UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Escuela Técnica Superior de Ingeniería Industrial

Applications of emotion perception driven by AI in the field of mental health, with special attention to low-resource languages

Trabajo Fin de Máster

Máster Universitario en Ingeniería Industrial-Màster Universitari en Enginyeria Industrial

AUTOR/A: López i Rios, Daniel

Tutor/a: Simarro Fernández, Raúl Harder Clemmensen, Line Katrine Alemany

CURSO ACADÉMICO: 2024/202

DTU Compute
Department of Applied Mathematics and Computer Science

M.Sc. Thesis

Applications of AI-driven Emotion Perception for Mental Health Care with a focus on Low-Resource Languages

Daniel López I Ríos (s222927)

Kongens Lyngby 2024

DTU Compute
Department of Applied Mathematics and Computer Science
Technical University of Denmark

Matematiktorvet
Building 303B
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Preface

This thesis was prepared at the department of Applied Mathematics and Computer Science at the Technical University of Denmark in fulfillment of the requirements for acquiring the degree of Master of Science in Engineering, MSc Eng.

It is assumed that the reader has basic knowledge in the areas of statistics, linear algebra, machine learning and deep learning.

Kongens Lyngby, July 31, 2024

Daniel Lopez

Abstract

This thesis investigates the improvement of speech emotion recognition systems through the integration of different interpretation techniques, focusing on low-resource languages. The research addresses two main areas: performance variations across languages, and the influence of emotional representations on system generalization.

Three speech emotion recognition models were implemented, all following from the same pre-trained model: wav2vec 2.0. This model was adapted and retrained to read emotions following different systems of interpretation. A classification model for discrete emotions, a regression model for continuous emotions, and a multiobjective model combining both. The evaluation assesses system performance across English and non-English datasets to refine emotion recognition capabilities in various linguistic and cultural contexts.

The findings show that incorporating multiple emotional representations helps to stabilize predictions but does not fully resolve generalization issues. Performance discrepancies between English and non-English datasets and significant biases towards specific emotions are noted. The study underscores the need for more sophisticated models that can address the linguistic and cultural diversity of non-English populations. Future research should focus on enhancing cross-corpus training and studying the applicability of this technique with other modalities. This research contributes to advancing general emotion recognition for broader inclusivity in human-computer interaction technologies.

Acknowledgements

I extend my heartfelt thanks to my supervisors, Line Katrine Harder Clemmensen, Sneha Das, and Paula Petcu, for their invaluable guidance and insightful feedback throughout the duration of the project. Their expertise and dedication have been fundamental in shaping the project and steering me towards the goal.

Additionally, I would like to thank Ari Goldhar Menachem, with whom I worked on the general project on which this thesis is based. His dedication and insight have been invaluable for the completion of this project.

Contents

CHAPTER 1

Introduction

1.1 Introduction and motivation

Recent advancements in deep learning and the popularization of large language models (LLMs) have brought forth a revolution in AI that promises immense potential for the future. Countless industries and organizations across various fields have already shown an interest in applying intelligent chatbots to enhance internal processes and improve client interactions. These innovations are reshaping how we interact with technology, offering smarter, more efficient solutions. The newly developed capacity for seamless interaction between computers and humans, primarily through chatbots, is at the heart of this transformative potential.

For human-computer interaction to be genuinely effective, communication must be fluid and natural, mirroring human-to-human interactions as closely as possible. A critical aspect of this communication is the expression and recognition of emotions. Emotional expression and recognition enable more meaningful and empathetic interactions, which are essential for user satisfaction and engagement. Consequently, there is a growing market need for sophisticated emotion recognition systems capable of understanding and responding to human emotions accurately.

Emotion recognition technology has a wide array of potential applications. For instance, it can assist healthcare professionals in understanding patients' emotional states, thereby improving the quality of care. In education, emotion recognition can help children with communication challenges express their feelings and intentions more effectively. In security, it can enhance profiling and screening processes. One particularly promising application is in mental health care, where companies like Tetatet are developing AI applications to provide mental health support. One example among them is a chatbot that can perform emotional awareness exercises, recognize emotional cues to engage users in meaningful conversations and generally help manage mental well-being, among many utilities. Emotion recognition is naturally crucial for many of those applications.

Despite these advancements, deep learning systems face significant limitations, particularly their reliance on large datasets, which are expensive and challenging to create. Most AI models are primarily trained in English, benefiting from the abundant resources available for this language. However, this creates a significant challenge when these models are applied to other languages, as they often suffer a well-known drop in performance due to the lack of training data. This issue is even more problematic for low-resource languages with fewer speakers and resources, often overlooked in the training of AI models. This situation presents a real risk of excluding these communities from the technological benefits of emotion recognition systems. Bridging this gap is essential to ensure equitable access to advanced AI technologies across different linguistic groups.

The goal of this thesis is to analyze different methods to create end-to-end unimodal systems for audio input with satisfactory accuracy and robustness in English and other languages not originally trained in. With this, this project aims to impulse this technology and its potential applications and help make them accessible and beneficial to a broader range of linguistic communities, thereby enhancing their inclusivity.

1.2 Research objectives

My research aims to explore various solutions to the challenge of emotion recognition in low-resource languages. Among the numerous techniques available in deep learning, I will primarily focus on unimodality due to its simplicity compared to multimodal techniques and its potential to escalate and adapt to more complex applications later on. This study will examine the effectivity of multiple systems with different speech emotion representation, their robustness in handling non-trained languages, and the biases exposed in their behaviour depending on the language and the emotion to classify. To achieve this, I have formulated two key research questions that this thesis aims to address:

1. What are the performance disparities of emotion recognition systems when applied to low-resource language datasets?
2. How does the emotional representation of speech affect the performance of the models and their capability to generalize across languages?

1.3 Thesis structure

The thesis is organized as follows. Chapter 1 introduces the motivation and objectives of the research. Chapter 2 presents background concepts and previous work related to emotion recognition, feature extraction, emotional datasets, emotion representation and transfer learning methodologies. Chapter 3 describes the datasets used in this study. Chapter 4 details the speech emotion recognition model used in this study. Chapter 5 evaluates the experiments, comparing the performance of different models and their effectiveness. Chapter 6 discusses the results, interpreting the implications of the results. Chapter 7 concludes the thesis with a summary of findings and future research directions.

CHAPTER 2

Background and previous work

This chapter begins by exploring conversational agents and their applications in healthcare. The field of emotion recognition is introduced and described from its early stages, focusing primarily on single modalities, to modern advanced multimodal systems that enhance accuracy and reliability. The details of speech feature extraction are presented, emphasizing the challenges and breakthroughs that have shaped the field. Furthermore, the characteristics and challenges related to emotional data are discussed.

2.1 Conversational agents for health care

For several years, the potential of conversational agents, or chatbots, to make various aspects of our lives more efficient and convenient has been increasingly recognized. One notable area is health care, where the applications of this technology have been particularly emphasized. One of the earliest examples, ELIZA, was introduced by Joseph Weizenbaum in 1966 [30]. This system was one of the first to demonstrate how a machine could function as a psychotherapist. With the rapid advancements in artificial intelligence and natural language processing, the introduction of systems like Apple's Siri and Amazon's Alexa, along with the increase in highly capable large language models, has caused renewed interest in this field.

In a comprehensive 2018 review, Laranjo et al. [12] evaluate studies describing conversational agents used in health care. These agents vary in complexity, from finite-state systems that guide users through a pre-determined dialogue to more sophisticated agent-based systems that engage in complex interactions. Modern systems, such as ChatGPT, exemplify agent-based systems where each agent can reason about its own actions and beliefs and dynamically adapt the dialogue based on the conversation's context.

An intermediate system type is a frame-based system, which uses frames to collect the necessary information to accomplish a specific task. Unlike finite-state systems, these are not limited to a fixed dialogue path but are guided toward filling the information "slots" required to complete the interaction.

The significance of these systems extends beyond mere task completion. For instance, a study by Fitzpatrick et al. [9] found that a non-task-oriented frame-based conversational agent could significantly reduce symptoms of depression among college students over two weeks by engaging them in dialogues based on cognitive behavioral therapy principles. Moreover, a diagnostic study [21] highlighted the potential and challenges of using conversational agents for diagnosing major depressive disorders (MDD). The agent's ability to accurately diagnose MDD improved with the severity of symptoms, achieving a sensitivity of 73% in cases with severe symptoms, while maintaining a specificity above 95% across all severity levels.

These studies underscore the potential for conversational agents not only to perform specific tasks but also to understand and respond to human emotions, making them invaluable in sensitive applications like mental health. The development of emotion recognition capabilities in these systems could enhance their effectiveness by enabling more empathetic and contextually appropriate interactions. This enhancement is crucial for applications where understanding user emotion is key, such as in therapy or customer service. By integrating advanced emotion recognition technologies, conversational agents could revolutionize how we interact with machines, making these exchanges more human-like and supportive.

2.2 Emotion recognition

Research on emotion recognition is vital as it equips computers with the capability to interpret human emotions accurately and respond intelligently to human needs, enhancing human-computer interaction. One area where this technology holds particular significance is in the field of mental health care, enabling psychiatrists and psychologists to better understand the emotional conditions of their patients, which can facilitate more personalized and precise healthcare services. Consequently, emotion recognition has increasingly evolved into an area of high interest within artificial intelligence, capturing widespread interest due to its potential to fundamentally change how humans and computers engage with one another.

