models disagree. If one modality has a prediction margin at least 0.12 larger than the other, its contribution receives a second small increase. The audio contribution is multiplied by 1.15, while the text contribution is multiplied by 1.10. The value 0.12 is used as a minimum separation so that very small margin differences are ignored, while the multipliers are kept close to 1 because this is intended only as an additional adjustment and not as the main weighting decision. These values are again predefined design settings.

After the adjustments, the modality weights are limited so that neither model completely dominates the fusion. Under normal conditions, the audio weight is restricted between 20% and 80%, while the text weight receives the remaining contribution. For translated text, the maximum text weight is reduced to 30%, and when the agreement between the two translations is below 0.18, it is further limited to 16%. These limits reduce the influence of uncertain translated text while still allowing both modalities to contribute.

The final audio and text distributions are combined using logarithmic fusion. If both models predict the same strongest emotion, that emotion receives a small agreement bonus of 1.04 before the final normalization. This value is kept close to 1 so that agreement provides only a small reinforcement. If no valid text result is available, the audio distribution is used directly.

All numerical values used in the fusion mechanism are predefined design parameters. They control the influence of model performance, prediction quality, modality disagreement, and translation uncertainty. They were not learned by a separate multimodal classifier or optimized on a labelled multimodal dataset, and therefore they are treated as implementation choices rather than optimal values.

5.7 Emotion-Based Danger Detection

After the hybrid emotion distribution is produced, it is passed to the final emotion-based danger detector, which produces either a SAFE or DANGER indication. The detector uses the complete hybrid emotion distribution rather than only the emotion with the highest value. Therefore, several emotions can contribute to the final decision at the same time.

Different contributions are assigned to the seven emotions. Fear receives the highest weight of 1.00, while anger receives 0.70. Smaller contributions are assigned to sadness 0.30 and disgust 0.20. Neutral, happy, and surprise have a weight of 0.00 and therefore do not increase the detector

value. These values were predefined to make fear the strongest contributor, followed by anger, while sadness and disgust have a more limited effect.

Each value represents the probability of the corresponding emotion in the final hybrid distribution. For example, the weight 1.00 means that the complete fear probability contributes to the detector, while anger contributes 70% of its probability. Sadness and disgust contribute only 30% and 20%, respectively. This weighting structure was used so that the final decision is influenced mainly by strong fear and anger predictions, while the presence of sadness or disgust has a smaller effect.

The final detector value is compared with a fixed threshold of 0.40. A value equal to or greater than 0.40 produces a DANGER indication, while a lower value produces SAFE. The relationship between the threshold and the emotion weights also controls how easily each emotion can affect the decision. For example, fear alone can reach the threshold when its probability is approximately 40%, while anger would require approximately 57% because its weight is 0.70. In contrast, sadness and disgust cannot reach the threshold by themselves, since their maximum possible contributions are 0.30 and 0.20.

The threshold of 0.40 and the emotion weights are predefined implementation choices. They were not obtained through the training or optimization of a separate DANGER/SAFE classifier, since no danger-labelled dataset was used for this stage. The output should therefore be interpreted as an indicative emotion-based danger detection result rather than as confirmation that a real dangerous situation is occurring.

5.8 Live Microphone Inference

In addition to stored audio files, the final system was implemented to accept speech directly from a microphone. For each execution, a recording of six seconds is captured. This duration was used as a fixed implementation setting so that each live test provides a short and consistent speech input. It was not selected through a separate optimization process.

The recording uses the default sampling rate of the available microphone and is captured as a mono signal using 32-bit floating-point values. It is then saved as a WAV file using 16-bit PCM format. The original sampling rate is kept during recording, since the audio preprocessing stage later resamples the signal to the 16 kHz required by the audio emotion model. The implementation also checks that the recording is not empty or almost silent before continuing with the remaining processing stages.

After the recording is completed, the same processing pipeline described in the previous sections is applied automatically. The recorded file is processed by the audio emotion model, while Whisper performs language detection and transcription. The resulting text is prepared for the text emotion model according to Section 5.5, and the audio and text outputs are combined using the dynamic fusion mechanism of Section 5.6. The final hybrid emotion distribution is then given to the emotion-based danger detector described in Section 5.7.

For each recording, the implementation displays the original transcript, the text used by the text emotion model, the predictions of the audio and text models, the hybrid emotion prediction, the final detector value, and the resulting SAFE or DANGER indication. This makes it possible to inspect the main outputs of the complete system directly when new speech is recorded through the microphone.

CHAPTER 6

EXPERIMENTAL SETUP & RESULTS

6.1 Experimental Setup

This chapter presents the experimental evaluation of the main components of the proposed system. The evaluation is divided into four experiments, allowing the different stages of the system to be examined separately.

The first experiment evaluates the audio emotion recognition model using speech recordings from speakers that were not included during training. The second experiment evaluates the text emotion recognition model on its separate test set. These two experiments provide the quantitative evaluation of the individual emotion recognition models.

The third experiment examines the behavior of the hybrid emotion recognition system, where the audio and text predictions are combined using the dynamic fusion process described in Chapter 5. Representative recordings are used to compare the individual predictions with the final hybrid result. A separate quantitative accuracy is not reported for this stage because a labelled paired multimodal test set was not used in this work.

Finally, the fourth experiment examines the emotion-based danger detector. Representative examples are used to show how different hybrid emotion distributions affect the final SAFE or DANGER indication. Since no independent test set containing ground-truth SAFE and DANGER labels was used, this experiment examines the behavior of the detector rather than reporting its classification accuracy.

The experiments presented in this chapter are summarized in Table 6.1.

