

DSpace Institution

DSpace Repository

<http://dspace.org>

Computer Science

thesis

2024-11

Dimensional Amharic Speech Emotion Recognition Model Using Deep Learning

ASSEFA, BELAY AWEKE

<http://ir.bdu.edu.et/handle/123456789/16439>

Downloaded from DSpace Repository, DSpace Institution's institutional repository



BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
FACULTY OF COMPUTING

MSc Thesis on:
Dimensional Amharic Speech Emotion Recognition Model Using Deep
Learning

BY: - ASSEFA BELAY AWEKE

Nov, 2024
Bahir Dar, Ethiopia



BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF RESEARCH AND GRADUATE STUDIES

Dimensional Amharic Speech Emotion Recognition model using Deep Learning

By

Assefa Belay Aweke

A thesis submitted

In Partial Fulfilment of the Requirements for the Degree of Master of Science in
Computer Science

Advisor: - Tesfa Tegegne (Assoc. prof)

©Assefa Belay Aweke

Nov, 2024
Bahir Dar, Ethiopia

DECLARATION

This is to certify that the thesis entitled “Dimensional Amharic Speech Emotion Recognition Using Deep Learning”, submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science under the Faculty of Computing, Bahir Dar Institute of Technology, is a record of original work carried out by me and has never been submitted to this or any other institution to get any other degree or certificates. The assistance and help I received during this investigation have been duly acknowledged.

Assefa Belay Aweke
Name of the Candidate


Signature

Nov 12, 2024 G.C
Date

©2024


Assefa Belay Aweke

ALL RIGHT RESERVED

BAHIR DAR UNIVERSITY
BAHIR DAR INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
COMPUTING FACULTY

Approval of thesis for defense result

I hereby confirm that the changes required by the examiners have been carried out and incorporated in the final thesis.

Name of Student: Assefa Belay Signature  Date Nov, 12, 2024

As members of the board of examiners, we examined this thesis entitled "Dimensional Amharic Speech Emotion Recognition Model Using Deep Learning" by Assefa Belay. We hereby certify that the thesis is accepted for fulfilling the requirements for the award of the degree of Masters of Science in Computer science.

Board of Examiners

Name of Advisor	Signature	Date
<u>Dr. Tesfa Tegegne (Assoc.prof)</u>	<u></u>	<u>NOV 19, 2024</u>
Name of External examiner	Signature	Date
<u>Dr. Michael Melese (PhD)</u>	<u></u>	<u>Nov, 12, 2024</u>
Name of Internal Examiner	Signature	Date
<u>Mr. Tamir Anteneh (Asst.Prof)</u>	<u></u>	<u>NOV 14, 2024</u>
Name of Chairperson	Signature	Date
<u>Mr. Alemu Kumilachew (Asst.Prof)</u>	<u></u>	<u>NOV 14, 2024</u>
Name of Chair Holder	Signature	Date
<u>Mss. Kidist Meshesha</u>	<u></u>	<u>NOV 14, 2024</u>
Name of Faculty Dean	Signature	Date
<u>Dr. Tesfa Tegegne (Assoc.prof)</u>	<u></u>	<u>NOV 15, 2024</u>



This thesis work is dedicated to my grandparents
Mr. Gagnaw Mengistie and Mrs. Simegn Birlie.

ACKNOWLEDGMENT

First of all, I would like to give limitless gratitude to the Almighty God and Saint Mary for allowing me alive and keeping me safe from different emergencies. Next, I would like to extend my heart gratitude to Tesfa Tegegne (Assoc. prof) for his guidance, support, thorough analysis, and invaluable feedback throughout the thesis development process. This thesis is not reached at this stage without your support. I also need to thank Mekdela Amba University for giving me this scholarship opportunity and Bahir Dar University for fulfilling my fundamental requirements. Last but not least, I would like to thank my family, especially my elder brother Mr. Getasew Belay for their support, advice, and motivation.

ABSTRACT

Amharic language is the official language of Ethiopia and is spoken by millions of people within the country and outside the country. This is a more important language as it is the second most widely spoken Semitic language next to Arabic. Hence, developing dimensional SER is a promising work. Recently, researchers had researched emotion recognition based on a categorical approach. However, a categorical approach such as classifying emotion based on each class is unable to represent each emotion as emotion have more than 68 classes. The researcher also approves that dimensional emotion recognition can represent more nuanced than categorical emotion. Although dimensional emotion recognition is promising, the valence result is lower than arousal and dominance as the people use the same sound to describe their pleasure and displeasure. The researchers were challenged to recognize angry and happy emotions as humans speak the same way. To handle this challenge, the researcher uses different mechanisms such as linguistic and acoustic features. In this research, we annotated ASED dataset with VAD annotation by annotation team. This study aimed to conduct dimensional Amharic SER model to overcome the above problems. Furthermore, we used deep learning models such as LSTM and BiLSTM and identified the most suitable deep learning models to recognize Amharic emotions dimensionally. Our model performance on categorical speech emotion recognition is 98% and mean square error for dimension of valence, arousal, and dominance with bimodal feature is 0.0081, 0.0655, and 0.0239 respectively. The mean absolute error of valence, arousal, and dominance is 0.00049, 0.0321, and 0.0061 with concordance correlation coefficient of 1.000, 0.8738, and 0.9775 respectively.

Keywords: Emotion Recognition, Amharic Language, VAD Annotation, Deep Learning, Dimensional Emotion Recognition, Speech Processing, Text Processing

LIST OF ABBREVIATIONS

ASED	Amharic Speech Emotion Dataset
ASER	Amharic Speech Emotion Recognition Model for acoustic feature
BiLSTM	Bimodal LSTM
CASER	Categorical Amharic Speech Emotion Recognition Model
CNN	Convolutional Neural Network
DASER	Dimensional Amharic Speech Emotion Recognition Model for bimodal feature
ECOC	Error Correcting Output Code
EL-HDAF	Ensemble learning by high-dimensional acoustic features
Emo-DB	Berlin Emotional Speech Database
FSVM	Fuzzy SVM
HAF	Hybrid acoustic features
KNN	K- Nearest Neighbor
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MLP	Multilayer perceptron
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
PA-Net	Parallel attention recurrent network
RAVDESS	Ryerson Audio-Visual Database of Emotional Speech and Song
RECOLA	Remote collaborative and affective interactions
ResNet50	Residual Network
RNN	Recurrent Neural Network
SBL-DF	Speech emotion recognition using supervised Bayes learning on digital features
SER	Speech Emotion Recognition
SEWA	Sentiment analysis in the wild
SVM	Support Vector Machine
TESS	Toronto Emotional Speech Set
VGG	Visual Geometry Group

TABLE OF CONTENTS

ACKNOWLEDGMENT	vi
ABSTRACT	vii
LIST OF ABBREVIATIONS	viii
LIST OF TABLES	xii
LIST OF FIGURES	xiii
CHAPTER ONE: INTRODUCTION	1
1.1 Introduction	1
1.2 Background.....	1
1.3 Motivation	2
1.4 Statement of the problem.....	3
1.5 Objectives of the study	5
1.5.1 General Objective.....	5
1.5.2 Specific Objectives.....	5
1.6 Scope of the study.....	5
1.7 Significance of the study	6
1.8 Organization of the thesis.....	7
CHAPTER TWO: LITERATURE REVIEW	8
2.1 Overview	8
2.2 Speech Emotion Recognition	8
2.2.1 Theoretical Foundations of SER	8
2.3 Discrete Speech Emotion Recognition	10
2.2.1. Challenges and Limitations in Discrete SER	10
2.3. Dimensional SER	11
2.3.1. Challenges and Limitations in Dimensional SER.....	13
2.4. Challenges and Considerations in Amharic SER	13
2.4.1. Dataset Availability and Annotation Challenges	13
2.6 Speech processing.....	14
2.7 Feature extraction technique.....	14
2.6.1. Mel Frequency Cepstral Coefficients (MFCC).....	16
2.6.2. Zero Crossing Rate (ZCR)	16

2.6.3. Root Mean Square (RMS).....	16
2.8 Deep learning Algorithm	16
2.7.1. Long Short-Term Memory (LSTM).....	17
2.7.2. Bidirectional Long Short-Term Memory (BiLSTM)	19
2.9 Evaluation Metrics and Methodologies in Speech Emotion Recognition	20
2.8.1. Discrete Emotion Recognition	20
2.8.2. Dimensional Emotion Recognition	20
2.10 Related Work.....	21
2.10. Summary literature review.....	24
CHAPTER THREE: Methodology (Materials and Methods).....	28
3.1 Overview	28
3.2 Dataset preparation	28
3.3 Dataset Annotation	29
3.3.1. Discreate To Dimensional Model Relationship	31
3.4 System Architecture	32
3.4.1 Categorical Amharic SER Model Design	32
3.4.2 Dimensional Amharic Speech Emotion Recognition Model Design.....	37
3.5 Text preprocessing.....	39
3.5.1 Tokenization and Normalization	40
3.5.2 Text Sequencing	40
3.5.3 Sequence Padding	40
3.5.4 Word Embedding	41
3.6 Model Development	42
3.6.1 Categorical Amharic SER Model Development	42
3.6.2 Dimensional Amharic SER Model Development with Acoustic Feature... 43	
3.6.3 Dimensional Amharic SER Model Development with Linguistic and Acoustic Feature.....	44
3.7 Development Tools.....	46
3.8 Evaluation Metrics.....	48
3.8.1 Evaluation Metrics for Categorical Amharic SER model.....	48
3.11.2. Evaluation Metrics for Dimensional Amharic SER model.....	49

3.9	Summary.....	51
CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION.....		52
4.1	Overview	52
4.2.1.	Dataset Splitting	52
4.2.2.	Model Hyperparameter.....	53
4.3	Model Configuration	54
4.3.1	Dimensional Amharic SER using acoustic feature with BiLSTM.....	55
4.3.2	Dimensional Amharic SER using acoustic and linguistic feature with BiLSTM	56
4.4	Experimental Results.....	56
4.4.1.	Categorical Amharic SER model using acoustic feature	57
4.4.2.	Dimensional Amharic SER model using merely acoustic feature with LSTM and BiLSTM	60
4.4.3.	Dimensional Amharic SER model using bimodal feature and BiLSTM algorithm	64
4.4.4.	Overall Model Performance	69
4.5	Discussion.....	69
4.5.1	Experimental Result of Categorical SER	71
4.5.2.	Experimental Result of Dimensional SER using Acoustic Feature	71
4.5.3.	Experimental Result of Dimensional SER using Acoustic and Linguistic Feature.....	72
4.6	Error Analysis.....	73
4.6.1	Categorical Amharic SER model performance error analysis	73
4.6.2	Dimensional Amharic SER using merely acoustic feature and bimodal features model performance error analysis	73
4.7	Answering the research questions	73
CHAPTER FIVE: CONCLUSION AND RECOMMENDATION		76
5.1	Conclusion	76
5.2	Future Work and Recommendation.....	78
REFERENCE.....		80

LIST OF TABLES

Table 2.1: literature review summary.....	24
Table 3.1: Amharic SER dataset	29
Table 3.2: Categorical and dimensional emotion relationship	32
Table 4.1: Dataset splitting for training, validation, and testing	53
Table 4.2: Hyper parameter used in this study	54
Table 4.3: overview of model performance categorical and dimensional along with acoustic feature only and bimodal feature	70

LIST OF FIGURES

Figure 2.1: Valence arousal (2D) emotion representation.....	11
Figure 2.2: Three-dimensional space emotion representation.....	12
Figure 2.3: LSTM network layer.....	18
Figure 2.4: BiLSTM architecture with activation function [34]	19
Figure 3.1: ASER dataset visualization.....	29
Figure 3.2: Categorical Amharic speech emotion recognition model design	33
Figure 3.3: Data augmentation using white noise addition technique	35
Figure 3.4: Data augmentation using time shifting technique.....	36
Figure 3.5: Data augmentation using pitch shifting technique.....	37
Figure 3.6: Dimensional SER model design without linguistic feature	38
Figure 3.7: Dimensional SER model with linguistic feature.....	39
Figure 3.8: Architecture of categorical SER model to classify five emotions	43
Figure 3.9: Architecture of dimensional SER using acoustic feature	44
Figure 3.10: Architecture of dimensional SER using both acoustic and linguistic feature	46
Figure 4.1: Dimensional Amharic SER using acoustic feature model configuration.....	55
Figure 4.2: Dimensional Amharic SER model configuration using both acoustic and linguistic feature with BiLSTM	57
Figure 4.3: Training and testing accuracy of categorical model using feature transformation technique.....	58
Figure 4.4: Confusion matrix analysis for categorical Amharic SER.....	58
Figure 4.5: The model performance on test data correctly and incorrectly classified on each emotion class	60
Figure 4.6: Training and validation loss of dimensional Amharic SER using acoustic feature and LSTM algorithm	61
Figure 4.7: Parallel comparison of the model performance with prediction score and actual score for three dimensions.....	62
Figure 4.8: Dimensional Amharic SER model performance using BiLSTM algorithm along with acoustic feature	63
Figure 4.9: Parallel comparison of the model performance with prediction score and actual score for three dimensions.....	64
Figure 4.10: Training and validation loss of dimensional Amharic SER model using bimodal feature and LSTM algorithm.....	65
Figure 4.11: Evaluation metrics for three dimensions.....	65
Figure 4.12: Training and validation loss of the model using bimodal feature with BiLSTM	66
Figure 4.13: Training and validation loss on separate dimension of Amharic SER using acoustic and linguistic feature.....	67

Figure 4.14: Evaluation metrics on each dimensions using bimodal feature with BiLSTM 69

CHAPTER ONE: INTRODUCTION

1.1 Introduction

This chapter lays the groundwork for the research by providing a comprehensive overview. We begin by establishing the background of the research, outlining the existing knowledge, and context that led to this specific investigation. Following this, we explore the motivation for undertaking this research, highlighting the gap in knowledge or unanswered questions that this study address.

Next, we clearly define the problem statement, outlining the specific issue or challenge that this research investigates. This is followed by the research objectives, which detail the precise goals and outcomes the study seeks to achieve. We then define the scope of the research, specifying the boundaries and limitations of the investigation. Finally, the chapter concludes by emphasizing the significance of the study, explaining the potential impact and contribution of the research to the field.

1.2 Background

Emotion recognition is an important part of human communication and interaction, with applications in psychology, banking, call center, video monitoring [1], in-car board systems [2], computer tutoring [3], teaching systems [4], and mental health [2]. The rising usage of technology in our daily lives has increased the demand for accurate and efficient emotion identification systems. Deep learning algorithms have yielded encouraging breakthroughs in the field of emotion recognition in recent years, particularly in the recognition of facial expressions and speech. However, research on dimensional emotion recognition for languages with minimal linguistic resources, such as Amharic, is not present.

Amharic is Ethiopian official language, spoken by millions of people in the country and around the world, and it is a widely used Semitic language next to Arabic [5]. Despite its widespread use, there are few resources and studies on emotion recognition in Amharic [5]. This study intends to close this gap by developing a dimensional emotion identification system for Amharic using deep learning techniques.

Dimensional emotion recognition refers to emotion recognition based on underlying characteristics such as valence (positive or negative), arousal (intensity), and dominance (power control) [3], [6], [7], [8]. In contrast to the traditional categorical approach, which classifies emotions into a limited range of discrete categories [5], this technique provides for a more complex and fine-grained understanding of emotions. Deep learning systems, such as CNN [9] [10] and RNN [8], [10] have shown significant promise in collecting complex patterns and characteristics in a variety of data formats, such as images, audio, and text. We intend to conduct this study using these sophisticated capabilities.

We used the ASED dataset made public by Ephrem et al [11]. Due to the modest size of the ASED dataset and need to simulate the real-world speech, we augment the speech data and annotate it with dimensional emotion labels. Then, we designed architecture to build the model train and evaluate its performance.