In the early stages of emotion recognition research, the focus was predominantly on unimodal emotion recognition systems, such as voice [8, 28, 36, 10, 19], text [31, 26], or facial expression recognition [34, 29, 14, 32]. These methods, when employed alone, are often subject to inaccuracies due to insufficient data available for training and vulnerability to environmental noise. This recognition of limitations led to the development of multimodal emotion recognition frameworks. Multimodal emotion recognition systems integrate data across various modalities, enabling the system to extract the most useful and discriminative features from each and to quantify dependencies among different modal features. The approach has shown a significant increase in the accuracy of emotion detection systems [33, 27, 15, 17].

Although multimodal systems have become the norm in recent research, these systems often employ separate unimodal systems for each modality which are then blended together by using a fusion module. Figure 2.1 illustrates a typical multimodal system with three categories: speech, text and face. Any improvement in one of the separate categories could then potentially benefit the multimodal system's capabilities as a whole, and thus it justifies the ongoing research into improving the separate unimodal models. As more research is being conducted into deep learning technologies, emotion recognition has

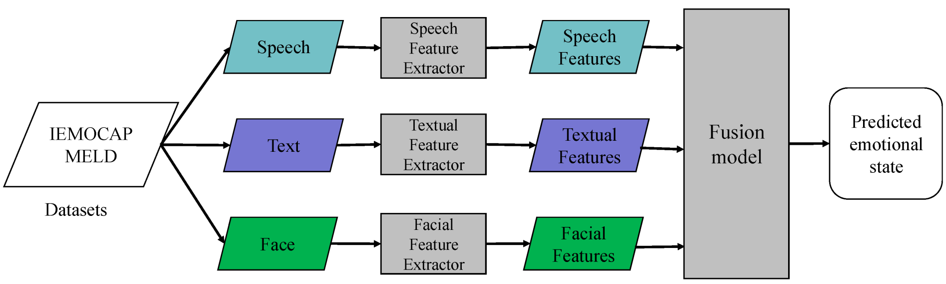


Figure 2.1: Overview of a typical multimodal emotion recognition framework [16].

become one of the leading topics in artificial intelligence research [18]. This involves the design of neural network architectures and specialized loss functions, with cross-entropy loss, in particular, being employed extensively [16].

2.3 Speech feature extraction

At the heart of emotion recognition technologies is the principle of identifying emotional content from a specific signal or modality. The careful identification of speech features that capture emotional variations represents a critical challenge in the field of emotion recognition. Typically, speech emotion features are divided into two main categories: hand-crafted features and deep speech emotion features. Hand-crafted features refer to the features used to describe speech signals that are designed by people based on prior knowledge and professional experience [16]. Deep features, on the other hand, are extracted using modern deep-learning techniques.

These methods involve using deep learning [13] models to extract a set of feature vectors that represent a signal's deep speech features. One of the strengths of deep learning is that it can learn high-level feature representations for emotion recognition automatically and often outperform traditional methods based on hand-crafted features. Common deep learning-based feature extraction methods include Deep Belief Networks, Convolutional Neural Networks, Recurrent Neural Networks, and Long Short-Term Memory networks [35].

A number of self-supervised frameworks have been developed and are seeing increased use. These models learn high-quality representations from unlabeled data. Among these are wav2vec [25] and the improved wav2vec 2.0 [1]. These unsupervised models use raw audio waveforms to obtain generalized speech features that can be applied to various downstream tasks. Other recent self-supervised models include HuBERT [11] and WavLM [5].

Despite significant progress in the field, deep feature extraction of speech features still faces obstacles. Deep learning models depend heavily on extensive datasets for training, which might not be readily available or practical to gather in certain situations. Moreover, although self-supervised models are generally efficient, they often fail to capture subtle emotional variations in speech [16]. Furthermore, even though the existing models perform well, they require considerable computational power, which can be a barrier for real-time applications. Future studies need to tackle these issues to improve the utility and performance of deep feature extraction in this field.

2.4 Emotional datasets

Datasets may have one or more modalities; the most common of them include audio, visual, and text, but other modalities, such as motion capture data or physiological signals, are used in some datasets.

Annotations can take the form of discrete, categorical emotion classes, continuous, dimensional values, or both. Datasets with categorical labels often employ Ekman's six basic emotions [7] (happiness, sadness, surprise, fear, disgust, and anger), as well as a neutral state. Studies commonly reduce the number of emotions used before training or fine-tuning emotion recognition models in order to increase performance.

Other datasets have continuous dimensional emotions. In this case, emotions are represented in a two-dimensional space with valence (negative-positive) on the x-axis and arousal or activation (low-high) on the y-axis. For example, anger is characterized by negative valence and high arousal. An illustration of the dimensional representation of emotions is shown in figure 2.2.

Emotion annotation is always subject to some ambiguity and subjectivity, and this issue is more significant for dimensional labels. For this reason, most datasets have categorical annotations. Furthermore, emotional datasets are divided in terms of their recording conditions: "In-the-lab", where actors imitate emotions in a clean environment, facing the camera or microphone directly, and "in-the-wild", spontaneous emotional expressions that are not acted for the purpose of the dataset, e.g. taken from movies or TV shows [24].

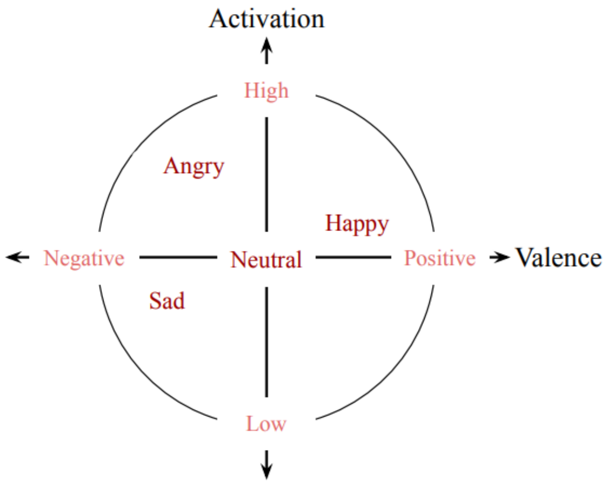


Figure 2.2: Illustration of the dimensional representation of emotions [6, 23].

2.4.1 Challenges related to data

It should be noted that most studies in emotion recognition were performed without thorough cross-corpus experimentation, thus their findings are specific to the datasets used. Creating a model that remains robust against variations in data conditions and distributions continues to be a persistent challenge across most machine learning fields. This issue is particularly crucial in the field of affective computing, as human emotional expression varies widely across ethnicities, cultures, languages, and even factors like age and gender [24].

A significant barrier in advancing emotion recognition technology is the scarcity of high-quality datasets. Obtaining data that is both comprehensive and representative of diverse emotional expressions across different demographics is a major challenge. How emotions are expressed varies greatly both between individuals and across different cultural backgrounds, nationalities, etc. Datasets from small countries are especially rare, posing an obstacle for the development of emotion recognition systems adapted to small ethnic populations and low-resource languages.

Secondly, ethical considerations play a significant role in acquiring emotional data. Ensuring the privacy and consent of participants, especially in sensitive contexts involving emotional expressions, limits the availability of such data. Datasets are often created using actors that are asked to express different emotions while facing the camera directly or speaking directly into the microphone, resulting in an artificial, enacted character of the expressed emotions. This varies greatly from true ”in-the-wild” emotional expressions, and thus poses another challenge for developing robust emotion recognition systems to be employed in real situations.

Moreover, the quality of data in terms of accuracy, resolution, and annotation is also an important factor, to a large extent due to the inherent subjectivity in emotion perception. Poorly annotated data, where the emotional labels do not accurately reflect the displayed emotions, can lead to models that are ineffective or biased in their predictions. For this reason, many datasets use annotations created by letting a number of people rate the perceived emotions, from which the average rating can be determined, instead of relying solely on the actors’ self-reported expressed emotions. Similarly, low-resolution data can limit the ability of models to distinguish subtle emotional cues, which are often crucial for accurate emotion recognition.

Furthermore, the advancement of emotion recognition systems for low-resource languages faces a barrier due to a lack of sufficient resources in terms of data and labels for languages beyond the most commonly spoken ones [6].

In conclusion, the lack of high-quality, diverse, and ethically sourced data is a significant barrier to developing reliable and universally applicable emotion recognition systems. Overcoming this barrier requires a large effort to create representative high-quality datasets for a broad range of languages and different demographics.

2.5 Dimensional labels for transfer learning

Das et al. [6] propose using the dimensional representation of emotions to improve the generalization of emotion recognition models across various languages, including those that are low-resource. Speech emotion recognition systems typically face challenges in generalization due to the wide range of linguistic features and emotional expressions across languages. Supervised learning approaches work well when there is a wealth of labeled data, but they are less effective for languages that lack extensive and annotated datasets.