Experiment	Description
Experiment 1	Audio emotion recognition
Experiment 2	Text emotion recognition
Experiment 3	Hybrid emotion recognition using stored WAV recordings
Experiment 4	Emotion-based danger detection using live microphone recordings

Table 6.1: Overview of the experimental evaluation

6.2 Evaluation Metrics

The quantitative evaluation of the audio and text emotion recognition models was based on accuracy, balanced accuracy, precision, recall, F1-score, macro-F1, and weighted-F1. Accuracy represents the proportion of test samples that are classified correctly. Balanced accuracy gives equal importance to the performance of each class and is useful when the number of samples is different between the emotion categories [26].

Precision shows how often a predicted emotion is correct, while recall shows how many of the actual samples belonging to an emotion are correctly identified. The F1-score combines precision and recall into a single measure. Macro-F1 calculates the F1-score separately for each emotion and then takes their average, giving the same importance to every class. In contrast, weighted-F1 takes the number of samples in each class into account. For this reason, comparing the two values helps show whether the overall performance is affected by classes containing more samples [27].

Confusion matrices were also used for the audio and text models. They provide a more detailed view of the predictions by showing which emotions are correctly recognized and which emotion categories are more frequently confused with each other.

The hybrid emotion recognition stage was examined using representative examples that compare the audio, text, and final hybrid predictions. The emotion-based danger detector was evaluated in the same way by examining its detector values and corresponding SAFE or DANGER outputs. Quantitative classification metrics were not calculated for these two stages because no ground-truth labelled test set was available for the complete multimodal system or for the SAFE/DANGER decision.

6.3 Experiment 1: Audio Emotion Recognition

The first experiment evaluates the audio emotion recognition model described in Section 5.3 on its final speaker-independent test set.

The final evaluation was performed on 1,622 samples from the RAVDESS and CREMA-D datasets. The separation between training and testing was based on speaker identity, meaning that the speakers included in the test set were not used during training. This setup provides a more demanding evaluation, since the model is required to recognize emotions from voices it has not previously encountered. The custom recordings used during model development were not included in this final test set.

Out of the 1,622 test samples, 1,196 were classified correctly. The model achieved an accuracy of 73.74%, a balanced accuracy of 76.18%, a macro-F1 of 73.80%, and a weighted-F1 of 73.01%.

Evaluation Set	Samples	Correct Predictions	Accuracy	Balanced Accuracy	Macro-F1	Weighted-F1
Audio test set	1622	1196	73.74%	76.18%	73.80%	73.01%

Table 6.2: Evaluation results of the audio model

The relatively close values of accuracy, balanced accuracy, macro-F1, and weighted-F1 show that the overall performance of the model remains fairly consistent across the different evaluation measures. However, the overall scores alone do not show which emotions are easier or more difficult to recognize, so the results are also examined separately for each emotional category.

The performance for each emotional category is presented in Table 6.3.

Emotion	Precision	Recall	F1-score	Support
anger	0.86	0.90	0.88	259
disgust	0.73	0.78	0.75	259
fear	0.68	0.62	0.65	259
happy	0.76	0.66	0.71	259
neutral	0.72	0.92	0.81	263
sad	0.67	0.49	0.57	259
surprise	0.69	0.97	0.81	64

Table 6.3: Detailed results per emotional category for the audio model

The results show noticeable differences between the seven emotions. Anger achieved the highest F1-score, at 0.88, with a precision of 0.86 and a recall of 0.90. Neutral also performed well, reaching an F1-score of 0.81 and a recall of 0.92. Surprise achieved the same F1-score of 0.81 and a particularly high recall of 0.97. However, the surprise category contains only 64 test samples, considerably fewer than the other categories, so this result should be interpreted with some caution.

Fear and sad were the most difficult emotions for the model, with F1-scores of 0.65 and 0.57 respectively. Fear achieved a recall of 0.62, while sad had a recall of 0.49. The performance of fear is especially relevant to the final danger detection stage, since fear has the highest contribution to the emotion-based danger detector.