1.3 Motivation

We are excited to conduct this research by motivating with the following core ideas. To begin with, emotion is our day-to-day activity, and understanding how it is expressed in Amharic communication make known the complexities of human interaction within Ethiopian culture. Emotion is a universal experience, shaping our interactions, decisions, and well-being. By exploring into the domain of Amharic speech emotion recognition, we capture the essence of emotions in Amharic and contribute to a deeper understanding of how emotions manifest themselves in everyday life.

Amharic, culturally rich language, offers a unique opportunity to explore the diverse range of emotions expressed by its speakers [11]. Through this research, we decipher the linguistic cues, acoustic patterns, and cultural nuances that underlie emotional expressions in Amharic. Moreover, this research aligns with the principles of cultural preservation, representation, and inclusivity. Amharic is not just a means of communication; it carries the heritage, traditions, and history of Ethiopia. By researching Amharic SER, we contribute to the preservation and representation of this important cultural aspect. The findings of this research can empower Ethiopians to express their emotions more authentically in the digital world, fostering cultural pride and promoting diverse linguistic perspectives on a global scale. By exploring the acoustic features, prosodic variations, and

linguistic cues specific to Amharic, we expand the knowledge base and methodologies in SER with dimensional approaches, benefiting both academia and industry applications. The dimensional approach which represents emotion in three dimensions such as valence, arousal, and dominance can represent emotion more nuancedly than categorical emotion as emotion is subjective [12] [13] [1].

Amharic, being a dynamic and socially rich language, offers a novel chance to investigate the different scope of feelings communicated by its speakers. Amharic isn't simply a method for correspondence; it conveys the legacy, customs, and history of Ethiopia. The discoveries of our exploration enable Ethiopians to communicate their feelings all the more genuinely in the computerized world, encouraging social pride and advancing different etymological points of view on a worldwide scale. Besides, by connecting the examination hole in Amharic discourse feeling acknowledgment, we mean to propel the more extensive field of feeling acknowledgment.

The benchmark of the ASED dataset established by Ephrem et al. inspired us to use it and apply a modern dimensional approach to put our contributions.

1.4 Statement of the problem

This study seeks to address issues that contribute to the field of AI. To begin with, present speech emotion detection models and approaches are often constructed for commonly spoken languages such as English, German, French, and so on. These models are not work for under-resourced languages [5][2]. The Amharic language has distinct phonetic, prosody, and cultural context, demanding the use of specialized models to accurately capture and interpret emotions.

Furthermore, recent research in speech emotion detection focuses mostly on categorical classification [5] [11] which limits the ability to capture the nuances and complexities of emotions. The use of dimensional models allows for a fine-grained examination of emotions, providing greater insights into people's emotional states [10][14][6]. The development of dimensional speech emotion recognition models in the Amharic language is currently not present, that limits the study of emotional dynamics in this language.

Recognizing emotions expressed in Amharic speech is important because Amharic is the working language of Ethiopia and frequently spoken Semitic languages [5]. It is useful in many areas of human-computer interaction, including psychology, social sciences, mental health, education, and driving [4][10]. The lack of dimensional emotion recognition for Amharic limits the deployment of such applications. As a result, executing dimensional Amharic SER is useful for assisting Amharic native or non-native speakers and incorporating technology. In addition, investigating the best feature capable of capturing the nuance of Amharic speech feature is promising too. Lack of standardized Amharic speech emotion recognition has the following potential challenges[5], [11].

- Difficulty in developing effective communication and language processing technologies for Amharic speakers.
- Inability to accurately interpret and understand the emotional content of Amharic speech.
- Impediments in developing personalized and tailored emotional support systems for Amharic speakers.
- Hindrance in creating inclusive and culturally sensitive mental health and well-being applications for Amharic speakers.
- Limited progress in creating emotionally intelligent virtual assistants and chatbots for Amharic speakers.
- Difficulty in accurately capturing and analyzing emotional feedback in user experience testing for Amharic language applications and products.
- Limitations in developing inclusive and culturally relevant educational tools that can adapt to the emotional responses and needs of Amharic-speaking learners, potentially affecting learning outcomes.
- Difficulty in accurately interpreting emotional cues in legal and justice systems for Amharic speakers, potentially impacting fair and effective legal processes.

Generally, to address those challenges, this research answers the following research question.

- To what extent does the deep learning approach improve SER?
- To what extent does the dimensional SER vary from the categorical SER?

- What is the effect of linguistic feature on the performance of dimension prediction in dimensional Amharic SER?

1.5 Objectives of the study

1.5.1 General Objective

The general objective of this research is to develop dimensional Amharic speech emotion recognition model using deep-learning approach.

1.5.2 Specific Objectives

To accomplish our study, we have the following specific objectives.

- ★ Prepare and annotate the freely available datasets ASSED based on valence, arousal, and dominance dimensions.
- ★ Develop categorical SER model for Amharic language.
- ★ Develop a dimensional SER model with merely acoustic feature and with both acoustic and linguistic feature for Amharic language.
- ★ Investigate the difference and similarity of dimensional and categorical SER.
- ★ Evaluate the developed dimensional emotion recognition model.

1.6 Scope of the study

The research involves gathering a representative speech of emotional states and linguistic differences found in the Amharic. This information is essential for training and testing the speech-emotion detection system, which ensures its performance across emotional expressions and linguistic patterns. However, the study did not cover the unique context of emotional expression to conduct a substantial grammatical or phonetic examination of Amharic speech. Instead of doing a complete grammatical examination of the language, the emphasis is on understanding and recording emotions in Amharic speech. Furthermore, the study excludes a great detail about non-speech-based emotional recognition modalities, including facial expression analysis, textual emotion, or physiological cues. The study focuses on the identification of emotions expressed through spoken language in Amharic and excludes emotion recognition in non-speech audio signals such as music or non-verbal

vocalizations. To accomplish our study, we include five emotion classes namely happy, sad, fear, normal, and angry.

1.7 Significance of the study

SER in Amharic holds huge potential across a wide range of sectors. By analyzing customer interactions in call centers, companies can gain valuable insights into service quality and customer satisfaction levels. This technology enables organizations to identify and address customer concerns more effectively, leading to improved overall customer experience. Moreover, the integration of SER in chatbots and virtual assistants allows for a more personalized and adopted interaction with users. The system can adapt their responses based on the emotional cues detected in user voices, creating a more engaging and human-like experience for customers.

In the dominion of mental health and well-being, SER play an important role in mood-tracking applications. By analyzing changes in a user emotional state through their voice, mental health apps can provide personalized support and interventions. This technology can support individuals to understand and manage their emotions, leading to improved mental well-being and self-awareness. Additionally, SER can be utilized in speech therapy to assess and monitor patients' emotional expression during therapy sessions. Therapists can track improvements in patients' ability to convey emotions through speech, enhancing the effectiveness of therapy sessions and facilitating better outcomes for individuals with communication disorders.

Additionally, the application of SER in market research and consumer sentiment analysis offers valuable insights into consumer behavior and preferences. By analyzing emotional cues in customer feedback and interactions, companies can gain a deeper understanding of consumer sentiment and make informed business decisions. This technology can help businesses tailor their products and services to better meet the emotional needs and preferences of their target audience. Additionally, in the field of education, SER can be integrated into tutoring systems to create emotionally intelligent learning experiences for students. By adapting teaching styles based on students' emotional states, educational

platforms can enhance engagement and learning outcomes, ultimately develop a more personalized and effective learning environment.

Furthermore, in the context of public speaking and presentation skill coaching, a dimensional Amharic SER system can provide valuable feedback and insights to speakers. By analyzing the emotional delivery and nuances in the speaker voice, the system can offer personalized coaching and recommendations to improve the speaker ability to convey confidence, engagement, and persuasion. This not only enhances the speaker's presentation skill but also adopts effective communication and connection with the audience.

1.8 Organization of the thesis

The rest of this thesis is structured as follows. The second chapter reviews the literature in the areas of feature extraction, SER in both categorical and dimensional approach, speech preprocessing, text preprocessing, and finding gaps that should be addressed in our research. Chapter three presents the system architecture for three models, along with dataset, feature extraction strategies, pre-processing approaches, and prediction algorithms. In Chapter Four, the performance of the investigation is assessed through discussions and the presentation of the experimental results. In the last chapter, the results are summarized, suggestions are made, and potential directions for further research are discussed.

CHAPTER TWO: LITERATURE REVIEW

2.1 Overview

In this chapter, we focus on the analysis and review of literature related to SER on both categorical and dimensional approaches. We examined various studies and research papers to gain insights into the concept of emotion recognition. Additionally, we explored the literature on speech processing, text processing, feature extraction, and prediction methods used in the overall activities of this study. By reviewing the relevant literature, we enhanced our understanding of SER in dimensional approaches and improve the performance of our analysis.

2.2 Speech Emotion Recognition

SER is a field of study that focuses on automatically identifying and analyzing the emotional content expressed in human speech. It involves using computational techniques, machine or deep learning algorithms, and linguistic and acoustic features extracted from speech signals to classify and recognize the emotional states conveyed by speakers. It has gained significant attention due to its potential applications in various domains, including human-computer interaction, affective computing, psychology, and social sciences [4] [10]. By enabling machines to understand and reply to human feelings, SER has the potential to improve the effectiveness of communication, enhance user experiences, and enable more personalized and adopted interactions.

2.2.1 Theoretical Foundations of SER

SER is grounded in several theoretical frameworks that gives insights into the nature and mechanisms of emotional expression in speech. These theoretical foundations contribute to the development of models and techniques for analyzing and interpreting emotional content in speech signals. This section explores some of the key theoretical frameworks that support the field of SER.

Appraisal Theory: suggests that emotions arise from an individual evaluation or appraisal of a particular event or situation. According to this theory, emotions can be characterized by different dimensions; valence (negative or positive), arousal (activation or intensity), and dominance (power). In SER, appraisal theory provides a foundation for understanding

how emotional states are reflected in speech and how acoustic features can be used to capture and classify these emotions [15].

Dimensional Models of Emotion: Dimensional models propose that emotions can be represented as vectors in a multi-dimensional space. The most common dimensions used in SER include valence, arousal, and dominance [13]. Dimensional models offer a continuous and nuanced representation of emotions, allowing for a more fine-grained analysis and recognition of emotional states in speech.

Psycholinguistic theory: this theory examines the interaction between language and emotions, focusing on the linguistic features and patterns associated with different emotional states. This theory explores how speech prosody, phonetic cues, intonation, speech rate, and vocal expressions contribute to the transmission of emotions. It provides insights into the mechanisms through which emotional information is encoded and decoded in speech, guiding the expansion of computational models for SER [16].

Social and Cultural Factors: Emotions are shaped by social and cultural contexts, and understanding the influence of these factors is important in SER. Social and cultural theories explore how societal norms, cultural practices, and individual differences impact emotional expression and perception. Cultural-specific models in SER consider the unique patterns and expressions of emotions within specific linguistic and cultural communities, such as Amharic SER.

Geneva Emotion Wheel: expands on the basic emotions model and categorizes emotions into a broader range of categories. It consists of eight primary emotion categories, including joy, guilt, sadness, fear, anger, disgust, shame, surprise. Each primary category is further divided into secondary and tertiary categories, providing a more detailed classification of emotions based on their specific characteristics[17].

Ekman's Six Basic Emotions Model: Ekman model is well-known and influential categorical emotion models. It identifies six universal basic emotions: happiness, disgust, sadness, fear, anger, and surprise. These emotions are characterized by distinct speech signal and specific physiological patterns. In SER, Ekman's model serves as a foundation

for recognizing and classifying speech signals into these six basic categories of emotions[18].

2.3 Discrete Speech Emotion Recognition

In the study of SER, there exist two primary approaches: Discrete and Dimensional SER. Discrete SER involves the identification and categorization of emotions in speech into predefined, distinct emotion categories. This method focuses on classifying emotions into clear emotional labels, contrasting with Dimensional SER, which capture the continuous and nuanced aspects of emotions expressed in speech. While Dimensional SER examines into the complexities of emotions, Discrete SER provides a straightforward classification into specific emotional categories, each with its unique characteristics and distinctions. According to the renowned Ekman theory of emotions, there exist six fundamental and universal basic emotions that are universally recognized across cultures.

2.2.1. Challenges and Limitations in Discrete SER

While discrete SER has its merits, it also faces several challenges and limitations. Some potential challenges are the following.

Subjectivity and Variability: Emotions are subjective experiences that can vary significantly among individuals. The same emotion manifest differently in different speakers due to factors such as cultural background, personality traits, and contextual influences. Defining a universal set of discrete emotion labels that accurately captures the richness and diversity of emotional expressions is a challenge [12].

Ambiguity and Overlapping Emotions: Speech signals often contain mixed emotions or emotions that exhibit overlapping characteristics. Accurately categorizing such complex emotional states into discrete labels can be challenging. For instance, a speech segment contains elements of both anger and sadness, making it difficult to assign a single discrete emotion label.

Limited Emotion Coverage: Discrete SER models typically focus on a predefined set of emotion labels. However, this limited coverage not encompass the entirety of emotional

experiences or capture lesser-known or context-specific emotions. It restricts the system's ability to recognize and differentiate emotions that fall outside the predefined categories.

2.3. Dimensional SER

Dimensional SER uses two or three dimensions to represent emotions. Dimensional SER in 2D is an approach that uses a two-dimensional space, characterized by valence and arousal, to classify emotions expressed through speech [19]. A deeper and more complex understanding of emotional states is made possible by this approach, which suggests that emotions are not discrete entities but rather can be represented on a continuous spectrum. Arousal denotes the intensity of the emotional experience, ranging from calmness to excitement, while valence conveys the pleasantness of an emotion, ranging from negative like sad to positive like happiness. Dimensional SER goes beyond typical categorical approaches, which frequently compress complex emotional experiences to discrete labels like happy, sad, or angry. Instead, Dimensional SER plots emotions throughout this two-dimensional space [13] [1].

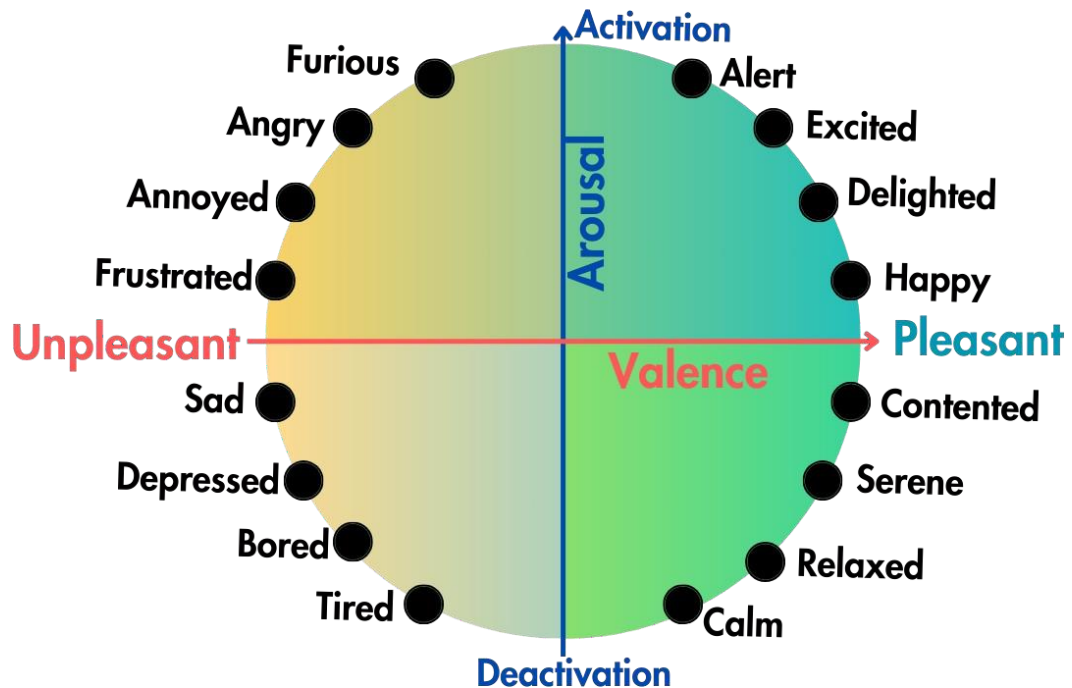


Figure 2.1: Valence arousal (2D) emotion representation

Three-dimensional SER is an improved technique that uses three dimensions such as valence, arousal, and dominance [12]. This approach builds upon the traditional 2D

approaches, adding a layer of complexity to better represent the multifaceted nature of human emotions. Understanding and interpreting emotional states is much improved when the dominant dimension is added to the traditional 2D model of SER, converting it into a 3D representation. To begin with, it used to enhance emotional nuance like dominance, the emotional spectrum becomes more nuanced. Emotions can now not only reflect their positivity or negativity (valence) and intensity (arousal) but also convey a sense of control or submission [12] [1] [13]. For example, feelings like anger (high arousal, negative valence) and fear (high arousal, negative valence) differentiated by their dominance levels, leading to more accurate emotion classification. In addition, dominance dimension allows for better differentiation between emotions that have similar valence and arousal levels but differ in terms of power dynamics. Furthermore, the dominant component measures the extent of assertiveness or submission, resulting in a more contextually aware recognition system that comprehends the interpersonal dynamics at work in addition to emotional states. To sum up, 3D model with dominance accommodates richer representations, such as feeling proud (positive, moderate arousal, high dominance) or nervous excitement (positive, high arousal, fluctuating dominance), which is harder to classify in a 2D dimensional space.