The dimensional approach, which focuses on continuous scales of activation (emotional energy) and valence (emotional positivity or negativity), provides a more universal framework that is beneficial for accommodating the subjective nature of emotion perception. This subjectivity can vary significantly across different cultural and linguistic contexts. For instance, perceptions of what might be considered a 'neutral' emotional tone can vary between languages, influenced by inherent phonetic and cultural differences. Most conventional models train using fixed class labels, which do not effectively capture these variations.

By adopting a model that utilizes activation and valence to capture emotional intensity and polarity, Das et al. aim to offer a more adaptable approach. This method not only addresses the challenges of scarce labeled data in many languages but also improves the consistency of speech emotion recognition systems across diverse linguistic environments.

Das et al.'s semi-supervised approach, which integrates dimensional metrics of activation and valence, has demonstrated superior performance compared to traditional methods. Their results show that the model not only enhances the accuracy of emotion classification but also ensures that the learned emotional representations are more consistent and transferable across different language datasets [6]. This supports their proposal that dimensional representation can significantly improve the generalization capabilities of speech emotion recognition systems in linguistically diverse settings.

2.6 Chapter summary

This chapter examines the integration and implications of conversational agents and emotion recognition technologies, particularly in healthcare settings. It describes various methodologies in emotion recognition, including multimodal frameworks that analyze combined data from multiple modalities such as speech, text, and visual data for more accurate emotional assessments. Specific feature extraction techniques are covered, noting the progression from hand-crafted to deep learning methods. Furthermore, the critical role of high-quality, diverse datasets in refining these technologies is discussed. We explore the significant challenge of developing robust emotion recognition systems for low-resource languages and present the dimensional representation of emotions as a potential solution to this problem. This approach, by focusing on universal dimensions of emotional expression such as activation and valence, aims to improve the generalization capabilities of these systems across various linguistic environments. The chapter also covers the challenges of data scarcity, ethical considerations, and the technical barriers in deploying these systems in real-world settings.

CHAPTER 3

Datasets

As with any task involving the design of frameworks based on machine learning models, access to large amounts of high-quality data is crucial. This chapter presents the emotional datasets that were used in the implementation of the emotion recognition system. The chapter is structured as follows: First, a brief introduction to emotional datasets and their various properties and characteristics. Subsequently, the specific datasets employed in this research are introduced and detailed comprehensively.

3.1 Datasets used in this study

3.1.1 IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset is an audio-visual dataset collected by the Signal Analysis and Interpretation Lab (SAIL) at the University of Southern California [3]. The dataset consists of scripted and improvised dialogues. The modalities included are audio, visual, facial motion capture, and text data, thereby serving as an important source for thorough investigation of emotional states in interactive settings. The dataset employed ten American actors (five female and five male) speaking English. IEMOCAP includes 4784 improvised and 5255 scripted conversations, in total approximately 12 hours of audio-visual data, providing a wide spectrum of emotional contexts. The dialogues being both scripted and improvised results in the dataset containing a variety of emotional content, improving its representativeness of real-world emotional communication [16]. The recordings have an average duration of 4.5 s and an average word count of 11.4. The dataset contains both ten categorical emotion labels (happiness, sadness, anger, surprise, fear, disgust, frustration, excitement, neutrality, and other) and the continuous dimensional labels valence, arousal, and dominance. The data was labelled by six expert raters overall with no less than three raters per video. Although the IEMOCAP is an "in the lab" dataset the authenticity of the expressed emotions is increased due to the dialogues being improvised. This is beneficial in helping deep learning models perform well on real data. IEMOCAP is limited by its English-only data, reducing its applicability for flexible models intended to work across different languages and cultures. Furthermore, the dataset is characterized by a severe class imbalance, with some emotional classes having only very few instances, which can impact trained models' performance for those classes. The copy of IEMOCAP obtained for this study was, unfortunately, not complete, as some files were corrupted. In terms of audio files, the loss was minor, as only one out of five sessions were affected. For video data, however, the dataset contains only videos of full minutes-long dialogues, as opposed to the audio, where both data and labels were provided as individual utterances. For this reason, the IEMOCAP dataset was used exclusively as an audio dataset in this study.

3.1.2 CREMA-D

This audio-visual dataset [4] contains a total of 7442 video clips with an average duration of 2.5 seconds. The data consists of the interpretations of 91 actors (48 male and 43 female) from different nationalities and ethnicities speaking in English. They were asked to speak a selection of 12 sentences with six different emotions and four different levels of intensity. The annotators classified the emotions into one of six categorical emotion labels based on the combined audio-visual presentation, based solely on audio data, and solely on video data. 95% of the clips have more than 7 ratings. Contrary to other datasets, CREMA-D also includes the confidence level of the annotators about their ratings. This dataset gives us the opportunity to test the model against the self-reported emotion labels (the actor's representation of the emotion they were asked to perform) and the perceived emotions represented by the annotators' ratings. I tested the model with both methods since it could give insight into the capabilities of the model.The data in CREMA-D is more diverse than IEMOCAP with a higher number of actors that vary in terms of ethnicity and age (from 20 to 74).

3.1.3 Emo-DB

Speech dataset in German [2], with 535 utterances in total, each one lasting for a couple of seconds. 10 actors (five male and five female) performed 10 sentences with one of seven different emotions. Annotations are self-reported, i.e. the emotions the actors were asked to express.

3.1.4 ShEMO

Speech dataset in Persian [20], with 3000 semi-natural utterances of a few seconds each one, equivalent to 3 hours and 25 minutes of speech. The utterances were captured from radio talks by 87 native Persian speakers and rated by 12 participants, also native Persian speakers, into six emotional categories.

3.2 Chapter summary

Chapter 3 introduces and describes the datasets used in this study, emphasizing their critical role in developing robust emotion recognition systems. The chapter begins with an overview of emotional datasets, discussing their various properties, modalities, and types of annotations (categorical and dimensional). It then provides detailed descriptions of the specific datasets employed in the research, such as IEMOCAP and CREMA-D. Each dataset's characteristics, including the type of emotional content, recording conditions, and annotation methods, are thoroughly examined. The chapter also addresses the challenges of data processing and preparation, highlighting the importance of aligning data quality and annotation accuracy with the research objectives.

CHAPTER 4

Audio modality: Speech emotion recognition

4.1 Wav2vec 2.0

Wav2vec 2.0, developed by Meta (formerly Facebook AI), is a state-of-the-art model for processing audio data, particularly for automatic speech recognition tasks. Released in October 2020, it remains one of the most utilized self-supervised models for speech recognition. It allows the learning of useful representations from large amounts of unlabeled data before fine-tuning on smaller, labeled datasets, enabling adaptation to specific tasks.

The basic version of wav2vec 2.0 used in this thesis was trained exclusively on unlabeled English language data. Leveraging extensive unlabeled audio data is significant because it is more readily available and easier to prepare compared to labeled data. By pre-training on this vast amount of unlabeled data, the model learns robust representations of the underlying audio features, making it highly effective for subsequent fine-tuning tasks such as emotion recognition.

The inner workings of wav2vec 2.0 involve several key components and processes that enable it to effectively learn representations from raw audio data:

1. **Feature encoder:** The process begins with the feature encoder, in this case a multi-layer CNN designed to process raw audio waveforms into latent speech representations, capturing local temporal dependencies. The encoder produces a high-dimensional representation that serves as the input for subsequent stages.
2. **Latent representation and masking:** Once the audio is encoded into latent representations, wav2vec 2.0 employs a masking strategy inspired by masked language modeling in natural language processing. A certain proportion of the latent representations are randomly masked, and the model predicts the masked portions based on their surrounding context. This approach forces the model to learn robust, context-aware features that capture the underlying structure of the audio signal.
3. **Transformer network:** The masked latent representations are then fed into a Transformer network, which is known for its ability to model long-range dependencies and contextual information. The Transformer processes the entire sequence of latent representations, building a contextualized representation for each time step. This contextualization is crucial for understanding the temporal dynamics and patterns within the audio data.
4. **Quantization module:** A distinctive feature of wav2vec 2.0 is its quantization module, which discretizes the continuous latent representations into a finite set of learned speech units. This step involves using a Gumbel-Softmax operation to enable differentiable quantization, allowing the model to jointly learn discrete speech units and their contextual representations. The quantization helps in reducing the model's complexity and improving the robustness of the learned features.
5. **Contrastive loss and fine-tuning:** During pre-training, wav2vec 2.0 is optimized using a contrastive loss function. This loss encourages the model to distinguish the true latent representation

of masked time steps from a set of distractors. Additionally, a diversity loss is applied to ensure the model utilizes the quantized speech units uniformly, promoting a richer and more balanced representation.

After the pre-training phase, the model undergoes fine-tuning on labeled data for specific downstream tasks.

Wav2vec 2.0's open-source nature makes it accessible to researchers and developers worldwide, contributing to its rapid adoption and continuous improvement. Given its advantages, wav2vec 2.0 was selected as the pre-trained model for this thesis, offering a more effective solution than creating a new model from scratch.

To adapt the speech recognition model for emotion recognition, a new classification or regression head is required, and the model must be fine-tuned on an appropriate dataset. There have been many previous projects that implemented similar adaptations, so my initial approach involved exploring relevant platforms, such as Hugging Face, where there could potentially be an appropriate model for the thesis.