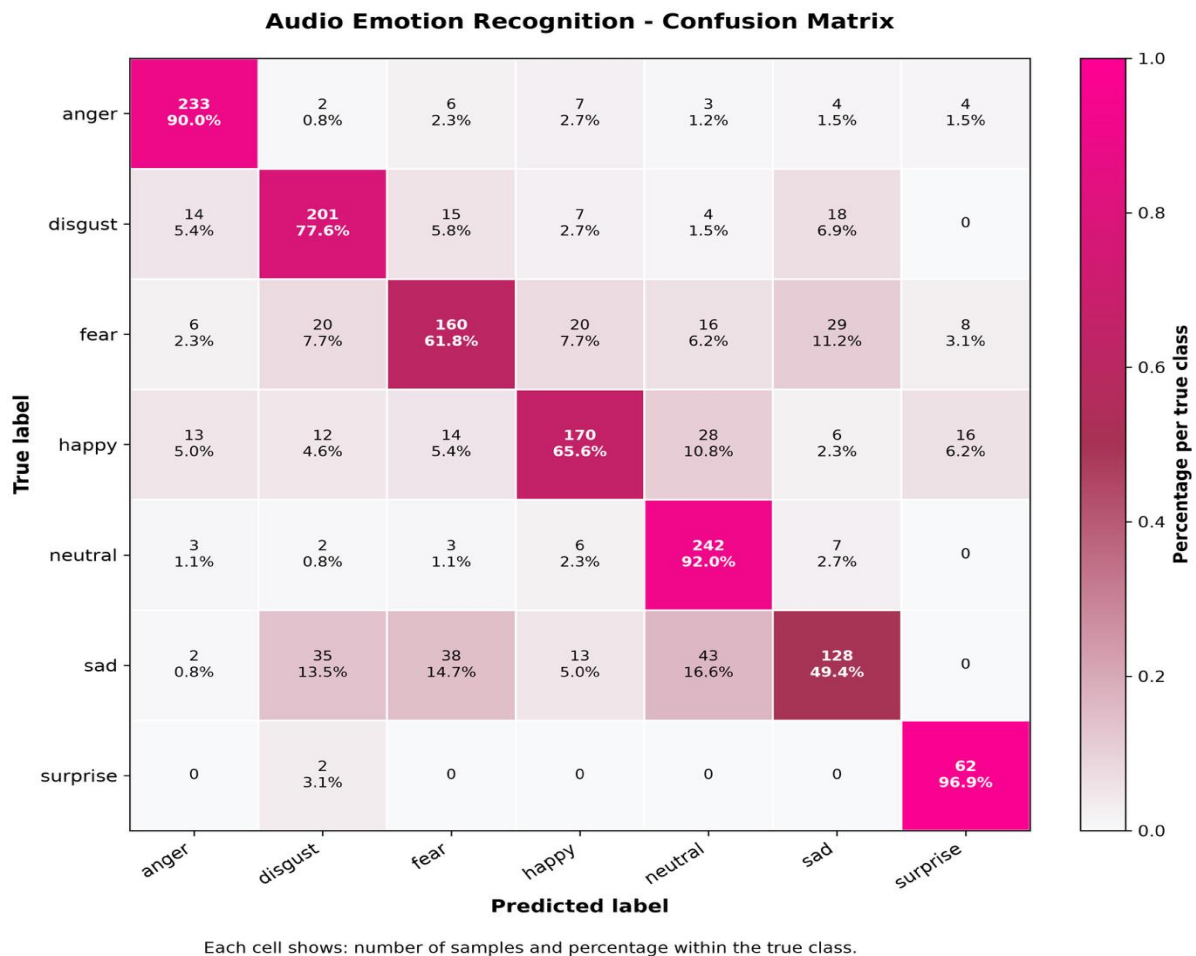


Figure 6.1: Confusion matrix of the speech emotion recognition model on the test set

The confusion matrix in Figure 6.1 provides a more detailed view of the classification results. Anger, neutral, and surprise show the highest percentages of correctly recognized samples, with 90.0%, 92.0%, and 96.9%, respectively. Fear is correctly recognized in 61.8% of the test samples, with its largest confusion occurring with sad 11.2%. Smaller confusion is also observed with disgust and happy, both at 7.7%. Sad is the most difficult category, with only 49.4% of its samples correctly recognized. Its main confusions are with neutral 16.6%, fear 14.7%, and disgust 13.5%.

6.4 Experiment 2: Text Emotion Recognition

The second experiment evaluates the text emotion recognition model described in Section 5.4 on its independent test set. After the training-only balancing procedure, the final dataset contained 17,070 training samples, 2,463 validation samples, and 4,926 test samples. The validation and

test sets were kept unchanged during balancing.

Evaluation Set	Samples	Accuracy	Balanced Accuracy	Macro-F1	Weighted-F1
Text test set	4926	71.50%	62.13%	59.01%	72.00%

Table 6.4: Evaluation results of the text model

Table 6.4 summarizes these results for the complete text-model test set.

The difference between macro-F1 and weighted-F1 is noticeable and should be interpreted together with the class distribution. Neutral accounts for 3,094 of the 4,926 test samples and therefore has a greater influence on weighted metrics. In contrast, macro-F1 gives equal importance to all seven emotional categories and provides a clearer indication of the variation in performance between classes.

A more detailed view of the performance for each emotion is shown in Table 6.5.

Emotion	Precision	Recall	F1-score	Support
neutral	0.85	0.79	0.81	3094
happy	0.59	0.73	0.65	422
sad	0.54	0.63	0.58	337
anger	0.57	0.48	0.53	405
fear	0.52	0.65	0.58	153
disgust	0.41	0.47	0.44	234
surprise	0.50	0.60	0.55	281

Table 6.5: Detailed results per emotional category for the text model

The neutral category shows the strongest performance, with a precision of 0.85, recall of 0.79, and F1-score of 0.81. This is also the largest category in the test set, containing 3,094 samples.

Happy achieves an F1-score of 0.65, while sad and fear both reach 0.58. Surprise follows with 0.55, anger with 0.53, and disgust shows the lowest F1-score at 0.44. Surprise reaches an F1-score of 0.55, while anger achieves 0.53. Disgust remains the most difficult category for the text model, with an F1-score of 0.44. These results show that the model does not perform equally across all seven emotional categories.

Fear achieves a recall of 0.65, meaning that approximately two thirds of the actual fear samples are correctly identified. Its precision is 0.52 and its F1-score is 0.58. This result is particularly relevant to the final danger detection stage, since fear has the strongest contribution to the emotion-based danger detector.



Figure 6.2: Confusion matrix of the text emotion recognition model on the test set

The confusion matrix provides a more detailed view of the class-specific behavior of the text model. Neutral is correctly recognized in 78.5% of the test samples, while happy reaches 72.5%. Sad and fear are correctly recognized in 62.6% and 65.4% of cases respectively. Surprise reaches 60.5%, whereas anger and disgust present greater difficulty, with recall values of 48.4% and 47.0%.

Another noticeable pattern is that samples from several emotional categories are classified as neutral. This occurs for 14.9% of happy samples, 23.7% of sad samples, 33.6% of anger samples, 19.6% of fear samples, 26.5% of disgust samples, and 24.6% of surprise samples.

This tendency is particularly relevant for fear and anger because these emotions have higher weights in the final danger detector. When the text model incorrectly classifies them as neutral, the emotional information provided by the text branch may have a smaller influence on the later danger detection stage. The effect of combining the text prediction with the independent audio prediction is examined separately in Section 6.5.