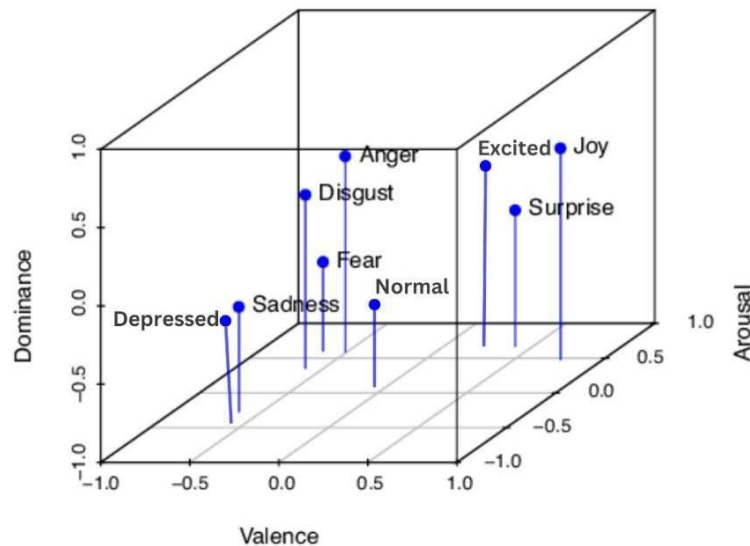


Figure 2.2: Three-dimensional space emotion representation

2.3.1. Challenges and Limitations in Dimensional SER

Dimensional SER faces several challenges and limitations that impact its performance and applicability. To begin with emotions are subjective experiences that can vary among individuals [12]. Different people may perceive and express emotions differently, leading to inter- and intra-speaker variability. This variability poses challenges in developing generalized models for dimensional SER that can accurately capture the diversity of emotional expressions. Second, annotating speech data with continuous emotional dimensions, such as valence, arousal, and dominance are challenging and often subjective. The process of assigning precise dimensional labels to speech recordings can be prone to errors and inconsistencies. In addition, developing accurate and robust dimensional SER models requires large-scale datasets with precise annotations of emotional dimensions. However, obtaining such datasets is challenging due to the time-consuming and resource-intensive nature of manual annotation. Moreover, there is no universally agreed-upon set of emotional dimensions for dimensional SER [20]. Different researchers and studies use varying dimensional models, such as valence-arousal, valence-tension, or valence-arousal-dominance. This lack of agreement makes it challenging to compare and generalize results across different studies and limits the establishment of standardized benchmarks.

2.4. Challenges and Considerations in Amharic SER

Amharic SER faces several challenges. To begin with, the availability of labeled speech datasets is limited, posing a challenge for training accurate and robust emotion recognition models in Amharic. In addition, emotions can be expressed differently across culture of Gondar, Gojjam, Wollo. and Shewa. In addition, annotating dimensionally for emotional Amharic speech data is complex. Moreover, Amharic has its own distinct acoustic characteristics, including phonetic and prosodic features. Understanding and capturing these specific acoustic patterns associated with different emotions in Amharic speech is essential for building accurate emotion recognition systems.

2.4.1. Dataset Availability and Annotation Challenges

To conduct research on dimensional SER, a significant number of datasets are needed to train the model effectively. However, publicly available datasets specifically for the

Amharic language are scarce on platforms like Kaggle, GitHub, or other similar sources. The only publicly shared dataset for Amharic language is provided by [11]. Annotating the available dataset based on dimensions such as valence, arousal, and dominance poses significant challenges. Emotions are subjective in nature and difficult to obtain accurate annotated emotions.

2.5 Speech processing

Speech preprocessing is a process of preparing the raw speech signals for effective analysis. The preprocessing typically consists of filtering the audio signal, segmenting speech from silence, and extracting relevant features that represent emotional states. Speech preprocessing follows the following basic steps [11].

1. Load audio waveform files: The speech preprocessing begins with loading audio files from a dataset, which contains the raw audio data.
2. Down Sampling: The audio is down sampled to a frequency of 16 kHz. This reduces the amount of data size by lowering the sampling rate, which helps in speeding up processing while retaining essential features of the audio.
3. Channel Conversion: The audio is converted to a mono channel, meaning that it is transformed from two channels to a single channel. This simplifies the data and sufficient for speech analysis.
4. Silent Segment Removal: Silent segments are identified and removed. This helps in focusing on the actual speech content, improving the efficiency of subsequent processing.
5. Fixed-Length Resizing: The audio segments are resized to a fixed length, which standardizes the input data for further analysis. This is important for deep learning models that require consistent input sizes.

2.6 Feature extraction technique

Feature extraction is the process of transforming raw input data into a set of meaningful features or representations that can effectively enhance the performance of deep learning models [4]. A key challenge in SER is the extraction of efficient speech characteristics that characterize the emotional content of speech [21]. According to Shilandari, Marvi, and Hadjiabdolhamid [22], the number of features is also important in improving accuracy. As

the number of features grows, so does the recognition rate until it reaches a peak, indicating the ideal number of features, and then begins to decline. According to the researchers [12], there are four types of local and global features employed in SER systems: prosodic features, spectral features, voice quality features, and Teager energy operator-based features. Prosodic and spectral properties are more typically employed in SER systems. Traditional feature extraction approaches, such as prosodic, are inefficient at identifying each emotion as a trait shared by several emotions. Prosodic features describe how we speak rather than what we say. It detects the loudness, pitch, intensity, and rhythm of speech, but not the emotional features [23]. Fairbanks and Hoaglin [24] discovered that happy and angry people use the same utterance. Failures in SER work is caused by feature extraction [25]. Traditional parameters, such as prosodic elements, have limits due to commonalities across different emotions [5]. They emphasize the adaptability of deep learning in determining relevant features for emotion recognition. The authors use modern feature extraction approaches like spectrum extraction such as MFCC, which turn audio data into feature vectors that deep learning models can handle. They emphasize the importance of feature extraction in creating a concise representation of input audio signals and then using these feature vectors to train deep learning models for emotion categorization. The authors [4] analyze several feature extraction approaches and conclude that MFCC performs better at collecting voice characteristics.

The authors [26] study the use of FSVM to extract speech emotion features, concentrating on short-time energy and resonance peaks. Using the Emo-DB corpus, the authors assess the performance of normal SVM and FSVM and discover that the FSVM approach has a better recognition rate than the SVM method. They emphasize the importance of short-term energy and resonance peaks in distinguishing various emotional states in speech signals. Short-term energy is connected with amplitude shifts and emotional tone, whereas resonance peaks are associated with sound discrimination and the capacity to distinguish sounds in loud situations. Finally, in this proposed system, we employ advanced feature extraction algorithms such as MFCC, ZCR, and RMS to better capture the emotional content of speech. Those three feature extraction algorithms are commonly used for speech related task such as SER [12].

2.6.1. Mel Frequency Cepstral Coefficients (MFCC)

The most commonly used speech feature is undoubtedly MFCC. MFCC is highly popular and robust due to their accurate estimation of speech parameters and the efficiency of their calculation model. They are widely recognized for their ability to capture essential characteristics of the speech signal, making them a standard choice in various speech processing applications [12]. The MFCC technique captures speech signals by transforming their short-term power spectrum into a linear cosine transform of the logarithmic power spectrum on a nonlinear Mel frequency scale. This method closely aligns the frequency scale with the human ear perception, making it highly effective for speech analysis and recognition tasks.

2.6.2. Zero Crossing Rate (ZCR)

ZCR is a key characteristic of a signal that indicates how often the signal crosses the zero axis, shifting among positive and negative. This feature is mainly valuable for detecting brief, loud sounds in a signal and for recognizing slight changes in its amplitude. ZCR is widely used in speech processing responsibilities such as speech synthesis, enhancement, and recognition, as it aids in determining whether human vocal is existed in an audio sample [12].

2.6.3. Root Mean Square (RMS)

RMS value is a useful technique for determining the average of values over time. For audio signals, the amplitude is squared, averaged over a specific period, and then average of square root is taken. This process yields a value that, when squared, is relational to the signal effective power [12]. It measures the reconstruction error between the original and reconstructed signals.

2.7 Deep learning Algorithm

Deep learning is subgroup of machine learning focus on using neural networks to model and resolve complex problems. These algorithms are designed to automatically learn and extract features from raw data by passing it through multiple layers of neurons, or deep layers. Each layer processes the data in increasingly abstract ways, allowing the model to

recognize intricate patterns and relationships that are difficult to capture for traditional algorithms [13] .

Deep learning is mainly effective for responsibilities involving huge amounts of unstructured data like speech recognition. They are trained using large datasets and powerful computational resources, adjusting the weights of the networks between neurons through a process called backpropagation [12]. This enables the model to improve its predictions or classifications over time by minimizing errors.

2.7.1. Long Short-Term Memory (LSTM)

LSTM networks are a specialized form of RNN architecture. Developed by Hochreiter and Schmidhuber in 1997, LSTM were created to overcome a major challenge faced by traditional RNN: the difficulty in capturing long-term dependencies, which arises from the disappearing and exploding gradient issues during backpropagation [27]. The initial issue, known as vanishing gradients, occurs when the gradients of the loss function concerning the network weights diminish to a very small value. As the gradients are propagated backward through the layers, they decrease in magnitude, which hinders the model ability to learn. This problem is especially severe in deep networks where the output and input are separated by many layers. When activation functions like sigmoid are used, their derivatives are limited to values between 1 and 0. During backpropagation, gradients are multiplied by these small derivatives at each layer. As the layers increase, this repeated multiplication leads to a rapid decrease in gradient size. Eventually, the gradients become so small and vanish by the time they reach the earlier layers [28]. The second problem, exploding gradients happen when gradients increase rapidly as they are passed back through the layers. This issue is the reverse of the disappearing gradient problem. It can be just as harmful to the training process. This is because, when weights are set to large values at the start, it can cause significant problems during training. This issue becomes more pronounced if the activation function produces large derivatives. As gradients flow back through the layers, they can increase rapidly, leading to an explosion in their size. This results in unstable updates, which can make the training process ineffective and difficult to control [29]. Managing weight initialization and selecting appropriate activation functions are important steps in preventing this problem.

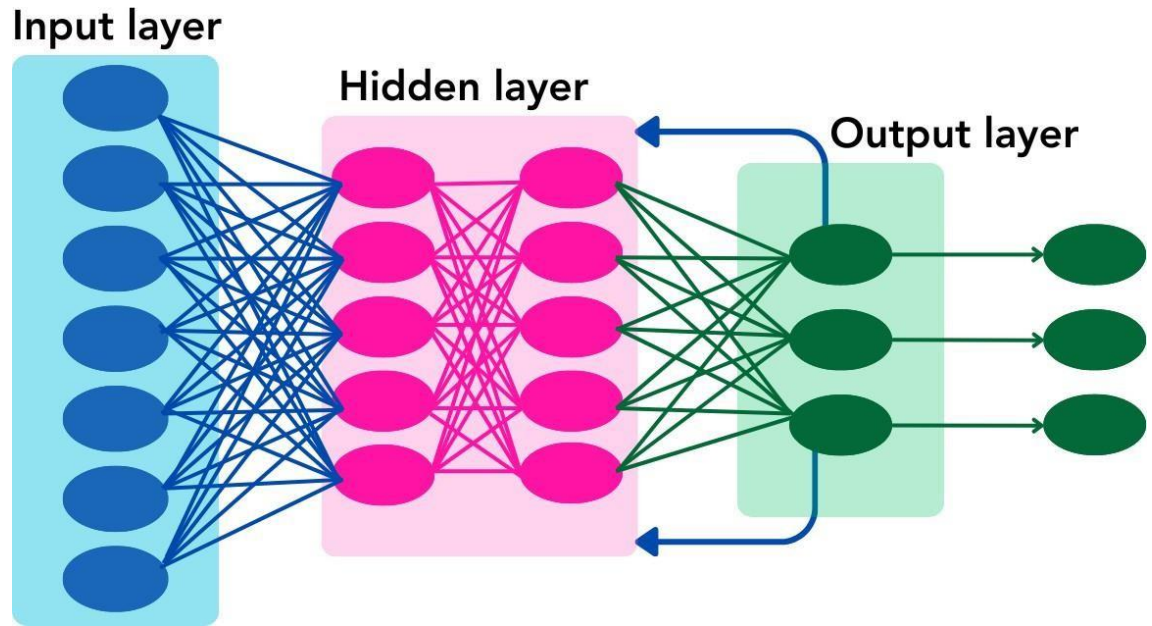


Figure 2.3: LSTM network layer

2.7.1.1 Key Features of LSTM

LSTM is truly exceptional due to their unique memory cell structure. It has gates that control the flow of information into and out of this memory cell. This design allows LSTM to keep and update information over long sequences. As a result, it is ideal for tasks that need to understand context over extended periods [30]. The core of an LSTM unit is its memory cell, which retains information over time. The cell state is the main highway through which information flows, and it undergoes modifications through regulated gates [31]. LSTM networks have three primary gates that control the flow of information. Those gates are forget gate, input gate, and output gate. The forget gate determines which information from the cell state should be discarded or retained. It uses a sigmoid layer that produces values between 0 and 1, where 0 means completely forget and 1 means completely retain [32]. The second gate known as input gate is responsible for decide which input values should be updated in the cell state. It features a sigmoid layer to control the input and works with a candidate layer that generates new potential values for the cell state [33]. Output gate dictates the output at the current time step, based on the cell state. The output gate uses a sigmoid function to decide what part of the cell state should be output, and this is typically combined with a tan function to push the output to a desired range [32].

2.7.2. Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM networks are an extension of traditional LSTM networks designed to capture information from both backward and forward sequences. While standard LSTM process sequences in a single direction, BiLSTM involve two LSTM running in parallel: one processes input from the first to the end of a sequence, while the other runs in reverse from the end to the beginning. The outputs from both directions are combined, allowing the network to have a more comprehensive understanding of context. This dual-directional approach is important in tasks like speech recognition where understanding both succeeding and preceding information is necessary [34]. The architecture of BiLSTM network with activation layer is represent in below Figure 2.4. Input sequences ($X_0, X_1, X_2, X_3, \dots, X_n$) are fed into two LSTM layers that operate in parallel. The forward LSTM layer processes the sequence from left to right, capturing information from the past to the present. Simultaneously, the backward LSTM layer processes the sequence from right to left, gathering context from the future to the past. The outputs from both directions are then combined, providing a more comprehensive understanding of the input sequence. This combined output ($y_0, y_1, y_2, y_3, \dots, y_n$) is passed through an activation function (f), which refines the output into the final.