4.2 Pre-trained models for speech emotion recognition

I explored the most recent and popular models on Hugging Face for pre-trained speech emotion recognition. Most of these models used wav2vec 2.0 as their base, with subsequent fine-tunings to improve performance for emotion recognition. These models were accessed directly through the Transformers library, simplifying implementation.

I tested these models using two datasets: EmoDB in German and CREMA-D in English, selecting 160 random utterances for testing in each case. My objective was to verify the reported performance on English datasets and assess robustness in other languages. Initial results showed subpar capabilities, with some models performing significantly lower than reported.

My second approach involved fine-tuning my own model while monitoring new additions on Hugging Face. My attempts at preparing my own emotion classification model encountered challenges, such as choosing an appropriate dataset for training, balancing the data, and adjusting hyperparameters. All of these took more effort and time than expected. Eventually, I found a newly released fine-tuned model by the SpeechBrain team that passed my performance tests. SpeechBrain [22] is an open-source PyTorch toolkit focused on speech applications and conversational AI.

4.2.1 SpeechBrain model for classification

This new model was fine-tuned in Google Colab using a V100 GPU and the IEMOCAP dataset due to its availability and quality. It used an 80/10/10 split ratio for training, validation, and testing subsets. A classification head was added to the base wav2vec 2.0 model, retrained by freezing the feature encoder components of wav2vec 2.0 but leaving the Transformers and classification head trainable, which allegedly increased performance and reduced training time. The model was trained with a batch size of 4, for 30 epochs, with a dynamic learning rate that started at 0.00001 for the wav2vec2 model and 0.0001 for the classifier head. The training also utilized an Adam optimizer. Using this setting the model reportedly achieved a 78.7% accuracy, with an average class accuracy of 75.3% on the testing subset. While this subset was not shared, limiting verification, successive experiments with my dataset have shown consistent results. After verifying its suitability for the purposes of this thesis, I proceeded to adapt and use the model for other experiments.

The IEMOCAP dataset originally contained utterances classified within 11 different labels, but the dataset was pruned to contain only five of them: neutral, anger, happiness, sadness and excitement, which was found equivalent to happiness and relabeled as such. This simplification improved accuracy by focusing only on the most basic and universal emotions and facilitated compatibility with other datasets, as these labels are the most commonly used.

The splits among the used IEMOCAP data were performed randomly across all utterances labeled with the specified emotions. Notably, the SpeechBrain team did not separate the training, validation, and testing subsets based on the speaker, and I maintained this approach. This method was chosen because, ultimately, the ultimate application of this model would be a personalized system with which the user consistently interacts. In such a scenario, it is reasonable to anticipate that the model will learn and adapt from these interactions, enhancing its functionality over time with the same user. Thus, training and testing on utterances from the same speaker in different contexts seems appropriate and aligns with the technological trend of AI systems becoming more personalized and adaptive.

4.2.2 SpeechBrain model for regression

The Speechbrain model is capable of performing emotion classification among four different emotions using the IEMOCAP dataset. That said, this dataset provides labels not only for emotional classes but also for dimensional values, such as valence and arousal. This feature allowed me to adapt the existing model to perform regression instead of classification, predicting these dimensional attributes from the inputs.

The IEMOCAP dataset includes the three dimensions commonly used for defining emotional recognition: valence, arousal, and dominance. However, following common practice in the literature, the model was adapted to predict only the first two dimensions since they are commonly considered enough to define an emotional state. Furthermore, the simplification may help improve the performance of the model in representing the attributes we are most interested in.

While I was unable to acquire a non-English dataset with dimensional data to test the model's ability to generalize across languages, further experiments could use the results of these experiments to investigate the generalization potential, so they are also included in the project.

To adapt the original model for regression, I made several modifications to the dataset handling and model structure:

- **Dataset inclusion:** I included all instances from the original IEMOCAP dataset since there was no longer a limitation by emotion classes. This allowed for a more extensive use of the available data.
- **Data treatment:** The data handling was modified to save the dimensional attributes (valence and arousal) instead of the classification labels.
- **Model adjustment:** The linear head added to the wav2vec model was retained, but the number of output units was reduced from four to two, corresponding to the valence and arousal dimensions.
- **Loss function and evaluation metric:** I replaced the original loss function and evaluation metric with Mean Squared Error (MSE), which is more suitable for regression tasks.
- **Softmax removal:** The softmax function, which was applied to the original output, was removed to accommodate the regression outputs.

Other characteristics of the model remained unchanged. The wav2vec 2.0 model was then fine-tuned from scratch using the same 80/10/10 split ratio for training, validation, and testing on the IEMOCAP dataset, as was previously done for the classification tasks.

4.2.3 SpeechBrain multiobjective model

At this point, I had developed both a classification model and an adapted regression model. These are the two classical approaches for emotion recognition, depending on the type of labeled data available. However, it is possible to combine these approaches into a multi-objective model. Similar to multimodal or multitask models, the rationale behind using a multi-objective implementation is that training the model on different but related objectives can enhance learning efficiency, reduce training time, and potentially improve accuracy by sharing representations across tasks.

Both the classification of four basic human emotions and the regression of two emotional dimensions are related representations of emotion recognition but stem from different methodological approaches. By training them together, I aimed to enrich the internal representations and define clearer classification boundaries. A better understanding of emotional dimensions such as arousal and valence could potentially help differentiate between emotional classes more effectively.

This approach could also enhance the model's generalization capabilities for cross-corpus and cross-lingual cases. Sneha Das et al. [6] suggest that emotional dimensions, particularly arousal, are easier to identify across languages than discrete emotional classes. If this holds true, a multi-objective model with a well-defined arousal dimension in its internal representation may perform better in classifying emotions in other languages compared to a purely classification-focused model.

To test this hypothesis, several modifications were made to the previous models, particularly in dataset handling and model structure:

- **Dataset inclusion:** For classification, I only used data corresponding to the emotions intended to classify. The process was similar to that used for the classification model.
- **Data treatment:** The data handling was modified to include both the dimensional attributes (valence and arousal) and the classification labels. All instances labeled as "excitement" were changed to "happy."
- **Model adjustment:** Two linear heads were added to the wav2vec 2.0 model: one for classification and one for regression. The classification head included a bias term, whereas the regression head did not. The number of outputs for each head remained the same as previously defined.
- **Softmax implementation:** The log_softmax function was applied to the classification output to normalize the results, offering a more stable implementation when paired with the Negative Log-Likelihood (NLL) loss.
- **Loss function and evaluation metric:** Averaging pooling was applied to the outputs of both heads. The regression head used MSE as the loss function and metric, while the classification head used NLL after the softmax. The global loss function for the model was the sum of these two values. Since NLL tends to produce slightly higher values, the model is expected to train the classification task more aggressively than the regression task.

Keeping other hyperparameters intact, such as the learning rate and the freezing of the feature encoder, the model was fine-tuned from scratch. Due to changes in the available environments on Google Colab, I used an L4 GPU instead of a V100, as used for previous audio model training.

4.3 Chapter summary

Chapter 4 explores the use of the wav2vec 2.0 model, developed by Meta, for speech emotion recognition. This model leverages extensive unlabeled audio data to learn robust representations, which are then fine-tuned on smaller labeled datasets. The chapter details the key components of wav2vec 2.0, including its feature encoder, masking strategy, Transformer network, quantization module, and the use of contrastive loss during pre-training. It also discusses the process of fine-tuning pre-trained models for emotion recognition, using various datasets to assess performance and robustness. The SpeechBrain model, which was fine-tuned using the IEMOCAP dataset, demonstrated significant accuracy and was selected as the base model for the speech emotion recognition model. Finally, the chapter describes how the model was further adapted for other experiments, highlighting the potential of combining classification and regression tasks to enhance emotion recognition capabilities across languages.

CHAPTER 5

Experiments and results

In this chapter, I present the tests done on the models described in past chapters. To evaluate all models in the most complete way possible I used a variety of datasets, from English to non-English ones. All experiments were performed with an ample quantity of test data and balanced class representation to represent effectively any bias or trend acquired by the models.

5.1 Speech model for classification

According to the Speechbrain developers' report on Hugging Face, the model achieves an accuracy of 78.7% on IEMOCAP. Due to a lack of access to the specific training and testing subsets used, I was unable to verify these numbers directly. Instead, I conducted additional experiments to evaluate the model's performance on IEMOCAP and other datasets, including two in different languages.

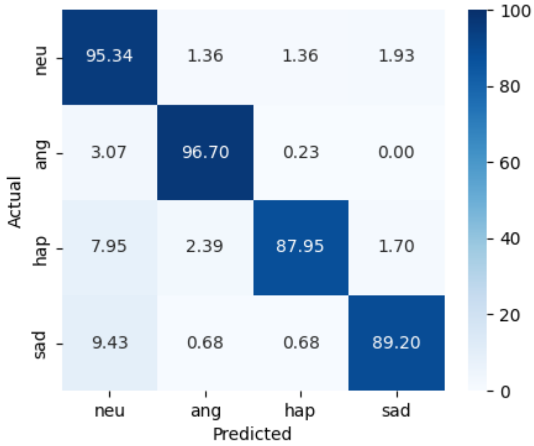
Experiment with IEMOCAP (see figure 5.1(a)): The model was tested in a random sampling of the IEMOCAP dataset, consisting of 494 utterances per class, totaling 1,976 utterances. This sample size was chosen to ensure equal representation of emotions while using the maximum available utterances from the partially accessible IEMOCAP dataset.