6.5 Experiment 3: Hybrid Emotion Recognition

The third experiment examines the behavior of the hybrid emotion recognition system. Unlike the previous experiments, where the audio and text models were evaluated independently, this

experiment focuses on how their predictions are combined for the same recording.

Since no labelled paired multimodal test set was used to evaluate the complete audio-text pipeline, a separate hybrid classification accuracy is not reported. Instead, representative recordings are examined to compare the audio prediction, text prediction, fusion weights, and final hybrid result.

For this experiment, previously stored audio recordings were used as input to the complete audio-text pipeline. These recordings were processed from WAV files rather than being captured through the live microphone interface.

File	Transcript	Text used	Language / Translation	Audio prediction	Text prediction	Fusion weights	Hybrid prediction
Fear_el.wav	Βοήθεια! Σε παρακαλώ, ας με βοηθήσει κάποιος.	Aid! Please someone help me.	el / google_translate	fear 55.31%	neutral 34.46%	audio 0.800, text 0.200	fear 43.83%
Anger_el.wav	Σου έχω πει να μην με ακουμπάς.	I told you not to touch me.	el / google_translate	anger 33.65%	neutral 47.29%	audio 0.700, text 0.300	anger 31.27%
Fear_en.wav	Help me, please. Help me, help me.	Help me, please. Help me, help me.	en / none	fear 61.44%	sad 31.04%	audio 0.773, text 0.227	fear 47.54%
Anger_en.wav	Stop touching me!	Stop touching me!	en / none	surprise 30.13%	anger 50.22%	audio 0.491, text 0.509	anger 38.12%
Happy_en.wav	I'm so happy to see you.	I'm so happy to see you.	en / none	sad 41.55%	happy 75.88%	audio 0.359, text 0.641	happy 47.72%
Neutral_el.wav	Σήμερα είναι Τρίτη.	Today is Tuesday	el / whisper_translate	anger 34.44%	neutral 67.22%	audio 0.700, text 0.300	neutral 28.27%
Sad_en.wav	I feel so lonely.	I feel so lonely.	en / none	sad 45.95%	sad 72.36%	audio 0.518, text 0.482	sad 60.91%
Surprise_el.wav	Ήρθες να με δεις!	You came to see me!	el / google_translate	surprise 39.78%	neutral 57.01%	audio 0.700, text 0.300	surprise 34.08%
Surprise_en.wav	Wow, you're	Wow, you're	en / none	neutral 24.48%	surprise 69.11%	audio 0.380, text	surprise 48.77%

av	here!	here!				0.620	
Disgust_el.wav	Τι κάνει αυτή εδώ.	What is she doing here?	el / google_translate	disgust 38.53%	neutral 62.32%	audio 0.700, text 0.300	disgust 26.44%

Table 6.6: Representative results of the hybrid emotion recognition system

The examples in Table 6.6 show that the contribution of the two modalities is not fixed. Depending on the current recording, the audio model may receive a larger weight, while in other cases the text model can have a stronger influence on the final result.

For example, in Fear_el.wav, the audio branch predicts fear with a score of 55.31%, while the text model predicts neutral with 34.46%. The fusion assigns weights of 0.800 to audio and 0.200 to text, resulting in a final hybrid prediction of fear with 43.83%. In this case, the audio information has the greater influence on the final result.

A different situation appears in Anger_en.wav. The audio model predicts surprise with 30.13%, while the text model predicts anger with 50.22%. The fusion weights are almost equally divided between the two modalities, with 0.491 for audio and 0.509 for text. The resulting hybrid prediction is anger with 38.12%, showing that the text prediction can have a slightly stronger influence when it receives a higher fusion weight for the current recording.

The same effect can be seen in Happy_en.wav. Although the audio model predicts sad with 41.55%, the text model predicts happy with a much higher score of 75.88%. In this case, the text branch receives a weight of 0.641, compared with 0.359 for audio, and the final hybrid prediction becomes happy with 47.72%.

There are also cases where both models agree. In Sad_en.wav, both the audio and text branches predict sad, with scores of 45.95% and 72.36%, respectively. Their combination keeps sad as the final emotion and produces a hybrid prediction of 60.91%.

Another pattern can be observed in the Greek recordings that require translation. In these examples, the contribution of the text branch does not exceed 0.300, which is consistent with the maximum weight defined for translated text in Section 5.6. For example, Anger_el.wav, Neutral_el.wav, Surprise_el.wav, and Disgust_el.wav all use audio and text weights of 0.700 and 0.300, respectively. In Fear_el.wav, the contribution of the text branch is reduced even further to 0.200. These examples show how the predefined limits for translated text affect the final fusion weights during inference.

Overall, the results show that the final hybrid prediction is not determined by a fixed contribution from audio and text. The modality weights vary between recordings, and either branch can have the greater influence depending on the current input. The examples also show the effect of the additional limits applied to translated text. Since this experiment is based on representative recordings rather than a labelled multimodal test set, these results describe the behavior of the implemented fusion mechanism and should not be interpreted as evidence that the hybrid system achieves higher classification accuracy than the individual audio or text models.

6.6 Experiment 4: Emotion-Based Danger Detection

The fourth experiment examines the final emotion-based danger detector. The detector receives the hybrid emotion distribution produced by the previous stage and uses it to generate the final DANGER or SAFE indication. The decision is based on the complete emotion distribution and the predefined emotion weights and threshold described in Section 5.7.

The predefined decision threshold of 0.40 was used for all examples. Values equal to or greater than this threshold produce a DANGER indication, while lower values produce SAFE. The threshold and the emotion weights are fixed design parameters and were not learned from a

separate danger-labelled dataset.