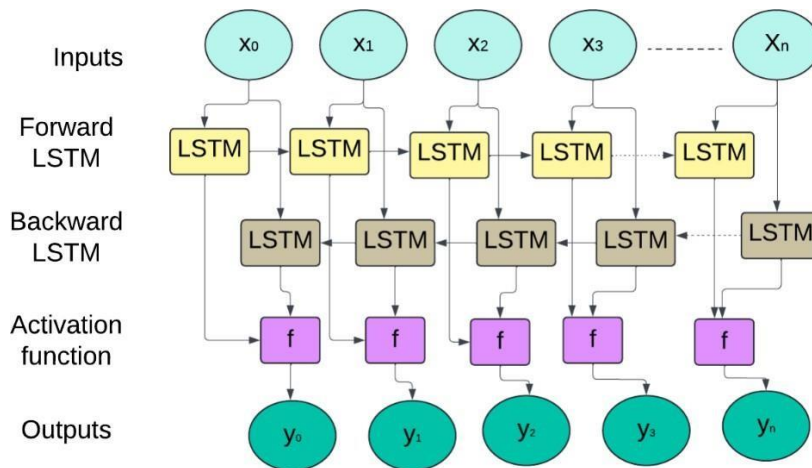


Figure 2.4: BiLSTM architecture with activation function [34]

2.8 Evaluation Metrics and Methodologies in Speech Emotion Recognition

SER involves the activity of automatically identifying and categorizing emotions conveyed through speech signals. Assessing performance of SER systems is important to assess model effectiveness and compare different approaches. This process requires the selection of appropriate evaluation metrics and methodologies. These metrics and methodologies determine how well the system achieves accuracy, robustness, and generalization to unseen data. In SER, two main categories of evaluation are commonly used: discrete emotion recognition and dimensional emotion recognition. Each category focuses on different aspects of emotion recognition and utilizes specific evaluation techniques. Understanding and utilizing suitable evaluation metrics and methodologies is important for advancing the field of SER and enhancing the effectiveness of emotion recognition systems.

2.8.1. Discrete Emotion Recognition

Discrete emotion recognition classifies speech signals into a set of predefined emotion categories. Commonly used evaluation metrics for discrete emotion recognition include accuracy, loss, and confusion matrix. Accuracy measures the overall correct classification rate, while loss measures model loss rate during classification. Confusion matrix provides to measure precision, F1 score, and recall. Precision and recall provide insights into the system ability to correctly identify specific emotions. F1 score is combined metric that considers both precision and recall. To measure the effectiveness of discrete emotion recognition systems, researchers often employ cross-validation techniques, such as k-fold cross-validation, to assess the generalization ability of the models on different subsets of the data. The evaluation process involves training the models on a portion of the dataset and testing them on the remaining unseen data, repeating this process multiple times to obtain reliable performance estimates [35].

2.8.2. Dimensional Emotion Recognition

Dimensional emotion recognition focuses on capturing the continuous dimensions underlying emotions, such as valence, arousal, and dominance. Evaluation metrics for dimensional emotion recognition include MSE, MAE, RMSE, and CCC. The first three are used to quantify the difference between predicted and true dimensional values, and the last CCC evaluates the overall agreement between the true and predicted dimensional ratings,

considering both the mean difference and the variance difference. To evaluate dimensional emotion recognition systems, researchers often use a separate validation set or employ cross-validation techniques, similar to those used in discrete emotion recognition, to assess the system performance in capturing the continuous nature of emotions and its generalization ability [12].

2.9 Related Work

SER is a useful research area as it can be applicable for many purposes [23]. Chen H. et al [4], [36] conduct SER to solve text emotion recognition problems. According to these scholars, the similar text has different emotions based on human vocal and implication. Their study compares various classification models for SER, providing a comprehensive analysis and use of deep learning model LSTM shows higher efficiency and performance compared to traditional ML models. However, the selected dataset is not large enough in terms of category (only four emotions such as happy, angry, sad, and neutral) and data size, which affects the reliability of the final model. Furthermore, while LSTM is identified as the most suitable model, there is no detailed analysis or explanation as to why it outperformed other models.

Dimitrova-Grekow et al [37] conducted research on seven emotions based on voice frequency. These authors used only the fundamental frequency parameter to classify the speech to emotions. They focused on checking either the vocal tone is sufficient information to recognize emotions. Their model performance was 89.74% accuracy for two emotions, 76.14% accuracy for three emotions, and 62.99% accuracy for four emotions. According to their experiment, number of emotions and accuracy of model performance have inversely proportional. However, happy and angry have the same utterance [23]. A study on speech emotion identification utilizing multiple classification models based on MFCC feature values is discussed by [4]. Emo-DB dataset was used in this study, which comprises audio and label files of emotional speech. The study compared the performance of six classifiers for predictive classification and found that LSTM had the highest accuracy among the six classifiers.

According to the researchers Schuller B. et al [38], combining linguistic and acoustic information for speech-emotion identification is more powerful than utilizing merely acoustic information. The authors offer a novel technique for acoustic feature analysis that combines a belief network-based emotional phrase-spotting algorithm with SVM. The study employs two emotional speech corpora, one for training and development and the other for testing, and produces an 8.0% improvement in mistake rates when compared to utilizing merely acoustic information. The study, however, only looks at a narrow set of emotions (anger, pleasure, disgust, fear, sadness, and surprise) and may not apply to other emotions.

Even if several researchers tried to recognize emotions at categorical labels, the dimensional approach for SER recently has gained attention as it can capture more nuanced emotions [14]. The categorical approach has the drawback of emotion recognition such as horror speech may contain both disgust and fear emotions and ambiguous class of emotion as emotion is subjective. i.e., Happiness can be classified as joy, pleasure, or excitement. To overcome this challenge Giannakopoulos et al [39] conduct a dimensional SER from movies. They have proposed a novel method for extracting emotional information from movies based on speech data. Those authors used a two-dimensional (valence and arousal) representation known as the emotion wheel and found better accuracy. However, their valence prediction is lower than arousal prediction. Dominance is a new dimension added to reflect the degree of control, influence, or power that a person perceives in a particular emotional state (submissive or dominant) [35]. These authors conduct research based on three-dimensional (dominance, arousal, and valence). They proposed two systems that use bimodal features such as text and acoustic to identify both dimensional and discreet emotions. A sequential system performs dimensional recognition first and then classify whereas a parallel system performs both dimensional and discrete emotions simultaneously. Their finding suggests that sequential system outperforms parallel systems and that dimensional emotion recognition is more accurate than discrete emotion recognition.

Even though dimensional emotion identification is superior to categorical emotion for recognizing and conveying emotion, valence prediction is more difficult than arousal and

dominance [12]. Arousal and dominance received more than 0.4 Concordance Correlation Coefficient (CCC) values, but valence received less than 0.1 CCC score, according to the researchers [40]. To address this issue, the authors [13] suggest a method for enhancing the low score of valence prediction using linguistic information. Their method combines acoustic and linguistic features, which yields a translation from words to vectors. The results improved valence prediction performance on both valence prediction and three dimensions.

Ben Letaifa L. [12] offers a voice emotion identification system based on deep learning methodologies and two efficient data augmentation techniques such as noise addition and spectrogram shifting. Three datasets namely TESS, Emo-DB, and RAVDESS were used to assess the system performance. Several feature extraction methods, including MFCC, ZCR, Mel spectrograms, RMS, and chroma, were used to pick the most relevant vocal features that describe speech emotions. The proposed system employs three distinct deep learning models such as MLP, CNN, and a hybrid model that blends CNN and Bi-LSTM. The authors found that the hybrid model had a greater accuracy rate. Atmaja et al [13] conducted a study on enhancing valence prediction in dimensional speech emotion recognition by incorporating linguistic information. Their findings indicate that utilizing IEMOCAP with an LSTM model resulted in over 80% accuracy, effectively doubling the performance compared to the existing state-of-the-art methods. They recommend further advancements to improve performance, aiming for levels closer to human annotation accuracy.

Since the majority of SER researches are on widely spoken languages like English, German, French etc, some scholars conduct research on under-resourced language. To distinguish emotions in Tamil speech, the researchers offer deep learning approaches such as LSTM and BiLSTM. The authors compare the performance of four distinct architectures with dropout layers, including Deep Hierarchical LSTM and BiLSTM, Deep Hierarchical BiLSTM and LSTM, Deep Hierarchical BiLSTM and BiLSTM, and Deep Hierarchical LSTM and LSTM, using a reduced Tamil emotional speech dataset. The results suggest that DHLB performs best, with an accuracy rate of 84% and the least amount of loss in each of the seven main emotions [2].

Ephrem et al [11] collected a new Amharic SER dataset (ASED) publicly available for the first time. Their dataset covers four dialects (Shewa, Wollo, Gondar, and Gojam) for five emotions namely happy, angry, neutral, fear, and sad from 27 unique sentences by 65 speakers (25 males and 40 females). Those datasets have a total of 2474 records evaluated with eight judges. Those authors had performed SER with algorithms VGG, Alex-Net, and ResNet50 using two feature extraction techniques such as Mel-spectrogram and MFCC. They conclude VGG performs best accuracy (90.73%) with relatively fastest training time and Alex-Net 91.13%. However, the size of the dataset is small to conduct a robust model. Furthermore, emotion recognition is not limited to five classes. In this research, we conduct a dimensional Amharic SER system. We use the bimodal feature acoustic and linguistic to get more accurate results on three dimensions [12] [13].

2.10. Summary literature review

We summarize the related work by specifying the name of the authors with published year, the title of the research, database, Methodology, the result out of 100%, and recommendation.

Table 2.1: literature review summary

Authors (year)	Title	Database	Methodology	Result (100%)	Recommendation and Gap
Ephrem A. et al (2023)	cross-corpus multilingual SER: Amharic vs. other languages	ASED RAVDESS Emo-DB URDU	Alex-Net VGG ResNet50	Average accuracy = 61.93, F1-score = 53.30	The study focused only on four languages, which may not be representative of all languages. The study did not explore other machine-learning models and low

					accuracy (61.93%)
Atmaja et al. (2020)	Improving Valence Prediction in Dimensional SER Using Linguistic Information	IEMOCAP	LSTM	Improve more than 80%, double the accuracy of the state-of-the-art	They recommend to improve performance to close to human annotation.
Chalapathi et al (2022)	Ensemble Learning by High- Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal	RAVDESS	EL-HDAF HAF SBL-DF	Accuracy for those three algorithm EL- HDAF, HAF, and SBL-DF are 94.84%, 89.41%, and 88.06% respectively.	Recommend making further comparative study with another existing algorithm
Peng Z. et al (2023)	Enhancing Dimensional Emotion Recognition from Speech through Modulation- Filtered Cochleagram and Parallel Attention	RECOLA SEWA	LSTM, PA-Net	LSTM accuracy is 88% and 59% for arousal and valence respectively whereas PA-Net accuracy is 93% and 63% for arousal and valence respectively.	They recommend integrating into pre-trained models and exploring transfer learning techniques to enhance model performance.

	Recurrent Network				
Fernandes et al (2021)	SER Using Deep Learning LSTM for Tamil Language	1450 records	DHLL DHLB DHBL DHBB	Accuracy DHLL = 81% DHLB = 84% DHBL = 83% DHBB = 81%	Focus on categorical emotions, and they use small datasets
Chen et al. (2023)	SER using multiple classification models based on MFCC feature values	Emo-DB	SVM, KNN, Semi-supervised graph-based classifier, ECOC classification model, Naïve Bayes, LSTM	Among six classifiers LSTM performs best with 93.2% and 73.03% accuracy in training and testing models respectively.	The model is over fitted because the training test is higher than the testing test for all six models. Eg. The training accuracy for SVM and ECOC is 100% but 68.88% and 64.44% test accuracy respectively
Hui Ma (2022)	SER Based on Fuzzy Support Vector Machine	Emo-DB	SVM FSVM	Accuracy of SVM and FSVM for four emotions (happy, angry, sad, and calm) is 80%,79%,77%, 70% and 88%, 83%, 82%, and	Recommend to work with complex emotion speech recognition. Only use SVM and recommend to improve by investigating

				80% respectively.	another algorithm. Work only for four categorical emotions but not dimensional emotions.
Barhoumi et al. (2023)	Real-Time SER Using Deep Learning and Data Augmentation	TESS Emo-DB RAVDESS	MLP CNN CNN + BiLSTM	For those three databases (TESS Emo-DB RAVDESS), the accuracy is MLP (99.90%,98.76%, 90.09%), CNN (99.95%,98.51%, 86.85%), and CNN + BiLSTM (100%, 99.50%, 90.12%) respectively.	They recommend investigating the use of physiological signals such as heart rate and skin conductance and doing it with other language rather than English.

CHAPTER THREE: Methodology (Materials and Methods)

3.1 Overview

This study explores the implementation of dimensional Amharic SER through the use of LSTM and BiLSTM models. The chapter provides an in-depth discussion on the key steps involved, including the process of preparing the required speech data, applying various preprocessing techniques, extracting relevant features, and developing the models for emotion recognition. Since our research involves comparing categorical and dimensional approaches to SER, we also develop and evaluate a categorical SER model. By incorporating this model into our study, we ensure a comprehensive analysis that allows us to assess the strengths and weaknesses of the categorical method in direct comparison to the dimensional approach. This evaluation provides deeper insights into the specific contexts in which each method might be more effective, contributing to a more understanding of SER overall.

3.2 Dataset preparation

In this study, we utilize the ASED dataset, which is publicly accessible [11]. This dataset represents the first freely available Amharic speech emotion corpus and was created using recordings from 65 individuals, consisting of 40 females and 25 males. The dataset contains a total of 2,474 speech samples. The emotional content of these samples was rigorously evaluated by a panel of eight judges. The availability of this dataset offers valuable resources for advancing research in Amharic SER. These datasets encompass five distinct emotion categories: happiness, sadness, anger, neutrality, and fear. The specific distribution and details of these emotion classes are presented clearly in Table 3.1 below. This table provides overview of the dataset structure, including the number of samples allocated to each emotion class and dialects such as Gojjam, Wollo, Shewa, and Gondar.

Table 3.1: Amharic SER dataset

No	Emotion	Amharic dialect				Number of records
		Gojjam	Wollo	Shewa	Gondar	
1	Neutral	104	208	140	70	522
2	Fearful	59	170	200	81	510
3	Happy	68	150	188	80	486
4	Sad	70	135	135	130	470
5	Angry	111	120	140	115	486
Total		412	783	803	476	2474

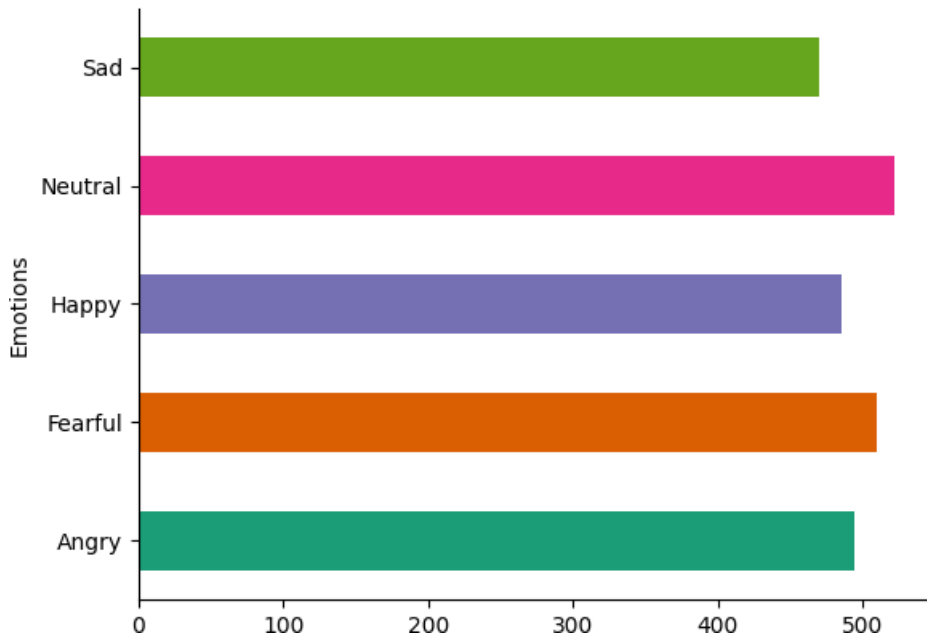


Figure 3.1: ASER dataset visualization

3.3 Dataset Annotation

For the categorical SER task, we did not perform any additional annotation since the dataset already contains high-quality, well-validated labels. These annotations were carefully assigned and rigorously evaluated by a panel of eight judges, ensuring their reliability and consistency. Given the thorough review process, the existing annotations were deemed sufficient for our analysis, allowing us to proceed directly with model development and

evaluation without requiring further refinement. However, for the task of recognizing emotions in a continuous format, the dataset required additional annotation along three key dimensions: arousal, valence, and dominance. These dimensions are essential for capturing the nuanced emotional states represented in a dimensional approach to emotion recognition. Unlike categorical models that classify emotions into distinct categories like happy or sad, continuous models evaluate emotions requires precise labelling in terms of arousal, valence, and dominance. Therefore, the annotation team undertook the task of annotating the dataset along these dimensions to enable a more detailed emotional expression.