The model's high performance on this dataset was expected, as the sample likely included training data, but it serves as a point of comparison to other datasets. The results showed strong capabilities in identifying anger, happiness, and sadness, with the main error source being the misclassification of happiness and sadness as neutral. The model easily differentiated between anger and neutral emotions but struggled slightly more (above 7%) between neutral and happiness or sadness.

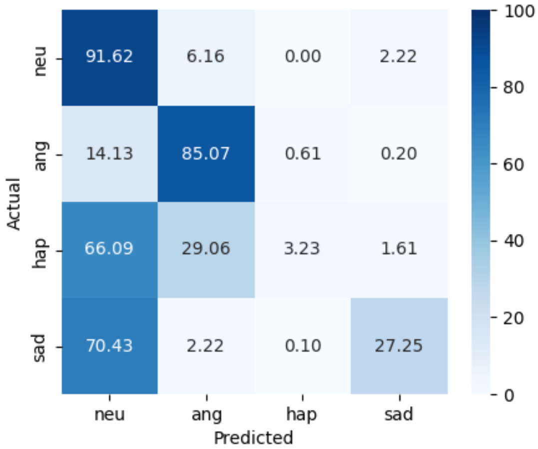
Experiment with CREMA-D: The CREMA-D dataset's self-reported emotion labels already had a balanced class representation. I removed utterances of unwanted classes, resulting in 1,087 inputs per class and 4,348 inputs in total. For human rated annotations, class frequencies varied, leading to a test set of 353 utterances per class, totalling 1,412 utterances.

As anticipated, the model's performance dropped on this different corpus. The CREMA-D tests mirrored the IEMOCAP results but showed more pronounced issues. The model recognized neutral and anger with above 75% accuracy but struggled with happiness and sadness, often misclassifying them as neutral or anger. Only on rare occasions does the model even predict an input as happiness. The model's classification of neutral and anger was more aligned with the rated annotations, suggesting it matched human rater performance. Two possible causes of this are a performance problem from the actors that struggled to imprint a specific emotion on the voice while concentrating on the visual interpretation, or human limitations to recognise emotions with just the voice as input.

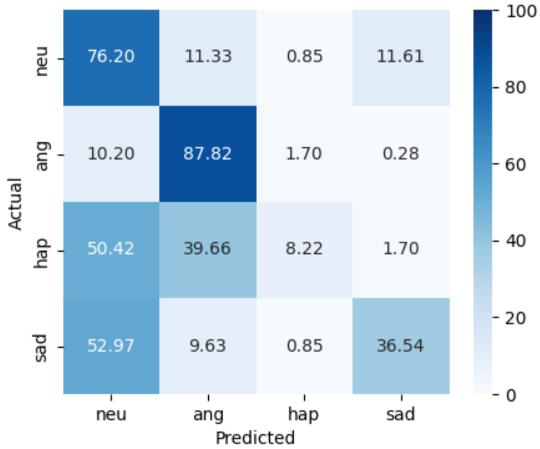
Experiment with Emo-DB: Due to the low amount of utterances on Emo-DB and the imbalance of the dataset, it was tested with 62 utterances per class, totalling 248 utterances. Surprisingly, the model performed better on this German dataset than on CREMA-D, an English dataset, with neutral being the easiest emotion to classify with above 90% accuracy. Both anger and happiness show acceptable results. However, sadness was frequently misclassified as neutral or happiness (31% and 60% respectively). Similar to the case of happiness in CREMA-D, it is rarely predicted.



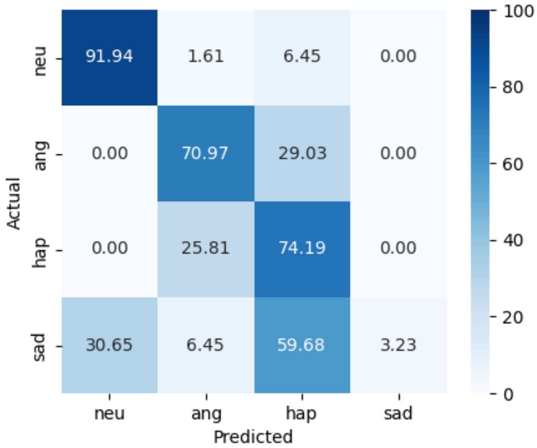
((a)) IEMOCAP



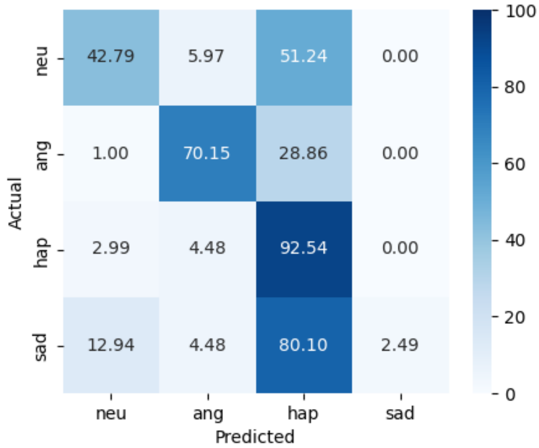
((b)) CREMA-D (self-reported)



((c)) CREMA-D (rated)



((d)) Emo-DB



((e)) ShEMO

Figure 5.1: Confusion matrices for the classification speech model tested on each dataset.

Experiment with ShEMO: This Persian dataset was tested with 201 utterances per class, totalling 804 inputs. The model performed worse on ShEMO than on Emo-DB (see table 5.1. While this is likely due to closer linguistic similarities between English and German, it is not possible to rule out the influence of a statistical shift in the data. The model showed good classification capabilities between neutral and anger but often misclassified inputs as happiness, especially those marked as sadness or neutral, where the true false positives reach 80% and 51% of the total inputs of both emotions. Similarly to the German dataset, sadness is rarely predicted and is often confused with happiness or neutral.

Overall, the model tended to classify English inputs as neutral, but this tendency shifted to happiness in other languages, even those as different as German and Persian. The model also predicted anger consistently but struggled to predict sadness, especially in non-trained languages.

	IEMOCAP	CREMA-D (self-reported)	CREMA-D (rated)	Emo-DB	ShEMO
Accuracy	0.92206	0.51656	0.52195	0.60081	0.51990
F1 score (macro)	0.92266	0.44441	0.46672	0.53310	0.46678

Table 5.1: Results of the classification speech model on each dataset.

5.2 Speech model for regression

This experiment evaluated the performance of the speech unimodal model for regression. The model was tested exclusively on the IEMOCAP dataset, as it was the only accessible dataset with labeled dimensional emotions. For context, it must be noted that in IEMOCAP each utterance was rated by at least three human annotators, with scores ranging from 0 to 5 for each dimension. Although the values are theoretically continuous, the annotators used discrete increments of 0.5. The different scores were then averaged to determine the final dimensional label for each utterance. As shown in figure 5.2, the

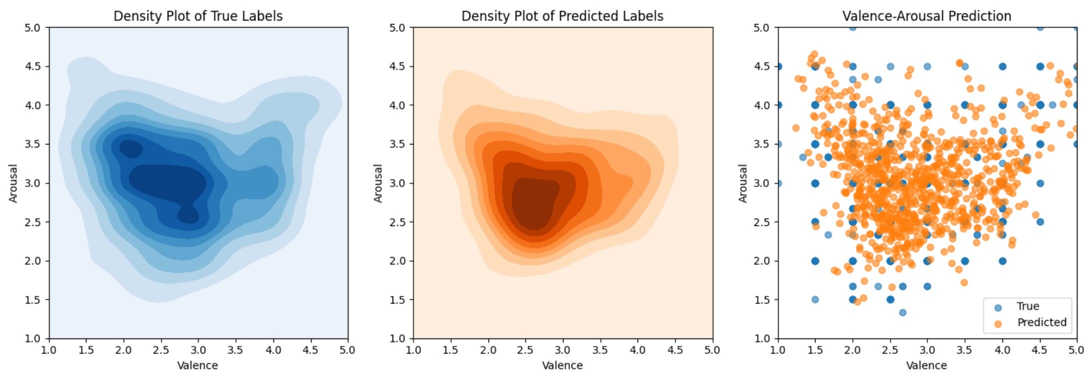


Figure 5.2: Distribution of true and predicted dimensional values.

distribution of the predictions followed, in general terms, the distribution of the true labels. This can also be seen when plotting the true and predicted dimensional values, as shown in figure 5.3. Intuitively, it can be understood that high absolute scores of valence will be related to high levels of arousal, as expressed emotions on the extreme of the valence spectre are generally shown in an emphatic manner.

In figure 5.4, the error distribution for valence and arousal is displayed. We can observe that both error distributions followed a normal pattern. Table 5.2 provides detailed characteristics of these distributions. Generally, the model predicted the arousal score slightly better than the valence, indicated by a smaller maximum error and reduced standard deviation. The slightly lower correlation for arousal suggests a slight misalignment between the prediction and true labels.