For this experiment, new speech samples were captured directly through the live microphone interface described in Section 5.8. Therefore, these examples correspond to separate recording sessions and are not the same audio files used in Experiment 3. As a result, even when similar sentences are spoken, differences in vocal expression, transcription, translation, and model predictions may lead to different hybrid emotion distributions.

Representative results from the complete system are presented in Table 6.7.

Transcript	Text used	Hybrid	Detector Value	Decision
Βοήθεια! Σε παρακαλώ, ας με βοηθήσει κάποιος.	Help! Please someone help.	fear 51.04%	60.63%	DANGER
Σου έχω πει να μην με ακουμπάς.	I have eaten you, don't listen to me.	anger 61.75%	53.94%	DANGER
Help me, please. Help me. Help me.	Help me, please. Help me. Help me.	fear 45.11%	55.47%	DANGER
Stop touching me!	Stop touching me!	anger 52.22%	50.19%	DANGER
I'm so happy to see you.	I'm so happy to see you.	happy 42.33%	18.76%	SAFE
Σήμερα είναι Τετάρτη.	Today is Wednesday.	neutral 47.78%	18.55%	SAFE
I feel so lonely.	I feel so lonely.	sad 56.42%	31.85%	SAFE
Τα πιάτα είναι άπλυτα.	The dish is simple!	disgust 29.00%	23.87%	SAFE
Wow! It is on my face!	Wow! It is on my face!	surprise 51.21%	15.85%	SAFE
Help! I'm afraid of the snakes. Help!	Help! I'm afraid of the snakes. Help!	fear 63.90%	72.16%	DANGER

Table 6.7: Representative results of the emotion-based danger detection stage

Table 6.7 shows how the final hybrid emotion distribution is used by the danger detector to

produce a DANGER or SAFE indication. In general, recordings with stronger contributions from fear or anger produce higher detector values and are therefore more likely to exceed the predefined 40% threshold.

For example, the Greek sentence beginning with “Βοήθεια! Σε παρακαλώ...” is mainly classified as fear by the hybrid system, with a value of 51.04%. The final detector value is 60.63%, which is above the threshold and therefore produces a DANGER indication. A similar result appears for “Stop touching me!”, where anger is the strongest hybrid emotion with 52.22%. In this case, the detector value is 50.19%, which also results in DANGER.

Different behavior can be observed for happy and neutral examples. “I’m so happy to see you.” produces a detector value of 18.76%, while the Greek neutral sentence “Σήμερα είναι Τετάρτη.” produces 18.55%. Both values remain below the threshold and therefore result in SAFE. Happy and neutral do not directly increase the detector value because their predefined weights are zero. However, the final value is not necessarily zero because the complete hybrid emotion distribution is used and smaller probabilities may also be assigned to emotions such as fear, anger, sadness, or disgust.

The sentence “I feel so lonely.” provides another useful example. The hybrid model predicts sad as the strongest emotion with 56.42%, but the detector value is only 31.85%, which remains below the threshold. This happens because sadness has a smaller predefined weight of 0.30. Therefore, the presence of a negative emotion does not automatically lead to a DANGER indication.

The examples also show that transcription and translation can affect the final result. In some Greek recordings, the English text does not fully preserve the original sentence. However, the danger detector does not use the translated text directly. It uses the final hybrid emotion distribution, which contains information from both the audio and text branches. As a result, the audio prediction can still influence the final decision when the transcription or translation is not fully accurate.

CHAPTER 7

DISCUSSION, INTERPRETATION & FUTURE WORK

7.1 Results Interpretation

The experimental results presented in Chapter 6 show different performance patterns for the audio and text emotion recognition models. The audio model achieved an accuracy of 73.74% and a macro-F1 of 73.80%, while the text model achieved an accuracy of 71.50% and a macro-F1 of 59.01%. Although their overall accuracy values are relatively close, the difference in macro-F1 shows that their performance is not equally distributed across the seven emotion categories.

The audio model showed more balanced performance between the different emotions. Anger and neutral were recognized more reliably, while fear and sad were more difficult for the model. The performance of fear is particularly important for the proposed system because fear has the highest weight in the final emotion-based danger detector. Errors involving this category can therefore affect the final DANGER or SAFE indication.

The text model showed greater differences between the emotion categories. Neutral achieved the highest F1-score, although it was also by far the largest class in the test set. Disgust and anger were among the more difficult emotions for the model. The noticeable difference between macro-F1 and weighted-F1 also shows that the overall performance is influenced by the unequal number of samples available for the different classes.

The hybrid examples presented in Chapter 6 show that the two modalities do not behave in the same way for every recording. In some examples, the audio prediction receives the larger contribution, while in others the text prediction has a stronger influence on the final hybrid result. There are also cases where both models predict the same emotion and their combination maintains or reinforces this result. These examples illustrate the reason for allowing the fusion weights to change dynamically instead of using the same audio and text contribution for every recording.

The text branch also introduces an additional source of uncertainty because it depends on transcription and, for Greek speech, translation into English. An incorrect transcript or translation can change the text that reaches the emotion classifier and therefore affect its prediction. The audio branch does not depend on these stages because it processes the original speech signal directly. For this reason, the two branches can sometimes produce different emotional predictions for the same recording.

However, the hybrid results should be interpreted within the limits of the experimental setup. No labelled paired multimodal test set was used to calculate a separate accuracy or F1-score for the complete audio-text system. The representative examples therefore show how the dynamic fusion mechanism behaves, but they do not demonstrate that the hybrid system achieves higher classification accuracy than either the audio or text model individually.