For the annotation team, we delivered a clear lecture explaining the three emotional dimensions: arousal, valence, and dominance to ensure a thorough understanding of the concepts. Additionally, we provided consistent and detailed training sessions, along with clear instructions to guide the annotation process effectively. To further support the annotators, we designed a user-friendly annotation tool with several key features. The tool allows annotators to easily assign scores for each dimension, automatically navigate to the specific speech segments awaiting annotation, seamlessly review any missing or incomplete annotations, and finally export the annotated data as CSV file. These combined efforts streamline the annotation process, enhance accuracy, and ensure that the workflow remains efficient and intuitive for all team members.

For this annotation purpose, we clearly define voice characteristics and guideline. Voice characteristics such as intensity, pitch, timber, duration, and loudness [35].

Intensity: It is proportional to the amplitude of the sound wave, which reflects the infraglottic pressure. This pressure is influenced by the spacing between the verbal cords when they are open. As the infraglottic pressure increases, the intensity of the sound also rises. Based on this relationship, high intensity is often linked to stress, while low intensity may be associated with feelings of depression or fatigue.

Pitch: Determined by the number of glottic cycles occurring within a given time frame. This feature is influenced by the properties of the vocal folds and the size of the pharynx, which directly impact the fundamental frequency of the voice.

Voice Quality (Timbre): It enables the differentiation between two sounds that have the same pitch and intensity. It is influenced by the structure of the resonant organs, which is why each individual has a unique voice.

Duration: Is influenced by the amount and speed of air being expelled. Greater lung capacity and a larger rib cage result in a longer duration.

Loudness: Is a perceptual characteristic of voice that corresponds to the amplitude of the sound wave. It represents how strong or weak a voice is perceived by the listener.

Based on these voice characteristics, the emotion could be classified as follows.

Happiness: Elevated pitch and intensity accompanied by an increase in the speed of speech [41].

Fear: A higher pitch with minimal variation, reduced energy, and a quicker speech rate that includes more frequent pauses [41].

Sadness: Sad emotions are characterized by a higher pitch, reduced intensity, but increased vocal energy (around 2000 Hz), extended duration with frequent pauses, and a lower first formant (the initial sound produced) [41].

Anger: Anger is expressed through a lower pitch, greater intensity, and increased energy (around 500 Hz) throughout the vocalization. It is also associated with a higher first formant and quicker onset times at the beginning of speech [41].

3.3.1. Discrete To Dimensional Model Relationship

For the annotation of our dataset, we used the work of Martín Iglesias as a foundational benchmark [35]. In his research, Iglesias effectively describes the relationship of seven distinct emotions: happiness, surprise, fear, disgust, sadness, anger, and neutral along with three dimensions arousal, valence, and dominance. While we did not implement Iglesias's work directly, we established it as a guiding standard for our annotation team, ensuring consistency and reliability in how emotions were classified in our dataset. On the following

Table 3.2, we illustrate the correlation between categorical emotion classifications and their dimensional counterparts.

Table 3.2: Categorical and dimensional emotion relationship

Emotions	Valence	Arousal	Dominance
Neutral	0.0	0.0	0.0
Angry	-0.43	0.67	0.34
Happy	0.76	0.48	0.35
Surprise	0.4	0.67	-0.13
Disgust	-0.6	0.35	0.11
Fear	-0.64	0.6	-0.43
Sadness	-0.63	-0.27	-0.33

3.4 System Architecture

In this section we present the model architecture of this study. It includes the categorical Amharic SER, dimensional Amharic SER with only acoustic feature and bimodal feature.

3.4.1 Categorical Amharic SER Model Design

Categorical Amharic SER task begins with loading the Amharic speech emotion dataset (ASED), which serves as the primary data source. The next step after preprocessing involves data augmentation using techniques such as pitch shifting, noise addition, and time shifting. Then feature extraction, focusing on three acoustic features such as MFCC, ZCR, and RMS.

Once the features are extracted, the dataset is split into 80% training data, 10% for validation, and and 10% testing data. The training data is fed into the core model, which consists of an acoustic feature input, followed by three LSTM layers and a final dense layer for further processing. The acoustic feature input represents sound characteristics that

capture important elements of speech data. These features are processed through three LSTM layers. The LSTM layers analyze the temporal patterns and contextual information within the speech sequences, effectively capturing how emotions unfold over time. After extracting and learning from these sequential features, the output is passed to a dense layer. The dense layer performs the final classification by applying learned weights to make predictions about the emotion category of the input data. It combines the information captured by the LSTM layers and maps it to distinct emotion classes such as happy, neutral, sad, fearful, and angry. Through this layered architecture, the model achieves accurate emotion recognition by using both temporal dynamics and feature interactions within the speech data. This model, labelled ASER (Amharic SER), is responsible for classifying the speech data into distinct emotion categories.

The system outputs one of five possible emotions: Happy, Sad, Angry, Fearful, or Neutral. The testing data is separately fed into the ASER model to evaluate its performance and validate its accuracy in emotion classification as shown on Figure 3.2.

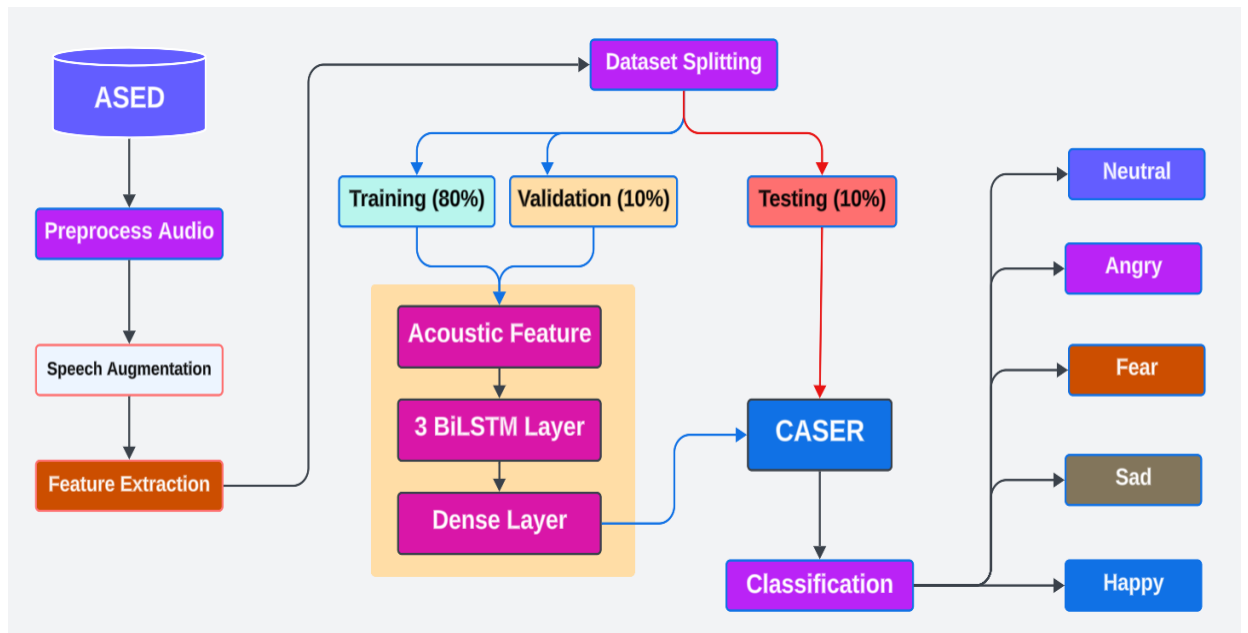


Figure 3.2: Categorical Amharic speech emotion recognition model design

3.4.1.1 Speech Preprocessing

Prior to feature extraction, several pre-processing steps were performed on the datasets. The audio recordings were first downsampled to 16kHz and converted to mono format. The majority of sound clips in the datasets are 5 seconds or less in duration, though a few

exceed this length. To standardize clip length, shorter clips were extended to 5 seconds by appending silence at the end, while longer clips were trimmed to precisely 5 seconds [11]. We follow the same speech preprocessing strategy like Ephrem et al [11].

3.4.1.2 Data Augmentation

Data augmentation is a widely used technique that enhances model robustness, supports better generalization, and helps prevent overfitting in SER (SER) models. This method not only improve the overall accuracy but also enhances data distribution consistency, resulting in reduced variance [42]. To enhance the effectiveness of our proposed methods, we employ three data augmentation techniques, which are: noise addition, time shifting, and pitch shifting.

Noise Addition

The noise addition technique is commonly used in audio processing and speech recognition to improve the robustness of models. This simulates real-world scenarios where audio signals may be affected by background noise, helping the model become more resilient to such disturbances [12]. To achieve this, we calculate the noise amplitude using the formula $\text{noise_amp} = 0.035 * \text{np.random.uniform()} * \text{np.amax(data)}$. Here, $\text{np.random.uniform}()$ generates a random value between 0 and 1, which is multiplied by 0.035 (recommended) to set the noise level relative to the signal strength.

Algorithm 1 Data Augmentation: Noise Addition

Ensure: Data \leftarrow ndarray, audio time series

1: Noise-amp $\leftarrow 0.035 \times \text{numpy.random.uniform}() \times \text{numpy.amax}(\text{Data})$

2: Noise $\leftarrow \text{Noise-amp} \times \text{numpy.random.normal}(\text{size}=\text{Data.shape}[0])$

3: Augmented-data $\leftarrow \text{Data} + \text{Noise}$

4: return Augmented-data

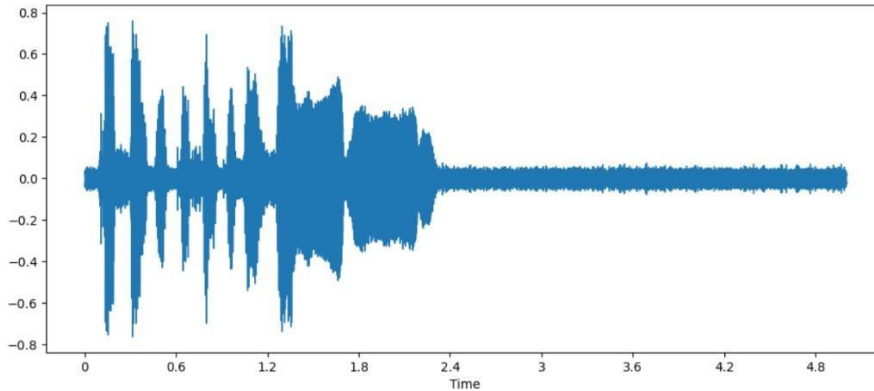


Figure 3.3: Data augmentation using white noise addition technique

Time Shifting

Time shifting is an audio processing technique that involves adjusting the timing of an audio signal without altering its duration or pitch. In this method the audio is moved forward or backward by a specific sample number. The `shift(data)` function is designed to introduce random variations in the starting point of the audio clip. To do this, we first generate a random value between -5 and 5 seconds using `np.random.uniform()`. This range is selected to enable substantial shifts both backwards and forwards in time. We then convert this value to milliseconds by multiplying it by 1000, ensuring accuracy in the time adjustments.

Algorithm 3: Data Augmentation: Time Shifting

Ensure: Data \leftarrow ndarray, audio time series

1: Shift-range \leftarrow `int(numpy.random.uniform(low=-5, high=5) \times 1000)`

2: Augmented-data \leftarrow `numpy.roll(Data, Shift-range)`

3: return Augmented-data

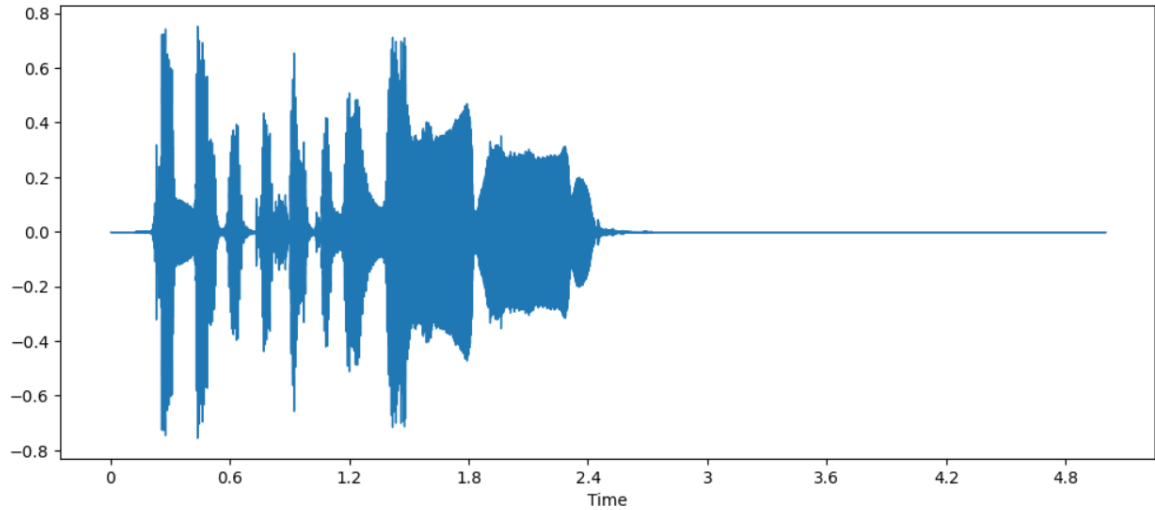


Figure 3.4: Data augmentation using time shifting technique

Pitch Shifting

In SER, the timing of speech can fluctuate based on the emotion conveyed. For example, a person articulates their words more rapidly or slowly when experiencing anger or sadness compared to when they feel neutral or happy. By adjusting the pitch along the time axis, we can replicate these variations and generate new training samples that differ slightly from the original data. This creates variations of the same audio sample, allowing the model to learn from diverse tonal characteristics. The `pitch(data, sampling_rate, pitch_factor=0.7)` function is utilized to modify the pitch of the audio signal, enabling our model to adapt to tonal variations. The implementation uses the librosa library `effects.pitch_shift()` method, which expertly adjusts the pitch of the audio while maintaining its original speed. In this case, the `pitch_factor` is set to 0.7, which indicates a downward shift in pitch, represented in semitones. By adjusting the pitch, the model can better recognize emotions across different vocal tones. By applying these methods, the model achieves better generalization and performance when exposed to unseen audio data.