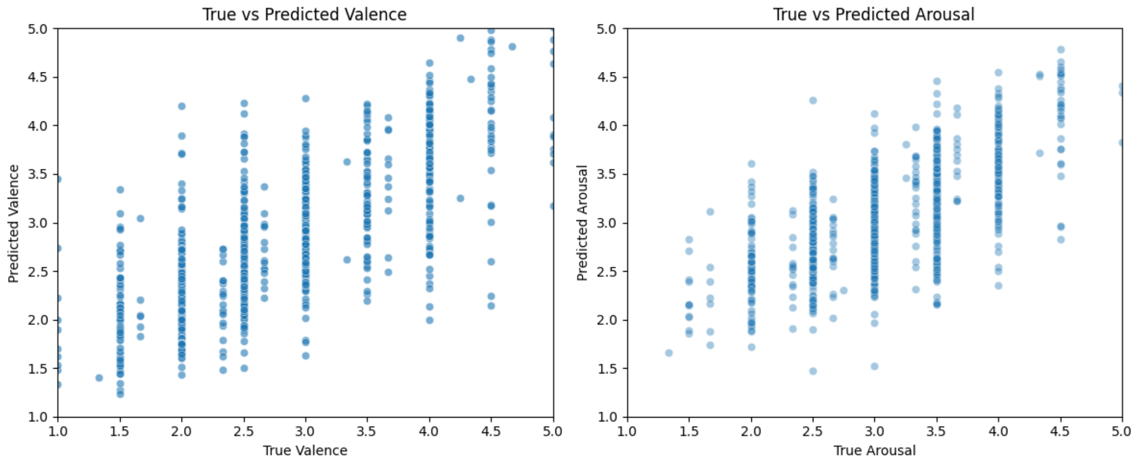


Figure 5.3: Correlation between true and predicted dimensional values.

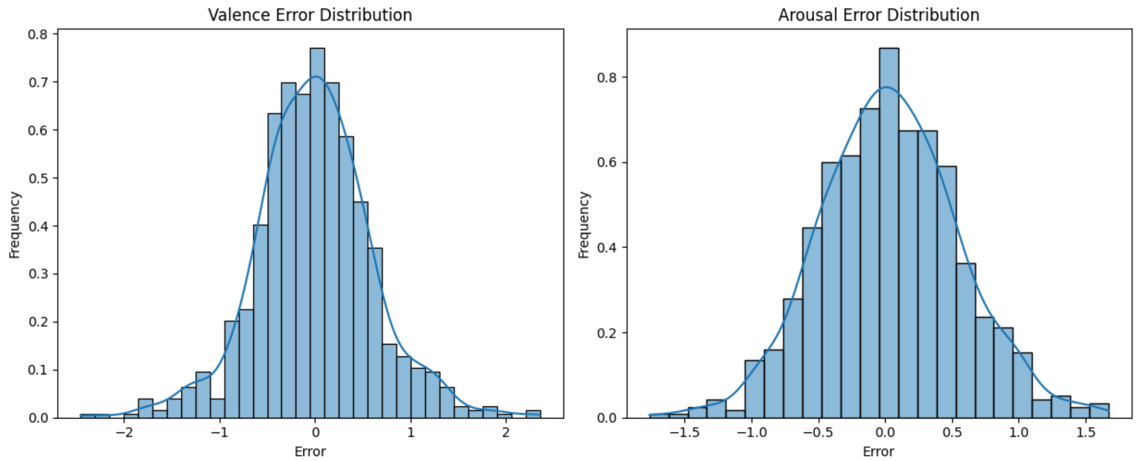


Figure 5.4: Error distribution for valence and arousal.

	Valence	Arousal
MSE	0.3696	0.2715
Max error	2.4541	1.7625
Mean	-0.0078	0.0352
Standard deviation	0.6079	0.5199
Correlation	0.7293	0.6693

Table 5.2: Results of the regression speech model on IEMOCAP.

5.3 Speech multiobjective model

This experiment assessed the performance of the multiobjective speech unimodal model for both regression and classification tasks. All tests were conducted under the same conditions and using the same subsets as the previous unimodal experiments for their respective tasks. This approach ensured consistency and comparability of results across different experimental setups. Figure 5.8 shows the confusion matrices for the multiobjective speech model tested on each dataset. Overall, the multiobjective model performed both tasks adequately but with slightly different performances compared to separately trained models. In regression, as for the case for the exclusively-regression model, the distribution of

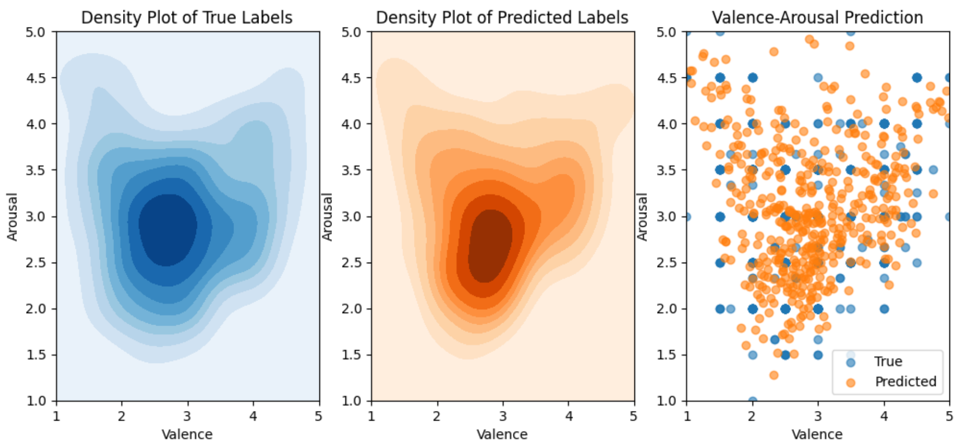


Figure 5.5: Distribution of true and predicted dimensional values.

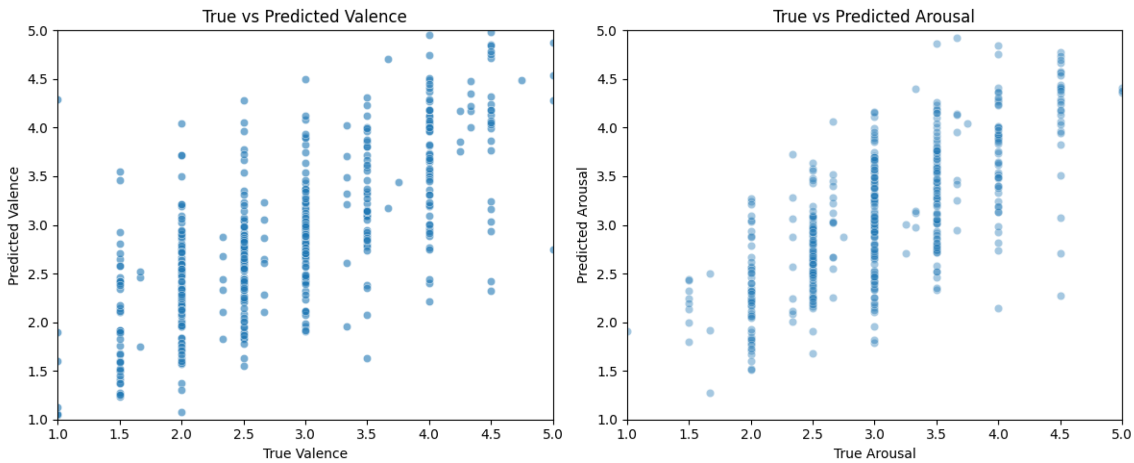


Figure 5.6: Correlation between true and predicted dimensional values.

predicted values follows the actual values, as shown in figure 5.5 and figure 5.6. It can be seen in table 5.3 that the model's performance was slightly worse, with a 0.075 increase in MSE for valence and a 0.034 increase in MSE for arousal. The only improvement was an increase in the correlation of arousal, correcting the previous slight misalignment.

In classification, as shown in table 5.4, the results were mixed. For English datasets, the multiobjective model provided more balanced results across all classes, increasing predictions for happiness, sadness

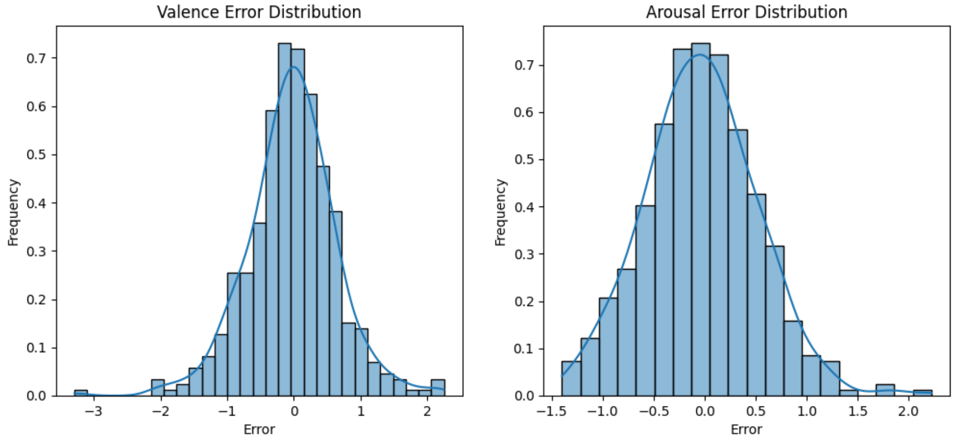


Figure 5.7: Error distribution for valence and arousal.

and anger but also raising the error rate for neutral. This tendency reversed for non-English datasets, particularly with ShEMO, where the model decreased predictions for happiness and sadness, often misclassifying these emotions as neutral. In the case of sadness, it was not predicted even once in all tested inputs.