The final experiment also shows how the hybrid emotion distribution is used by the emotion-based danger detector. Examples with stronger fear or anger contributions generally produced higher detector values and were more likely to result in a DANGER indication. However, the detector uses emotional information and does not have a complete understanding of the real-world situation. A strong fear or anger prediction may therefore produce DANGER even when the surrounding situation does not represent an immediate real-world threat. This is an important limitation of the proposed approach and is discussed further in Section 7.3.

7.2 Strengths of the Proposed System

One of the main strengths of the proposed system is the combination of speech and textual information for emotion recognition. The two modalities capture different aspects of the same spoken input: the audio branch represents characteristics related to vocal expression, such as intensity, rhythm, and pitch, while the text branch focuses on the verbal content of the utterance. Their combination therefore reduces the dependence of the system on a single source of emotional information and allows the final prediction to consider both acoustic and linguistic evidence.

Another important strength is the dynamic fusion mechanism. Rather than assigning fixed importance to the audio and text branches for every recording, the system adjusts their contribution according to the quality and reliability of the available predictions. This allows the relative influence of each modality to change from one input to another. As observed in the representative examples, there are cases in which the audio information has a stronger influence, cases in which the textual information becomes more important, and cases in which both modalities support the same emotional category.

The use of pre-trained models also strengthens the proposed approach. WavLM-base-plus is used as a frozen feature extractor for the audio branch, allowing pre-trained speech representations to be included without training the representation model from the beginning. For the text branch, RoBERTa-base is fine-tuned for the seven emotion categories used in this work. In this way, both branches make use of existing pre-trained representations while being integrated into the specific emotion recognition pipeline.

A further advantage is the ability to process both English and Greek speech within the same overall pipeline. English recordings can be transcribed and processed directly by the text model, while Greek recordings are transcribed and translated into English before text emotion classification. Although the additional translation stage introduces limitations, the bilingual processing capability extends the range of inputs that can be examined by the proposed system.

The implementation with live microphone input also provides a practical way to examine the complete processing pipeline beyond previously stored test files. A new recording can pass through speech transcription, text processing, both emotion recognition branches, dynamic fusion, and the final emotion-based danger detection stage. This makes it possible to observe the behavior of the complete system on newly recorded speech and to inspect the output produced by each individual stage.

Finally, the proposed architecture is modular. The audio emotion model, text emotion model, transcription and translation process, fusion mechanism, and emotion-based danger detection stage operate as distinct components within the same pipeline. This separation makes the contribution and behavior of each stage easier to examine and allows individual components to be modified, replaced, or improved in future work without requiring the complete system to be redesigned.

7.3 Limitations

7.3.1 Audio model

The performance of the audio emotion model was not equally strong across all emotional categories. This was reflected in the experimental results, where fear and sadness showed lower recognition performance than categories such as anger and neutral.

Errors involving fear are particularly relevant to the proposed danger detection stage because fear has the highest weight in the emotion-based danger detector. If fear is confused with another emotion, its contribution to the final detector value may be reduced, while an incorrect fear prediction may increase the detector value unnecessarily.

Another limitation is related to the available speech data. The audio model was developed using RAVDESS and CREMA-D together with the custom recordings, providing data from different speakers and sources. The custom dataset added further examples to the training process, although it was limited to three speakers. A larger custom dataset with more speakers could provide additional variation in future experiments.

The final evaluation was performed on speakers from RAVDESS and CREMA-D that were kept separate from training. This allowed the model to be tested on previously unseen speakers and provided a speaker-independent evaluation of its performance. However, the experiments were carried out using the available datasets and did not separately examine every possible recording condition, such as different microphones or stronger environmental noise.

7.3.2 Text model

The text emotion model has some limitations related to the type of information that it processes. By design, the text branch analyzes the written content of an utterance and does not use vocal characteristics such as tone, intensity, rhythm, or speaking style. As a result, emotional information that is expressed mainly through the way a sentence is spoken is not directly available to the text model. This is one of the reasons why the proposed system also includes the independent audio branch.

The experimental results also showed differences in performance between the emotion categories. Neutral achieved the strongest results, while categories such as anger and disgust were more difficult to recognize. The difference between macro-F1 and weighted-F1 also reflects the unequal number of samples available for the different classes, with the larger neutral category having a greater influence on the weighted results.

A further limitation is related to the type of text used during training and inference. The text classifier was trained using English samples from GoEmotions, while in the complete system it can also receive automatically transcribed speech and, for Greek recordings, translated English text. This creates a difference between the text used during model development and some of the inputs encountered by the complete system. The independent test set still provides a quantitative evaluation of the text classifier on the prepared GoEmotions data, while transcription and translation introduce an additional challenge when the model is used as part of the complete speech-processing pipeline.

7.3.3 Transcription and translation

Transcription and translation introduce an additional source of uncertainty before the text reaches the emotion classification model. Whisper is used to convert the original speech signal into text, and errors during this process can change, remove, or replace words from the spoken sentence. As a result, the text received by the emotion classifier may sometimes differ from the original spoken content.

For Greek recordings, an additional translation step is required because the text emotion model processes English input. The implementation uses both Google Translate and Whisper translation to produce possible English versions of the Greek speech. Although using two translation methods allows their outputs to be compared, the resulting English text may still differ from the original Greek sentence.

Examples of this behavior were observed during the experiments, where some Greek recordings produced inaccurate English translations. In these cases, the text emotion model processes wording that does not fully represent the original spoken content, which can affect its emotion prediction.

The dynamic fusion mechanism reduces the contribution of the text branch when its information is considered less reliable. However, transcription or translation errors can still influence the text

prediction and therefore the final hybrid emotion distribution. This may also affect the output of the emotion-based danger detector, since it operates on the final hybrid distribution.