Algorithm 4 Data Augmentation: Pitch Shifting

Ensure: Data \leftarrow ndarray, audio time series

Input: Sampling-rate \leftarrow Integer (22050 Hz)

Input: Pitch-factor \leftarrow Shift factor (default 0.7)

```
1: Augmented-data ← librosa.effects.pitch_shift(Data, Sampling-rate, Pitch-factor)
2: return Augmented-data
```

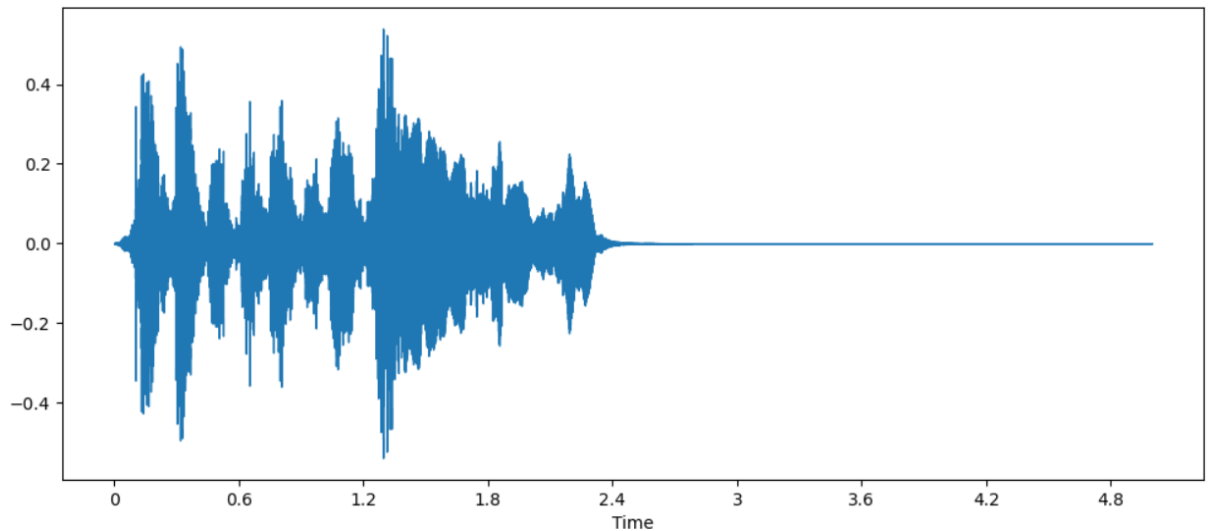


Figure 3.5: Data augmentation using pitch shifting technique

3.4.2 Dimensional Amharic Speech Emotion Recognition Model Design

For dimensional SER, we undertook the design of two distinct models: one that incorporates linguistic features and another that operates without it. This dual-model approach was strategically implemented to rigorously evaluate the impact of linguistic features on the accuracy and effectiveness of dimensional emotion prediction.

3.4.2.1 Dimensional Amharic SER model without Linguistic Feature

For dimensional SER without linguistic feature, we initiated the process by precisely preparing the VAD (Valence, Arousal, and Dominance) annotation data, which was developed by three trained psychologists. This step ensuring the accuracy and reliability of the emotional annotations, as these are important for effective recognition of emotions in speech. Once the annotation data was refined, we proceeded to load both the audio speech data and the corresponding parallel annotation files. This involved a systematic approach to ensure that each segment of speech data accurately aligned with its associated emotional attributes. Subsequently, we mapped the annotation values to each speech path. We follow the same speech preprocessing, feature extraction, and dataset splitting technique as our categorical SER model. Then after, the model training phase is begun. The model takes

acoustic feature as input and pass through three consecutive LSTM layer. The output of last LSTM layer feed to dense layer as shown on the following Figure 3.6.

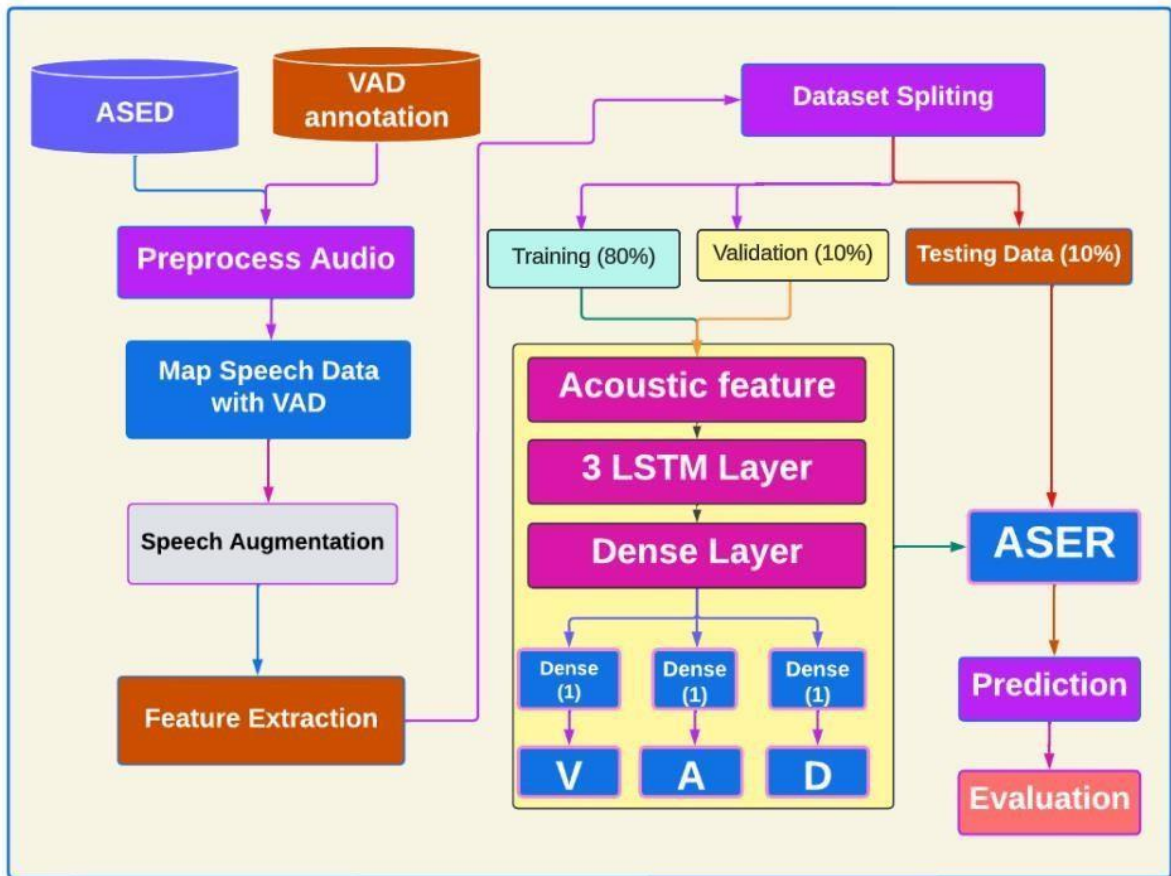


Figure 3.6: Dimensional SER model design without linguistic feature

3.4.2.2 Dimensional Amharic SER model with Linguistic Feature

The model we designed for dimensional SER that incorporates linguistic features bears similarities to the model that relies solely on acoustic features. The primary distinction lies in the addition of linguistic features, which enrich the model by providing essential context and meaning that complement the purely acoustic characteristics of the speech. To ensure the quality and accuracy of our data, we undertook a thorough process of manual transcription for each segment of augmented speech data. This transcription step was crucial, as it ensured that the underlying linguistic elements corresponding to the emotional content were accurately documented and aligned with the audio data. After the transcription process is completed, we proceed with text preprocessing and feature

extraction as essential steps in preparing the data for effective dimensional SER as shown on the following Figure 3.7.

Through this careful preparation and augmentation of data, we facilitate robust training of the model, enabling it to assimilate and generate insights from both acoustic and linguistic context.

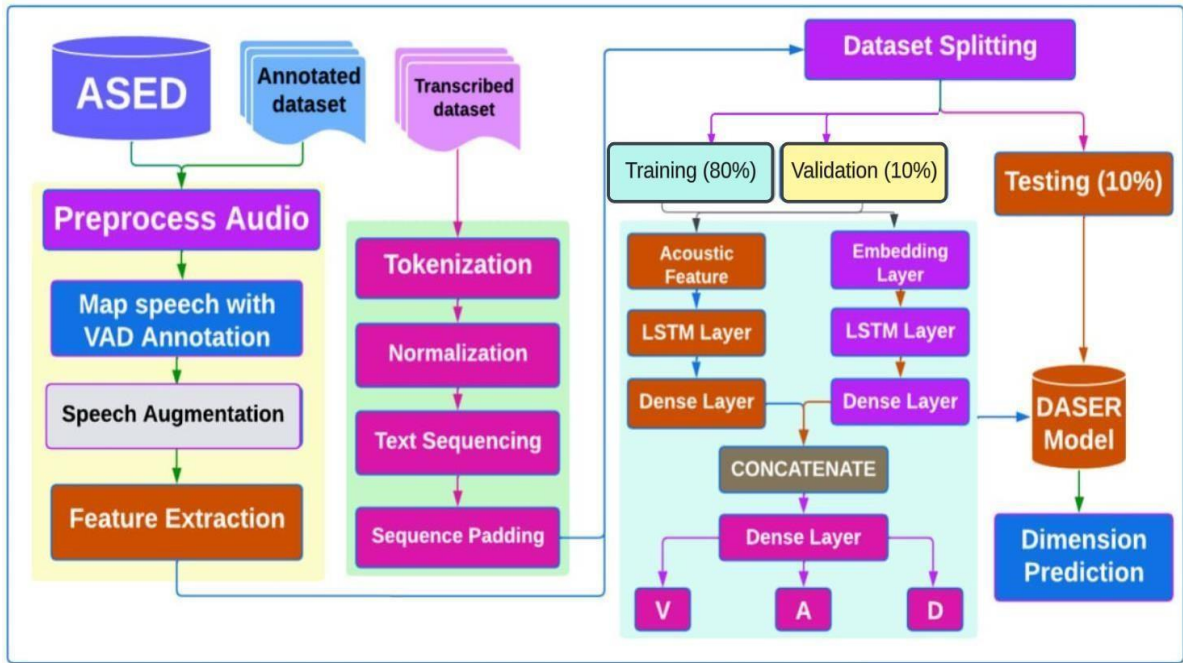


Figure 3.7: Dimensional SER model with linguistic feature

We developed a dimensional Amharic SER model that uses bidirectional LSTM (BiLSTM) networks. In our design, we implemented two versions of the model: one incorporating linguistic features and the other operating without linguistic feature. The architecture of both models closely follows the framework established in our LSTM model work, with the primary distinction being the substitution of the traditional LSTM algorithm with the BiLSTM variant. This enhancement allows for the processing of input sequences in both forward and backward directions, by this means we improving the context understanding of the data and potentially leading to better emotion recognition outcomes.

3.5 Text preprocessing

This section explains the process of converting text into a collection of features or attributes that encapsulate the information from the sequences of words intended for system input.

All text must undergo this transformation, as it is categorized as either part of the training set or the testing set. The initial step involves collecting all the text associated with each audio utterance, referred to as transcriptions. For transcription, we preferred for manual transcription to ensure greater accuracy.

3.5.1 Tokenization and Normalization

Once transcription is complete, text preprocessing begins with tokenization, which involves dividing the text into smaller units known as tokens. Normalization, which follows the tokenization process, is designed to standardize characters with similar meanings into a consistent form. For this step take it in to account during our manual transcriptions. We normalize as the following: alphabet family of “ሐ” and “ኀ” normalized to “ሀ”, alphabet family of “ሠ” normalized to “ሰ”, alphabet family of “ጸ” normalized to “ፀ”, and alphabet family of “ዐ” normalized to “አ”.

3.5.2 Text Sequencing

Text sequencing is the method of converting unprocessed text into a numerical format suitable for deep learning algorithms. In this process, each word is assigned a distinct integer value while maintaining the original order of the words. This transformation results in a numeric sequence represented as a vector, matrix, or tensor, which can be utilized for various natural language processing tasks. For this purpose, we employed the ‘texts to sequences’ function from the Keras Tokenizer class, which changes a collection of sentences into a series of sequences by replacing each word with its associated numerical token.

3.5.3 Sequence Padding

Sequence padding is the process of adding the necessary number of tokens to a text sequence to ensure all sequences are equal in length. This is important for processing text data in batches, as deep learning models require inputs with fixed dimensions. Padding guarantees that shorter sequences are extended to match the length of the longest sequence in the dataset by incorporating padding tokens, such as zeros or other specified values. For example, in our dataset the sentence “ዞር በል ጥፋልኝ ከአሁን በኋላ እንዳላይህ” have six tokens and another sentence “ነገ ስብሰባ አለኝ” have three tokens. Thus, these sequences need to be

padding to ensure that the shorter sentences match the length of the longest one in the dataset. Padding is done by adding zeros either at the beginning or the end of the sentence. In our scenario, we chose to append zeros to the end of the sentences.

3.5.4 Word Embedding

Deep learning models cannot directly process textual representations or strings because they lack the ability to perform mathematical operations on them. As a result, text must be converted into numerical vectors to allow the models to analyze and learn from the information it contains. The next step after preprocessing is vectorization, which involves generating a numerical representation of the corpus (collection of transcriptions) that maintains the semantic relationships between words. For vectorization process, embedding is more preferable technique and overcome the limitation of other vectorizer such as bag of words (BoW) and TF-IDF. Embeddings are a way to represent data as vectors in a lower-dimensional space [35]. For the development of the model, we incorporated an embedding layer into the LSTM and BiLSTM architecture. This layer is constructed using the Keras Embedding class and is configured with the following parameters for our specific case:

Input dimension: The vocabulary size or the highest integer index that the layer can encode is determined by the number of vocabularies in the tokenizer vocabulary, plus one for the out-of-vocabulary token.

Output dimension: The dimensionality of the dense embedding space determines the size of the output vectors.

Input length: The length of the input sequences refers to the uniform length that all sequences is padded to match, which corresponds to the maximum sequence length.

The embedding layer convert the input word indices into dense vectors within the embedding space. These dense vectors learned during training and capture semantic relationships between words.

3.6 Model Development

We developed LSTM and BiLSTM neural network model for dimensional SER with and without linguistic features. Both LSTM and BiLSTM are specifically designed to hold long-term dependances in sequential data, making them suitable for processing emotional speech and sentences in natural language text.

3.6.1 Categorical Amharic SER Model Development

The architecture of the categorical SER model is designed to effectively classify emotions from acoustic features extracted from speech signals. It begins by input layer that receives the acoustic features, which are derived from preprocessing techniques such as MFCC, ZCR, and RMS. These features capture the valuable attributes of the audio signal, offering a comprehensive representation for the layers that follow.

The core of the model consists of three stacked BiLSTM layers. The initial LSTM layer consists of 256 units, enabling the model to grasp intricate temporal dependencies within the data. This is followed by a second LSTM layer with 128 units, which continues to refine the learned representations. The final LSTM layer, which contains 64 units, is designed to prepare the model for classification. The use of LSTM is particularly advantageous in this context, as they are adept at handling sequential data and can remember information over long time intervals, which is important for understanding the nuances of speech.

Following the LSTM layers, the model moves to dense layers. The first dense layer features 64 units, and it is succeeded by a second dense layer with 32 units. These layers function to integrate the features obtained from the LSTM, allowing the model to learn more abstract representations. The final dense layer, with 5 units, corresponds to the different emotion classes: Happy, Angry, Sad, Fear, and Neutral. The following Figure 3.8 clearly shows the architecture of our categorical SER task.