Compared to the single objective classification model, the multiobjective model reduced its tendency to classify inputs as neutral in English but exhibited this behavior more in other languages. The model improved its ability to predict sadness in English but lost almost completely this ability in non-trained languages. Overall, the multiobjective model shows worse results in its regression task when tested on

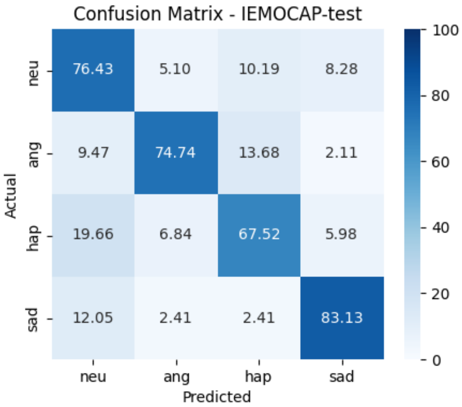
	Valence	Arousal
MSE	0.4443	0.3054
Max error	3.2874	2.2233
Mean	-0.0306	-0.0360
Standard deviation	0.6659	0.5515
Correlation	0.7005	0.7126

Table 5.3: Dimensional results of the multiobjective speech model on IEMOCAP.

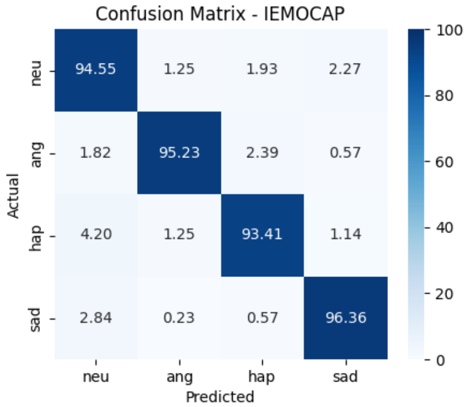
	IEMOCAP (full)	IEMOCAP (test)	CREMA-D (self-reported)	CREMA-D (rated)	Emo-DB	ShEMO
Accuracy	0.94886	0.75	0.51539	0.53754	0.61694	0.56095
F1 score (Macro)	0.94896	0.75331	0.45723	0.50064	0.53472	0.48596

Table 5.4: Classification results of the multiobjective speech model on each dataset.

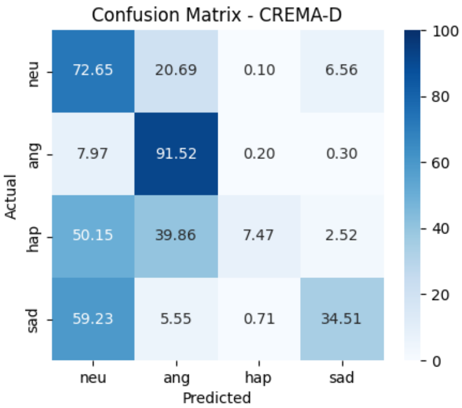
the IEMOCAP dataset, but an improvement on the classification task. While it does not perform as well on the testing subdataset of IEMOCAP and just slightly worse on the self-reported CREMA-D it is of note that it shows better results on all other tests, including in all the non-English datasets.



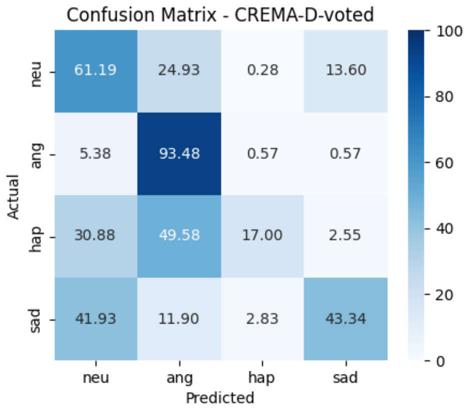
((a)) IEMOCAP (testing data)



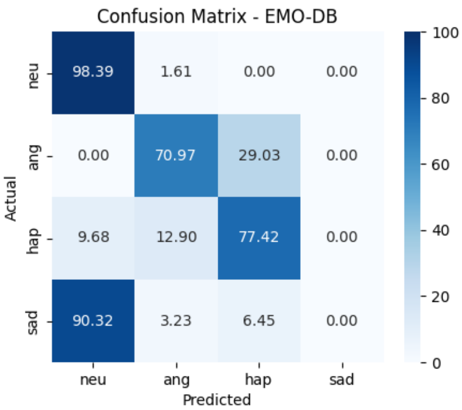
((b)) IEMOCAP (full dataset)



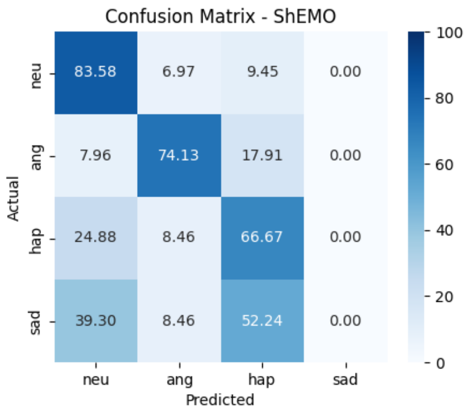
((c)) CREMA-D (self-reported)



((d)) CREMA-D (rated)



((e)) Emo-DB



((f)) ShEMO

Figure 5.8: Confusion matrices for the classification results of the multiobjective speech model tested on each dataset.

Dataset	Classification model		Multiobjective model	
	Accuracy	F1 score (max)	Accuracy	F1 score (max)
IEMOCAP (full)	0.92206	0.92266	**0.94886**	0.94896
IEMOCAP (test)	**0.78700**	-	0.75000	0.75331
CREMA-D (self-reported)	**0.51656**	0.44441	0.51539	0.45723
CREMA-D (rated)	0.52195	0.46672	**0.53754**	0.50064
Emo-DB	0.60081	0.53310	**0.61694**	0.53472
ShEMO	0.51990	0.46678	**0.56095**	0.48596

Table 5.5: Results of the multimodal and unimodal models tested on CREMA-D.

5.4 Chapter summary

In this chapter I present comprehensive experiments evaluating various models for emotion recognition across multiple datasets. The experiments include tests on unimodal speech models for classification and regression, as well as a multiobjective task combining both of them. The speech classification model was tested on IEMOCAP, CREMA-D, Emo-DB, and ShEMO datasets, showing varying degrees of accuracy and robustness, particularly highlighting challenges in recognizing emotions like sadness and happiness across different languages. The chapter further examines the performance of the regression model, although there is no baseline to compare against. Additionally, the chapter evaluates the impact of the multiobjective approach, emphasizing the improvements in robustness and accuracy achieved through this method.

CHAPTER 6

Discussion

The experiments conducted in this study underscore the complexity and multifaceted nature of emotion recognition systems. The performance of these systems is influenced by several factors, including the modality, the language of the dataset, and the internal representation of the emotional state. It has been proven in previous investigations that certain emotions are more readily recognized by specific modalities. For instance, Anger is more accurately detected through audio, while happiness is better recognized through video. Sadness, which generally performs poorly in unimodal models, shows significant improvement when both audio and video inputs are utilized (as mentioned in the CREMA-D paper [4]). All of these tendencies are observed in the model. Although this thesis did not focus on multimodality, it is important to mention that strong skews in emotion recognition are to be expected due to only using the speech modality.

Interestingly, experiments with non-English datasets challenge many of the assumptions drawn from the English dataset analyses. The models exhibit better performance, both in accuracy and f1 score, on non-English datasets compared to CREMA-D, but with notable differences in probability distribution. While it is not possible to affirm without further research if it is a characteristic of the IEMOCAP dataset with which all the speech models were trained, the CREMA-D dataset or the English language, it is clear that for non-CREMA-D/non-English tests the model behaved in a different way, consistent across all other datasets. The preference for predicting the neutral class in English datasets shifts to happiness in non-English datasets, often accompanied by a decrease in the accuracy of other emotions, such as neutral and anger. Sadness predictions are particularly affected, almost disappearing in non-English datasets. Despite individual differences across datasets, a consistent trend emerges across languages as diverse as German and Persian. This suggests the potential for developing more robust models that can adapt to these trends and improve performance across multiple languages. The experiments also highlight the benefits of incorporating multiple internal representations of emotions in the final prediction. The results demonstrate that using emotional dimensions to complement emotional classes helps define clearer classification boundaries. The multiobjective speech model, which combines classification and regression tasks, demonstrates a stabilizing influence, balancing the scores across different emotions. This effect is evident across languages, moderating the trend towards happiness, although the model seems to lose the ability to predict sadness entirely. The multiobjective speech model achieved lower accuracy only on the test subdataset of IEMOCAP and on the self-reported CREMA-D, which is of less interest to us than the rated one. On all other datasets, including non-English ones, the multiobjective model shows better accuracy. In the case of ShEMO, this difference is especially significant, improving the results by 4%. On all tests, the F1 score is consistently better by using the multiobjective task, which points to a more regular distribution of the correct prediction across emotions.