7.3.4 Hybrid Fusion

The hybrid fusion mechanism also has some limitations. The final audio and text weights are adjusted dynamically for each recording, but the rules and numerical parameters used for this adjustment are predefined design choices. They were not learned or optimized using a labelled multimodal dataset.

The representative examples in Chapter 6 showed that the mechanism can change the contribution of the two modalities depending on the current input. However, since no labelled paired audio-text test set was used, the complete hybrid system was not evaluated with a separate quantitative accuracy or F1-score. Therefore, the experiments demonstrate the behavior of the fusion mechanism but do not show that it performs better than the individual audio or text models.

7.3.5 Emotion-Based Danger Detection

The emotion-based danger detector converts the final hybrid emotion distribution into a DANGER or SAFE indication. The representative examples showed behavior consistent with the predefined design of the detector, with stronger contributions from fear and anger generally producing higher detector values, while examples dominated by neutral or happy remained below the decision threshold.

However, emotional information alone does not provide a complete description of the real-world situation. For example, “Help! I’m afraid of the snakes. Help!” produced a DANGER indication because the hybrid result contained a strong fear contribution. However, the presence of fear alone does not confirm that an immediate real-world danger is present. In the same way, because the detector depends only on the available emotional evidence, a situation with a weaker emotional response may remain below the predefined threshold. The final output should therefore be interpreted as an emotion-based indication of possible danger rather than as confirmation of an actual dangerous event.

The detector also depends on the output of the previous stages of the system. Errors in audio or text emotion recognition, transcription, or translation can affect the final hybrid emotion distribution and therefore influence the resulting DANGER or SAFE indication.

Finally, the emotion weights and the 0.40 decision threshold are predefined design parameters of the current implementation. They provide a clear and reproducible decision rule, but they were not learned or optimized using a dedicated dataset containing ground-truth DANGER and SAFE labels. A future evaluation using labelled danger-related data could be used to examine and refine these parameters.

7.4 Future work

Several directions could be explored in future work to further improve the reliability, generalization, and applicability of the proposed system.

One important direction is the use of larger and more diverse speech datasets. Additional recordings from speakers with different ages, accents, speaking styles, and recording conditions could improve the ability of the audio emotion model to generalize to voices and environments that were not represented during training. Evaluation under more realistic conditions, including background noise and different microphone characteristics, would also provide a more complete assessment of the model’s robustness.

Another important extension would be the creation or use of a dedicated dataset containing

labelled DANGER and SAFE situations. Such a dataset would make it possible to evaluate and tune the emotion weights and decision threshold using labelled data. It could also support the development of a dedicated danger detection model incorporating information beyond emotion alone.

The multilingual component of the system could also be improved. In the current implementation, Greek speech is translated into English before being processed by the text emotion model. A multilingual emotion recognition model, or a model trained directly on Greek emotional text, could reduce the dependence on translation and limit the effect of translation errors on the final prediction.

The hybrid fusion mechanism represents another area for future development. The current approach adjusts the contribution of the audio and text branches through predefined rules based on the available information. With an appropriate paired multimodal dataset, a future system could learn the fusion process directly from data and determine the relative contribution of each modality automatically.

Additional sources of contextual information could also be incorporated into the decision process. Emotion alone cannot always determine whether an actual dangerous situation is taking place. Environmental sounds, the semantic context of the situation, previous speech segments, and temporal changes in the speaker's emotional state could provide additional evidence for a more complete danger assessment.

Finally, the system could be extended into a continuous real-time application rather than processing isolated short microphone recordings. Such an implementation would require further evaluation of processing latency, computational requirements, robustness, privacy, and reliability.

These extensions could improve the evaluation and reliability of the proposed system and support the development of a more complete multimodal approach for emotion-based danger detection.

CHAPTER 8

CONCLUSION

8.1 Conclusion

This thesis presented a multimodal emotion recognition system that uses information from speech and text together with an emotion-based danger detector. The main objective was to investigate how emotional information obtained from two complementary modalities could be combined and subsequently used to identify indications of potential danger.

The proposed system integrates two emotion recognition branches. Speech is analyzed using WavLM-based representations combined with handcrafted acoustic features, while textual information is processed using a RoBERTa-based emotion classifier. Whisper provides the transcription required by the text branch, while Greek recordings are additionally translated into English. The outputs of the two emotion recognition models are then combined through a dynamic fusion mechanism that adjusts their relative contribution according to the available information.

The experimental evaluation showed the performance of both individual emotion recognition models. The audio model achieved an accuracy of 73.74%, a balanced accuracy of 76.18%, and a macro-F1 of 73.80%, while the text model achieved an accuracy of 71.50%, a balanced accuracy of 62.13%, and a macro-F1 of 59.01%. The class-level results also showed that the two models have different strengths and weaknesses across the seven emotion categories.

The representative hybrid experiments demonstrated how the relative contribution of the two modalities can vary between recordings. In some cases, the audio branch had a stronger influence on the final emotion prediction, while in others the text branch received a greater contribution. These examples show how the implemented fusion mechanism can use information from both modalities instead of applying the same fixed contribution to every recording.

The final hybrid emotion distribution was used by the emotion-based danger detector to produce a DANGER or SAFE indication. In the representative examples, stronger contributions from fear and anger generally produced higher detector values, while the examined neutral and happy examples remained below the predefined decision threshold. In this way, the implemented system provides a complete processing pipeline from speech input to audio-text emotion recognition and final emotion-based danger detection.

This work brings together acoustic emotion recognition, textual emotion recognition, speech transcription, bilingual processing, dynamic multimodal fusion, and an emotion-based danger detector within a single modular pipeline. Further evaluation using labelled multimodal and danger-related data, together with more varied real-world recordings, could provide a more complete assessment of the system and support its future development.