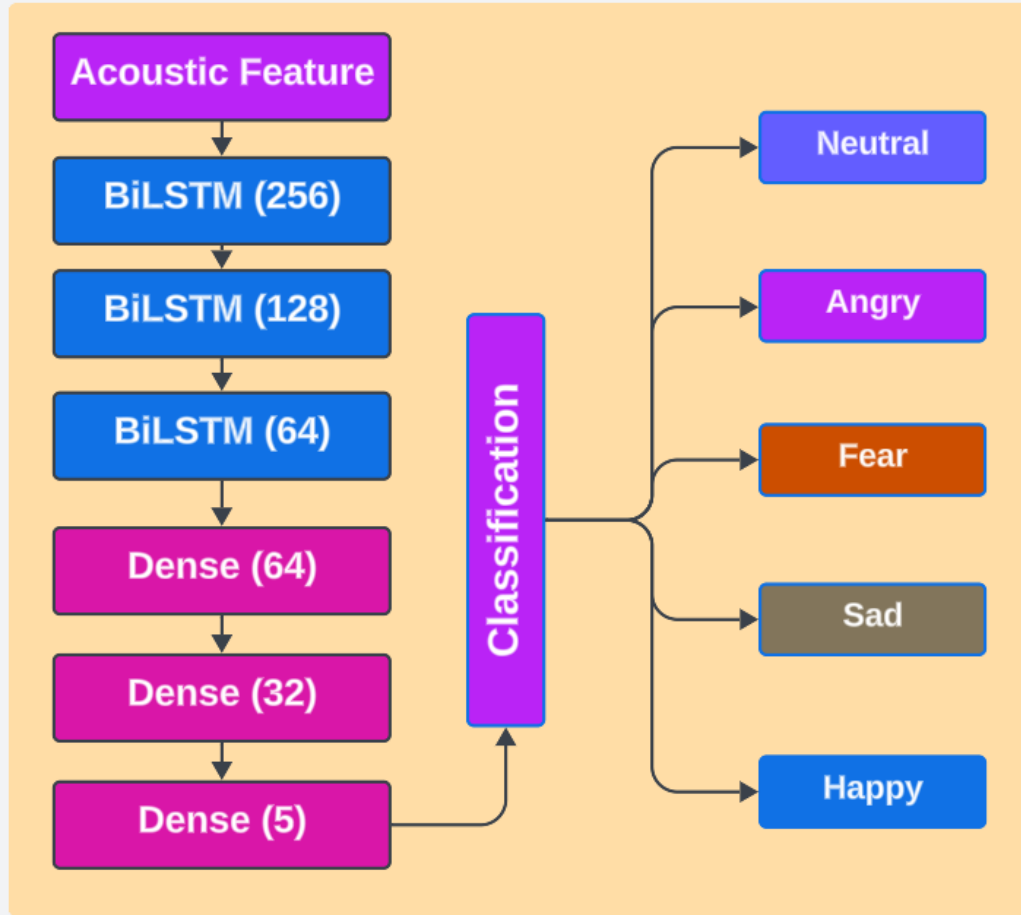


Figure 3.8: Architecture of categorical SER model to classify five emotions

3.6.2 Dimensional Amharic SER Model Development with Acoustic Feature

The architecture of the dimensional SER model is designed to assess emotions based on three key dimensions: Valence (V), Arousal (A), and Dominance (D). This model processes acoustic features extracted from speech signals to provide a nuanced understanding of emotional states. The model begins by taking acoustic feature as input and pass through three consecutive stacked LSTM layers like categorical SER model design.

After the LSTM layers, the model branches into three separate dense layers, each dedicated to predicting one of the emotional dimensions. The first dense layer outputs a single value for Valence (V), the second dense layer also outputs a single value for Arousal (A), and the third dense layer provides a single value for Dominance (D). The following Figure 3.9 shows the clear architecture of dimensional SER with acoustic feature.

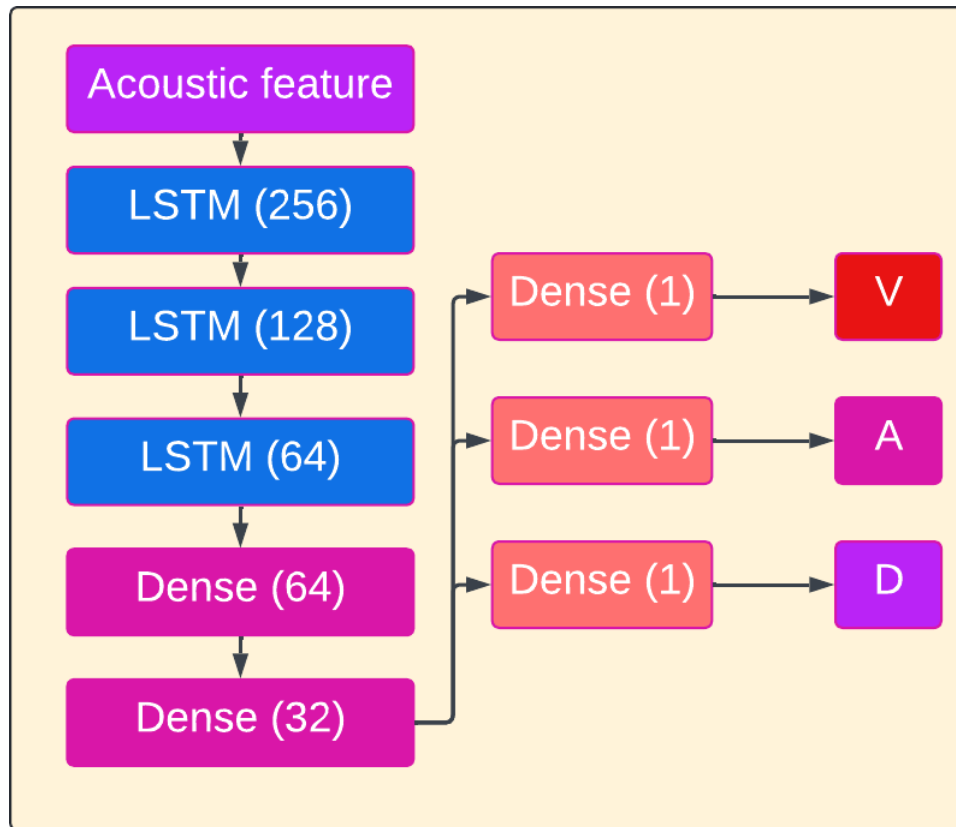


Figure 3.9: Architecture of dimensional SER using acoustic feature

3.6.3 Dimensional Amharic SER Model Development with Linguistic and Acoustic Feature

The architecture of the dimensional SER model integrates both linguistic and acoustic features to provide a comprehensive analysis of emotional states, represented by the dimensions of Valence (V), Arousal (A), and Dominance (D). This dual-network approach allows the model to use the strengths of both types of features, and enhancing its ability to capture the nuances of human emotion in speech.

The model begins with two parallel pathways: one for acoustic features and another for linguistic features. The acoustic feature pathway processes the audio data through a series of LSTM layers. The acoustic network procedure is similar to dimensional SER using only acoustic feature. Simultaneously, the linguistic feature pathway processes textual data through a similar structure of LSTM layers. This pathway also begins with a 256-unit LSTM layer, followed by a 128-unit layer and a final 64-unit layer, concluding with a dense

layer of 64 units. This parallel processing allows the model to extract meaningful features from both the acoustic and linguistic inputs.

After the individual pathways, the outputs from both the acoustic and linguistic networks are concatenated. This concatenation merges the learned representations, enabling the model to utilize the complementary information from both feature types. Following this, a dense layer of 64 units processes the combined features, leading to a second dense layer with 32 units that further refines the representation.

Finally, the model branches into three separate dense layers, each dedicated to predicting one of the emotional dimensions. The initial dense layer outputs a single value for Valence (V), the second dense layer outputs a single value for Arousal (A), and the third dense layer provides a single value for Dominance (D).

This architecture effectively combines acoustic and linguistic features, allowing for a richer and more nuanced understanding of emotional states in speech. By utilizing the strengths of LSTM networks and integrating diverse feature types, the model offers a robust framework for dimensional SER. The architecture is shown on the following Figure 3.10.

We created a dimensional Amharic SER model architecture that utilizes only acoustic features, as well as one that combines both acoustic and linguistic features, using bidirectional LSTM (BiLSTM) networks. The design of both models follows thoroughly to the framework we established for our LSTM model, with the main difference being that the LSTM algorithm has been replaced with the BiLSTM algorithm. This improvement enables the processing of input sequences in both backward and forward directions, which we enhance the context understanding of the data, potentially resulting in more accurate emotion recognition results.

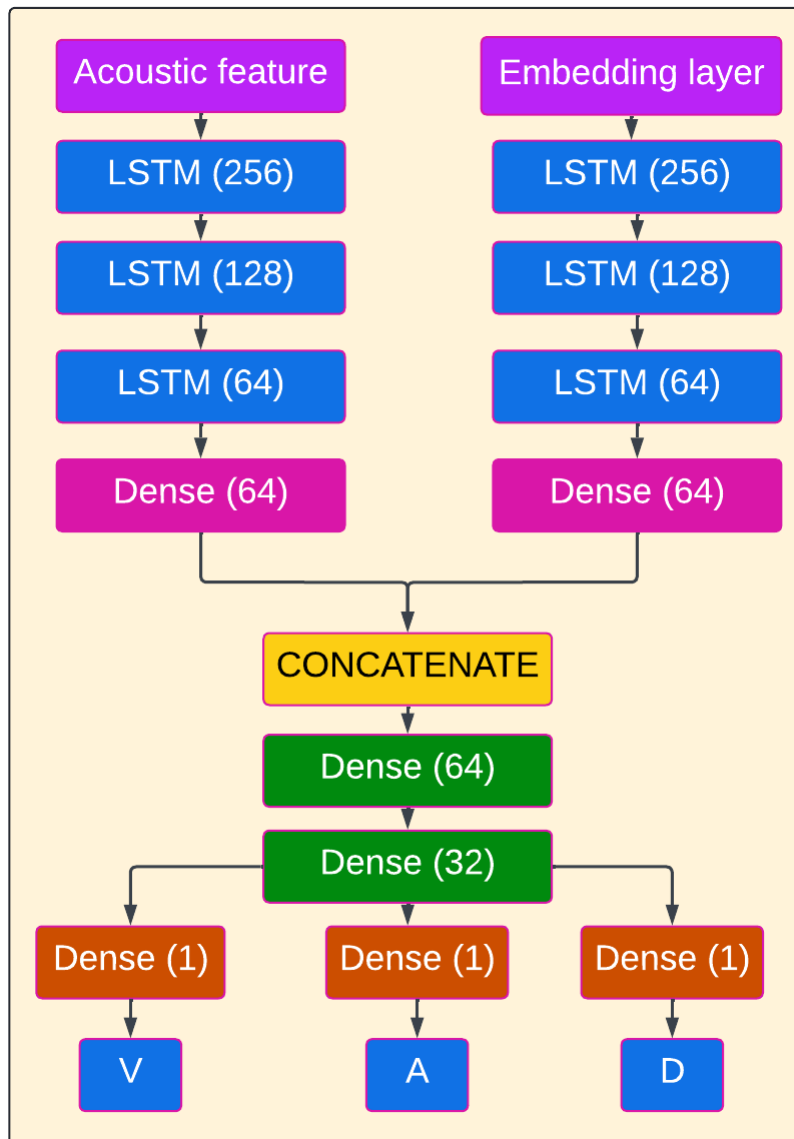


Figure 3.10: Architecture of dimensional SER using both acoustic and linguistic feature

3.7 Development Tools

In our work on developing both categorical and dimensional Amharic SER models, we use a variety of Python libraries and tools within the Google Colab platform, which supports Jupyter notebook. Below is a summary of these packages and their roles in our research.

Pandas and NumPy: These foundational libraries used for data manipulation and numerical operations. Pandas provides powerful data structures, such as DataFrames, which facilitate the handling of structured data important for preprocessing the datasets

used in SER tasks. NumPy offers support for arrays and matrices, along with a collection of mathematical functions to apply to these data structures. It is useful for handling numerical data in the modelling process.

Librosa: This specialized library is essential for audio analysis, allowing us to extract various features from audio files that are critical for understanding and identifying emotions conveyed in speech.

Seaborn and Matplotlib: Both tools are used for data visualization, helping to create informative plots and graphs that enhance the understanding of data patterns and facilitate the interpretation of our model results.

Scikit-learn: This comprehensive machine learning library provides utilities for data preprocessing and model evaluation, including standard scaling of features and splitting datasets into training and testing subsets. It also offers functions to generate confusion matrices and classification reports, which are important for assessing model performance.

IPython Audio and Display: These components enable the playback of audio files directly within the notebook environment, allowing for interactive exploration of the sound data used in our models.

Keras and TensorFlow: These powerful frameworks for building and training deep learning models are important to our implementation. We employ LSTM and BiLSTM, to create robust models for emotion recognition. Keras simplifies the model-building process while TensorFlow serves as the backend engine, providing high performance and scalability.

PyAudioAnalysis: This library is useful for in audio feature extraction and classification, providing various functionalities for segmenting and analyzing audio signals. It enables us to easily implement advanced audio analysis techniques, enhancing the feature extraction process for our SER models.

NLTK: is a Python library designed for developing NLP applications. It offers a variety of tools for processing and analyzing text data.

Python: is a versatile programming language extensively used for developing deep learning models, for natural language processing (NLP), and speech recognition applications. It offers a comprehensive array of libraries and tools that simplify data preprocessing, model creation, training, and assessment.

Lucidchart: is an online diagramming tool utilized for designing flowcharts, organizational charts, wireframes, and various other visual representations of data. It allows real-time collaboration and is commonly used in business, education, and project planning. We used Lucidchart for drawing a diagram.

Microsoft office, Google docs: utilized for documenting the study.

3.8 Evaluation Metrics

The models undergo testing across different experimental configurations during the development, training, and evaluation stages. We use different evaluation metrics for categorical and dimensional SER task.

3.8.1 Evaluation Metrics for Categorical Amharic SER model

We use accuracy, loss, and confusion matrix as a performance indicator that are used to evaluate classification of emotional speech.

Accuracy: indicates the model effectiveness in accurately identifying whether the emotional speech is correctly classified or not. It is measured by dividing the number of correct predictions by the total number of predictions.

Loss: is an indicator of how effectively a model can reduce its prediction error during training. It quantifies the gap between the predicted values and the actual values in the training dataset. A lower loss signifies better model performance.

Confusion matrix: is a chart that displays the number of true negative, false negative, false positive, and true positive generated by the model predictions [12]. It is important for calculating metrics such as precision, recall, and F1 score. These metrics enable us to assess the model performance concerning its true positive rate (recall), true negative rate (precision), overall accuracy, and overall effectiveness (F1 score). The algorithm that

performs higher performance is chosen based on these evaluation results. Mathematically these metrics computed as follows.

$$Accuracy = \frac{TP+TN}{TN+TP+FN+FP} \dots\dots\dots (1)$$

$$Precision = \frac{TP}{TP+FP} \dots\dots\dots (2)$$

$$Recall = \frac{TP}{TP+FN} \dots\dots\dots (3)$$

$$F1score = 2 * \frac{Precision*Recall}{Precision+Recall} \dots\dots\dots (4)$$

Where:

TP (True Positives): correctly predicted as positive.

TN (True Negatives): correctly predicted as negative.

FPP (False Positives): incorrectly predicted as positive.

FNN (False Negatives): incorrectly predicted as negative.

3.11.2. Evaluation Metrics for Dimensional Amharic SER model

To evaluate the effectiveness of our dimensional SER model, we used three key metrics: MAE, R² score, and MSE. Each of these metrics provides insights into the model accuracy and performance in making predictions.

MSE: This metric quantifies the average of the squares of the errors, which are the differences between predicted and actual values. By squaring the errors, MSE emphasizes larger differences more than smaller ones, making it particularly sensitive to outliers. A lower MSE indicates better model performance, as it signifies that the predictions are closer to the actual data points [43].

$$MSE = 1/n \sum_{i=1}^n (xi - yi)^2 \dots\dots\dots (5)$$

Where:

“n” is the sample of observations

“xi” is the actual value

“yi” is the predicted value

MAE: This metric measures the average of the absolute differences between predicted and actual values. Unlike MSE, which squares the errors, MAE treats all deviations equally by taking their absolute values. This makes MAE less sensitive to outliers, offering a straightforward interpretation of average prediction error. A lower MAE indicates a more accurate model [44].

$$MAE = 1/n \sum_{i=1}^n |x_i - y_i| \dots \dots \dots (5)$$

Where:

“n” is the sample of observations

“xi” is the actual value

“yi” is the predicted value

CCC: is an important metrics in evaluating dimensional SER systems, as it assesses the degree of agreement between predicted and actual values. The CCC ranges from -1 to +1, with a value of +1 indicating perfect agreement, 0 indicating no agreement, and -1 indicating perfect disagreement. It combines aspects of precision and bias, enabling investigators to understand not only how closely predictions correspond to real values but also any systematic deviations that exist. The mathematical formulation of the CCC takes into account the covariance of the variables and their variances, making it robust against linear transformations, which is mainly useful in contexts where different scaling employed. This makes the CCC an invaluable tool for emotion recognition systems that strive to accurately capture the nuances of human emotional expressions in continuous multidimensional space[45].

$$CCC = \frac{2.Cov(X,Y)}{Var(X) + Var(Y) + (Mean(X)-Mean(Y))^2} \dots \dots \dots (6)$$

Where:

“X” represents predicted values and “Y” represents actual values.