In summary, the results illustrate the complexities and challenges of emotion recognition across different languages. Incorporating multiple internal representations of emotions shows promise in improving model performance, but further research is needed to ensure their applicability across languages and datasets and to refine them further.

CHAPTER 7

Conclusions and future recommendations

7.1 Conclusions

This master thesis aimed to explore the enhancement of emotion recognition systems, the performance disparities of these systems in low-resource languages, and the impact of different emotional representations in speech systems and their generalization capabilities across languages. The findings provide clear answers to these research questions.

1. **What are the performance disparities of emotion recognition systems when applied to low-resource language datasets?**

   The performance of emotion recognition systems in low-resource language datasets showed notable disparities compared to English datasets. While speech models demonstrated better overall accuracy and F1 scores on non-English datasets than on CREMA-D, the distribution of predicted probabilities differed significantly. In non-English datasets, the models exhibited a strong preference for predicting happiness over other emotions, which led to a decrease in the accuracy of neutral and anger predictions and almost a complete absence of sadness predictions. This trend was consistent across diverse languages, including German and Persian (considered both poor-resource languages in this context due to not being used for training in any model), highlighting the need for models that can adapt to different linguistic contexts and maintain their performance.

2. **How does the emotional representation of speech affect speech systems and its capability to generalize across languages?**

   The incorporation of multiple internal representations of emotions, such as emotional dimensions alongside emotional classes, proved beneficial for speech systems. The multiobjective model, which combined classification and regression tasks, helped define clearer classification boundaries and balanced the prediction scores across different emotions. This stabilizing effect was observed across languages, moderating the tendency towards over-predicting neutral and happiness. However, there were some cases where the method did not help to balance emotions, such as sadness in non-English datasets. Speech models that included the multiobjective task achieved the highest accuracy on most datasets and F1 scores on all experiments. While further research is necessary to refine these approaches for better cross-linguistic applicability, these findings suggest that using multiple emotional representations can enhance the performance of the models.

7.2 Future recommendations

Despite the progress made in this thesis, several avenues for future research remain unexplored due to time constraints, limited resources, or being beyond the scope of this study. The results and discussions have highlighted areas that warrant further investigation to provide more conclusive evidence. Notably, the observed differences in behavior between English and non-English datasets and the varying performance of models across these datasets require more extensive experimentation. Testing the models with multiple datasets within the same language and across different languages will help establish solid foundations and mitigate the effects of particularly noisy datasets.

Both multimodality and multiobjective techniques have separately shown potential for enhancing emotion recognition models, but it would be interesting to investigate their applicability when combined on the same model. Further investigation would require applying the multiobjective task on the speech module of a multimodality model and check the evolution of its results

The primary aim of this project was to contribute to the ongoing research on emotion recognition due to its vast social and economic potential. Although it has developed some end-to-end applications and explored various methods to address current technological challenges, the results showed that we are still far from completely solving these issues. However, I believe that the technology is mature enough for limited real-life applications. Applying the ideas presented in this thesis to perfect or build applications in real-world environments is beyond the scope of this study, but I hope that this work will inspire and contribute to future advancements in this field.

References

[1] Alexei Baevski et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." In: *Advances in neural information processing systems* 33 (2020), pages 12449–12460.

[2] Felix Burkhardt et al. "A database of German emotional speech." In: *Interspeech*. Volume 5. 2005, pages 1517–1520.

[3] Carlos Busso et al. "IEMOCAP: Interactive emotional dyadic motion capture database." In: *Language resources and evaluation* 42 (2008), pages 335–359.

[4] Houwei Cao et al. "Crema-d: Crowd-sourced emotional multimodal actors dataset." In: *IEEE transactions on affective computing* 5.4 (2014), pages 377–390.

[5] Sanyuan Chen et al. "Wavlm: Large-scale self-supervised pre-training for full stack speech processing." In: *IEEE Journal of Selected Topics in Signal Processing* 16.6 (2022), pages 1505–1518.

[6] Sneha Das et al. "Continuous metric learning for transferable speech emotion recognition and embedding across low-resource languages." In: *arXiv preprint arXiv:2203.14867* (2022).

[7] Paul Ekman and Wallace V Friesen. "Constants across cultures in the face and emotion." In: *Journal of personality and social psychology* 17.2 (1971), page 124.

[8] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." In: *Pattern recognition* 44.3 (2011), pages 572–587.

[9] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial." In: *JMIR mental health* 4.2 (2017), e7785.

[10] Hongliang Fu et al. "Cross-Corpus Speech Emotion Recognition Based on Multi-Task Learning and Subdomain Adaptation." In: *Entropy* 25.1 (2023), page 124.

[11] Wei-Ning Hsu et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pages 3451–3460.

[12] Liliana Laranjo et al. "Conversational agents in healthcare: a systematic review." In: *Journal of the American Medical Informatics Association* 25.9 (2018), pages 1248–1258.

[13] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *nature* 521.7553 (2015), pages 436–444.

[14] Shan Li and Weihong Deng. "Deep facial expression recognition: A survey." In: *IEEE transactions on affective computing* 13.3 (2020), pages 1195–1215.

[15] Yong Li, Yuanzhi Wang, and Zhen Cui. "Decoupled multimodal distilling for emotion recognition." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pages 6631–6640.

[16] Hailun Lian et al. "A Survey of Deep Learning-Based Multimodal Emotion Recognition: Speech, Text, and Face." In: *Entropy* 25.10 (2023), page 1440.

[17] Fen Liu et al. "A multi-modal fusion method based on higher-order orthogonal iteration decomposition." In: *Entropy* 23.10 (2021), page 1349.

[18] Feng Liu et al. "Lgcct: A light gated and crossed complementation transformer for multimodal speech emotion recognition." In: *Entropy* 24.7 (2022), page 1010.

[19] Cheng Lu et al. "Progressively discriminative transfer network for cross-corpus speech emotion recognition." In: *Entropy* 24.8 (2022), page 1046.

[20] Omid Mohamad Nezami, Paria Jamshid Lou, and Mansoureh Karami. "ShEMO: a large-scale validated database for Persian speech emotion detection." In: *Language Resources and Evaluation* 53 (2019), pages 1–16.

[21] Pierre Philip et al. "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders." In: *Scientific reports* 7.1 (2017), page 42656.

[22] Mirco Ravanelli et al. *SpeechBrain: A General-Purpose Speech Toolkit.* arXiv:2106.04624. 2021. arXiv: `2106.04624 [eess.AS]`.

[23] James A Russell. "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6 (1980), page 1161.

[24] Elena Ryumina, Denis Dresvyanskiy, and Alexey Karpov. "In search of a robust facial expressions recognition model: A large-scale visual cross-corpus study." In: *Neurocomputing* 514 (2022), pages 435–450.

[25] Steffen Schneider et al. "wav2vec: Unsupervised pre-training for speech recognition." In: *arXiv preprint arXiv:1904.05862* (2019).

[26] Shadi Shaheen et al. "Emotion recognition from text based on automatically generated rules." In: *2014 IEEE International Conference on Data Mining Workshop*. IEEE. 2014, pages 383–392.

[27] Yuntao Shou et al. "Conversational emotion recognition studies based on graph convolutional neural networks and a dependent syntactic analysis." In: *Neurocomputing* 501 (2022), pages 629–639.

[28] Monorama Swain, Aurobinda Routray, and Prithviraj Kabisatpathy. "Databases, features and classifiers for speech emotion recognition: a review." In: *International Journal of Speech Technology* 21 (2018), pages 93–120.

[29] Michel F Valstar et al. "The first facial expression recognition and analysis challenge." In: *2011 IEEE international conference on automatic face & gesture recognition (FG)*. IEEE. 2011, pages 921–926.

[30] Joseph Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine." In: *Communications of the ACM* 9.1 (1966), pages 36–45.

[31] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. "Emotion recognition from text using semantic labels and separable mixture models." In: *ACM transactions on Asian language information processing (TALIP)* 5.2 (2006), pages 165–183.

[32] Hongling Yang et al. "Multimodal Attention Dynamic Fusion Network for Facial Micro-Expression Recognition." In: *Entropy* 25.9 (2023), page 1246.

[33] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. "Tag-assisted multimodal sentiment analysis under uncertain missing modalities." In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pages 1545–1554.

[34] Ligang Zhang and Dian Tjondronegoro. "Facial expression recognition using facial movement features." In: *IEEE transactions on affective computing* 2.4 (2011), pages 219–229.

[35] Shiqing Zhang et al. "Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects." In: *Expert Systems with Applications* (2023), page 121692.

[36] Yuan Zong et al. "Adapting Multiple Distributions for Bridging Emotions from Different Speech Corpora." In: *Entropy* 24.9 (2022), page 1250.