References

- [1] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A Dataset of Fine-Grained Emotions," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds., Online: Association for Computational Linguistics, Jul. 2020, pp. 4040–4054. doi: 10.18653/v1/2020.acl-main.372.
- [2] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face," *Entropy*, vol. 25, no. 10, p. 1440, Oct. 2023, doi: 10.3390/e25101440.
- [3] P. Ekman, "Are there basic emotions?," *Psychological Review*, vol. 99, no. 3, pp. 550–553, 1992, doi: 10.1037/0033-295X.99.3.550.
- [4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019, doi: 10.1109/ACCESS.2019.2936124.
- [5] T. Deschamps-Berger, L. Lamel, and L. Devillers, "End-to-End Speech Emotion Recognition: Challenges of Real-Life Emergency Call Centers Data Recordings," in *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, Sep. 2021, pp. 1–8. doi: 10.1109/ACII52823.2021.9597419.
- [6] S. R. Livingstone and F. 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*, vol. 13, no. 5, p. e0196391, May 2018, doi: 10.1371/journal.pone.0196391.
- [7] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset," *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, Oct. 2014, doi: 10.1109/TAFFC.2014.2336244.
- [8] S. Chen *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, Oct. 2022, doi: 10.1109/JSTSP.2022.3188113.
- [9] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," Jul. 26, 2019, *arXiv*: arXiv:1907.11692. doi: 10.48550/arXiv.1907.11692.
- [10] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavy, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in *Proceedings of the 40th International Conference on Machine Learning*, PMLR, Jul. 2023, pp. 28492–28518. Accessed: Sep. 05, 2026. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [11] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, "Speech emotion recognition in conversations using artificial intelligence: a systematic review and meta-analysis," *ArtifIntell Rev*, vol. 58, no. 7, p. 198, Apr. 2025, doi: 10.1007/s10462-025-11197-8.
- [12] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980, doi: 10.1037/h0077714.
- [13] X. Huang, Y. Wang, H. Huang, C. Zhou, X. Zhou, and Y. Chen, "A Speech Emotion Recognition Model Combining WavLM Pre-Trained Features and Attention Mechanism," *Electronics*, vol. 15, no. 13, p. 2855, Jan. 2026, doi: 10.3390/electronics15132855.
- [14] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: Sep. 05, 2026. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

- [15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.
- [16] F. Alqarni, A. Sagheer, A. Alabbad, and H. Hamdoun, "Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition," *Sci Rep*, vol. 15, no. 1, p. 17617, May 2025, doi: 10.1038/s41598-025-99515-6.
- [17] W. Lu, Y. Xu, and J. Gu, "Adaptive weighting in a transformer framework for multimodal emotion recognition," *Speech Communication*, vol. 176, p. 103332, Jan. 2026, doi: 10.1016/j.specom.2025.103332.
- [18] Y. Wu, S. Zhang, and P. Li, "Multi-modal emotion recognition in conversation based on prompt learning with text-audio fusion features," *Sci Rep*, vol. 15, no. 1, p. 8855, Mar. 2025, doi: 10.1038/s41598-025-89758-8.
- [19] Y.-C. Yoon and S. Kim, "MASKSER: A Robust Emotion Recognition Model Based on Voice Data and Noisy Transcripts," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, pp. 4154–4167, 2025, doi: 10.1109/TASLPRO.2025.3617234.
- [20] S. Ashraf, M. Al-Kfairy, B. Shah, S. Anwar, Z. Halim, and B. Hsiao, "Cross-Lingual Emotion Identification: A Collaborative Learning Approach for Cyber-Physical-Social Systems," *IEEE Transactions on Computational Social Systems*, pp. 1–13, 2026, doi: 10.1109/TCSS.2026.3708567.
- [21] J. Staš, S. Ondáš, and J. Juhár, "Performance Evaluation of Different Speech-Based Emotional Stress Level Detection Approaches," *IEEE Access*, vol. 13, pp. 112880–112904, 2025, doi: 10.1109/ACCESS.2025.3584534.
- [22] J.-T. Tzeng, S.-G. Leem, A. N. Salman, C.-C. Lee, and C. Busso, "Noise-Robust Speech Emotion Recognition Using Shared Self-Supervised Representations with Integrated Speech Enhancement," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Apr. 2025, pp. 1–5. doi: 10.1109/ICASSP49660.2025.10887569.
- [23] G. Jin, Y. Xu, H. Kang, J. Wang, and B. Miao, "DSTM: A transformer-based model with dynamic-static feature fusion in speech emotion recognition," *Computer Speech & Language*, vol. 90, p. 101733, Mar. 2025, doi: 10.1016/j.csl.2024.101733.
- [24] F. Portal, J. De Lope, and M. Graña, "A Performance Benchmarking Review of Transformers for Speaker-Independent Speech Emotion Recognition," *Int. J. Neur. Syst.*, vol. 35, no. 10, p. 2530001, Oct. 2025, doi: 10.1142/S0129065725300013.
- [25] S. Yang, J. Chen, and Y. Tao, "Performance Evaluation and Analysis of LLMs for Detecting Dangerous Speech in Emotional Disorder Data," in *2025 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, Nov. 2025, pp. 1–8. doi: 10.1109/CloudCom67567.2025.11331545.
- [26] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The Balanced Accuracy and Its Posterior Distribution," in *2010 20th International Conference on Pattern Recognition*, Aug. 2010, pp. 3121–3124. doi: 10.1109/ICPR.2010.764.
- [27] S. Farhadpour, T. A. Warner, and A. E. Maxwell, "Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices," *Remote Sensing*, vol. 16, no. 3, p. 533, Jan. 2024, doi: 10.3390/rs16030533.