3.9 Summary

This chapter focused on the research methods employed to create a deep learning model for both categorical and dimensional analysis of SER in Amharic. To verify the reliability of our results, we implemented an experimental research design, which enabled us to determine the causal relationship between independent and dependent variables while controlling for different variables. We explained the dataset used for our study and the preprocessing and feature extraction technique for both acoustic and linguistic feature. The model architecture for categorical Amharic SER model using acoustic feature and dimensional Amharic SER model using merely acoustic feature and both acoustic and text feature are clearly explained. To recognize emotion in Amharic speech, we employ BiLSTM and LSTM deep learning algorithm. Furthermore, the evaluation metrics for the categorical Amharic SER model, including accuracy, loss, and confusion matrix, as well as CCC, MAE, and MSE for the dimensional Amharic SER model, are thoroughly detailed. The development tools to accomplish this study also described. In conclusion, this chapter give a deep understanding of the research methods used to design the proposed model.

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION

4.1 Overview

In this chapter, we present experimental results of our proposed model, providing a comprehensive analysis of its performance. We discover the outcomes of our experiments on categorical Amharic SER and dimensional Amharic SER with merely acoustic feature and both acoustic and linguistic feature. Each experiment is clearly explained, allowing for a thorough understanding of how these different feature sets impact the model ability to accurately recognize emotions in Amharic speech. Furthermore, the research question of our study is answered in this chapter.

4.2. Model implementation

In this study, we addressed both categorical and dimensional SER tasks. To recognize emotional dimensions related to valence, arousal, and dominance, we utilized acoustic features exclusively. Additionally, we analyzed the impact of incorporating both acoustic and linguistic features on model performance, particularly focusing on enhancing the valence dimension. To accomplish this we load audio files, annotated CSV file, and transcription. The process of loading speech data, mapping with the associated annotated dataset having target variable of valence, arousal, and dominance value is explained in chapter three. We clearly presented how the speech data is integrated with the annotated CSV file, which serves as a mapping between the audio recordings and their corresponding emotional attributes. Additionally, the steps involved in text and speech preprocessing, ensuring that the data is prepared for analysis, data augmentation and feature extraction technique are thoroughly explained.

4.2.1. Dataset Splitting

To maintain a fair evaluation, we separated the available data into testing, training, and validation sets. This separation enabled us to train and evaluating its performance on the unseen portion. We follow the standard dataset splitting technique 80%, 10%, and 10% for training, testing, and validation respectively[46].

Table 4.1: Dataset splitting for training, validation, and testing

Emotions	Training	Validation	Testing	Total
Neutral	418	52	52	522
Fear	408	51	51	510
Happy	388	49	49	486
Sad	376	47	47	470
Angry	388	49	49	486

4.2.2. Model Hyperparameter

Hyperparameters are important and govern the training process of deep-learning architectures. It improves the model performance, convergence speed, and ability to predict or classify to unseen data. In our study, we employed various hyperparameters across different models, including LSTM and BiLSTM architectures for dimensional Amharic SER and BiLSTM for categorical Amharic SER.

Epochs: It tells the frequency with which the learning algorithm processes the entire training dataset. In our case, we utilized early stopping to prevent overfitting, which monitors the validation accuracy and halts training if no improvement is observed for a specified number of epochs (patience). The ReduceLRonPlateau function modifies the learning rate when the validation accuracy levels off. Lowering the learning rate allows the model to make more precise adjustments to the weights, which result in improved convergence. In this research, learning rate reduced by half if there is no enhancement in validation accuracy over three epochs, with a minimum learning rate established at 0.00001.

Batch size: It refers to the quantity of training samples employed in a single cycle of model training. A batch size 32, 64, 128, and 256 was applied across all experiments and achieved better accuracy on batch size 64.

Sequence length and embedding dimension: The Sequence Length defines the length of input sequences fed into the model. Embedding dimension specifies vector space in which

words or features are represented. In our model which utilized linguistic feature an embedding dimension of 26 was employed, which effectively captures the semantic relationships between features.

Activation function: It adds non-linearity to the model, enabling it to recognize complex patterns. we used the softmax function for categorical models, which is suitable for multi-class classification, while linear activation was applied in the dimensional models suitable for regression tasks.

Optimizer: is tasked with updating the model weights according to the loss function. In this research, we utilized the Adam optimizer, known for its adaptive learning rate features, which makes it suitable for a variety of issues. Those hyper parameters are summarized on the following Table 4.2.

Table 4.2: Hyper parameter used in this study

No	Hyperparameter	Categorical Amharic SER	Dimensional Amharic SER			
			Acoustic feature		Acoustic and Linguistic feature	
		BiLSTM	LSTM	BiLSTM	LSTM	BiLSTM
1	Epoch	Early stop	early stop	early stop	early stop	early stop
2	Batch size	32-256	32-256	32-256	32-256	32-256
3	Vocabulary	--	--	--	107	107
4	Embedding Dimension	--	--	--	26	26
5	Activation function	softmax	linear	linear	linear	linear
6	Optimizer	Adam	Adam	Adam	Adam	Adam

4.3 Model Configuration

The model configuration or summary provides a detailed explanation of the architecture of the model, outlining the various layers and connections, as well as the sizes of the parameters involved. This includes information about the types of layers used (such as

LSTM, BiLSTM, and Dense layers), the number of units in each layer, activation functions, and regularization techniques applied.

4.3.1 Dimensional Amharic SER using acoustic feature with BiLSTM

The BiLSTM model designed for the dimensional Amharic SER task comprises several key layers that work together to process and classify emotional content from speech data. The model begins with an input layer that accepts sequences of length 1620, followed by a reshape layer that adds an additional dimension to the input data, preparing it for further processing. The core of the model consists of three Bidirectional LSTM layers, which capture contextual information by processing the input sequences in both backward and forward directions. The first BiLSTM layer outputs a sequence of 512 features, while the second layer refines this to 256 features, and the third layer reduces it to a more compact representation of 128 features. Following the BiLSTM layers, two dense layers further process the features, outputting 64 and 32 units, respectively. The model concludes with three output layers, each responsible for predicting different dimensions of emotion: dominance, arousal, and valence, with each outputting a single score. In total, the model contains 1,359,555 parameters, all of which are trainable, ensuring that it can effectively learn from the training data to identify and classify emotions in Amharic speech. The BiLSTM model configuration for dimensional Amharic SER using acoustic feature is shown on the following Figure 4.1.

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 1620)	0	-
reshape (Reshape)	(None, 1620, 1)	0	input_layer[0][0]
bidirectional (Bidirectional)	(None, 1620, 512)	528,384	reshape[0][0]
bidirectional_1 (Bidirectional)	(None, 1620, 256)	656,384	bidirectional[0][0]
bidirectional_2 (Bidirectional)	(None, 128)	164,352	bidirectional_1[0][0]
dense (Dense)	(None, 64)	8,256	bidirectional_2[0][0]
dense_1 (Dense)	(None, 32)	2,080	dense[0][0]
valence (Dense)	(None, 1)	33	dense_1[0][0]
arousal (Dense)	(None, 1)	33	dense_1[0][0]
dominance (Dense)	(None, 1)	33	dense_1[0][0]

Total params: 1,359,555 (5.19 MB)
 Trainable params: 1,359,555 (5.19 MB)
 Non-trainable params: 0 (0.00 B)

Figure 4.1: Dimensional Amharic SER using acoustic feature model configuration

4.3.2 Dimensional Amharic SER using acoustic and linguistic feature with BiLSTM

The model designed for Dimensional SER integrates both acoustic and linguistic features through a comprehensive architecture comprising multiple layers. The input layer accepts two types of data: acoustic features with a sequence length of 1620 and linguistic features with 26 dimensions. Embedding layer serves to transform the linguistic features into a dense representation, enhancing the model ability to capture semantic relationships. Following the embedding layer, a reshape layer prepares the acoustic data for processing. The model employs several Bidirectional LSTM layers, starting with a layer that outputs 512 features for each of the 1620-time steps, effectively capturing complex temporal patterns. This is followed by additional Bidirectional LSTM layers that further refine the feature representation, reducing the output to 256 and then 128 features. Dense layers subsequently process these representations, with the final dense layers outputting 64 and 32 units, respectively. The model culminates in three output layers, each dedicated to predicting a specific dimension of emotion: dominance, arousal, and valence. With a total of 2,779,169 parameters, all of which are trainable, the model is designed to learn intricate relationships within the data, enabling effective emotion recognition in Amharic speech. The BiLSTM model configuration for dimensional Amharic SER using acoustic and linguistic feature is shown on the following Figure 4.2.

4.4 Experimental Results

In this subsection, we present the experimental results obtained from our proposed model for Amharic SER. Our evaluation includes two distinct approaches: the categorical approach, which focuses on classifying emotions into specific categories, and the dimensional approach, which assesses emotions based on underlying dimensions such as arousal, valence, and dominance. By analyzing the effectiveness of this model through these two approaches, we provide a comprehensive understanding of its effectiveness in recognizing emotions within Amharic speech.

Layer (type)	Output Shape	Param #	Connected to
input_layer (InputLayer)	(None, 1620)	0	-
input_layer_1 (InputLayer)	(None, 26)	0	-
reshape (Reshape)	(None, 1620, 1)	0	input_layer[0][0]
embedding (Embedding)	(None, 26, 26)	2,782	input_layer_1[0][0]
bidirectional (Bidirectional)	(None, 1620, 512)	528,384	reshape[0][0]
bidirectional_3 (Bidirectional)	(None, 26, 512)	579,584	embedding[0][0]
bidirectional_1 (Bidirectional)	(None, 1620, 256)	656,384	bidirectional[0][0]
bidirectional_4 (Bidirectional)	(None, 26, 256)	656,384	bidirectional_3[0][0]
bidirectional_2 (Bidirectional)	(None, 128)	164,352	bidirectional_1[0][0]
bidirectional_5 (Bidirectional)	(None, 128)	164,352	bidirectional_4[0][0]
dense (Dense)	(None, 64)	8,256	bidirectional_2[0][0]
dense_1 (Dense)	(None, 64)	8,256	bidirectional_5[0][0]
concatenate (Concatenate)	(None, 128)	0	dense[0][0], dense_1[0][0]
dense_2 (Dense)	(None, 64)	8,256	concatenate[0][0]
dense_3 (Dense)	(None, 32)	2,080	dense_2[0][0]
valence (Dense)	(None, 1)	33	dense_3[0][0]
arousal (Dense)	(None, 1)	33	dense_3[0][0]
dominance (Dense)	(None, 1)	33	dense_3[0][0]

Total params: 2,779,169 (10.60 MB)
Trainable params: 2,779,169 (10.60 MB)
Non-trainable params: 0 (0.00 B)

Figure 4.2: Dimensional Amharic SER model configuration using both acoustic and linguistic feature with BiLSTM

4.4.1. Categorical Amharic SER model using acoustic feature

Our research explores the question: To what extent does dimensional SER varies from categorical SER? To address this inquiry, we developed a categorical SER model capable of accurately classifying emotion categories. This model is important for our investigation, as it allows us to analyze and compare the effectiveness of categorical classifications against the dimensional approach.

Loss and Accuracy: The loss metric is utilized to assess the effectiveness of the model. It quantifies the difference between predicted outputs and the actual labels in the training dataset. A lower loss value signifies that the model predictions are more aligned with the true labels. The following Figure 4.3 displays a mathematical plot showcasing the accuracy and validation accuracy of the BiLSTM-based model. This result indicates a high level of performance. The model achieved an accuracy of 99.35%, and 99.14% for training and

validation accuracy respectively. In the training and validation loss graph, the training loss consistently decreased over the epochs, indicating effective learning, while the validation loss also showed a downward trend, suggesting good generalization. The accuracy graph demonstrates that both training and testing accuracy reached near-perfect levels.

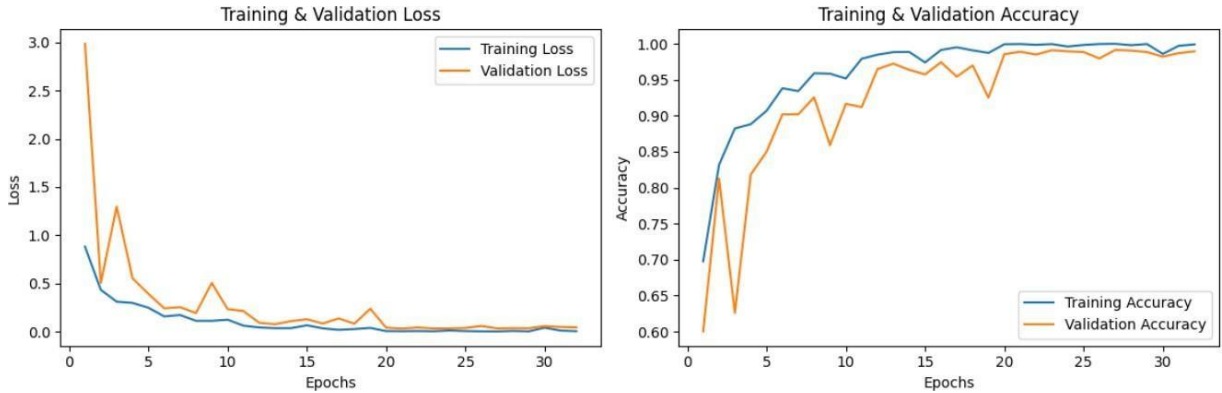


Figure 4.3: Training and validation accuracy of categorical model using feature transformation technique

Confusion matrix: this model has demonstrated exceptional performance as show on the following confusion matrix analysis.

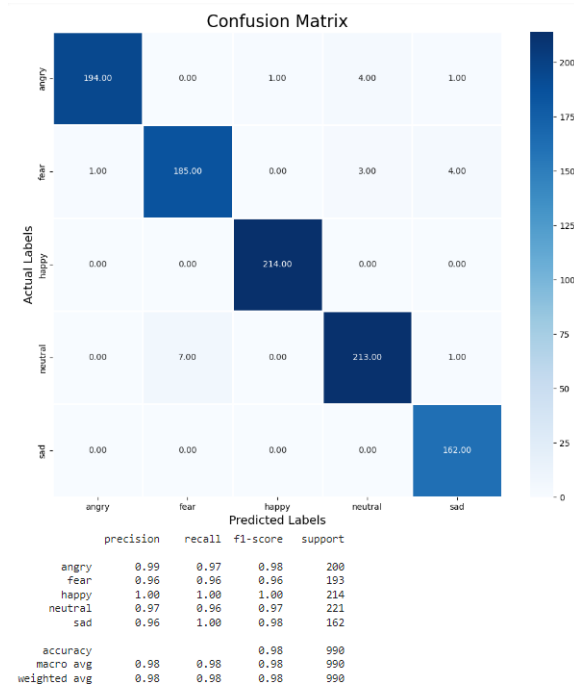


Figure 4.4: Confusion matrix analysis for categorical Amharic SER

The confusion matrix presents a thorough breakdown of the model performance for each emotion category: angry, fear, happy, neutral, and sad. Starting with the angry category, the model achieved 194 true positives (TP), which reflects its success in accurately identifying instances of anger. There was only one false positive (FP), where an instance was incorrectly categorized as angry but was actually fear, and six false negatives (FN), where actual instances of anger were misclassified one instance as happy, four instance as neutral, and one instance as sad. This resulted in impressive precision of 0.99 and a recall of 0.97, indicating that the model is highly reliable in detecting anger while also identifying a few instances incorrectly.

Turning to the fear category, the model identified 185 true positives but had 7 false positives, with instances misidentified as fear instead neutral. Additionally, eight instances of actual fear were misclassified as 1 angry, 3 as neutral, and 4 as sad, which contributed to a precision of 0.96 and a recall of 0.96. Although the performance is solid, the presence of these misclassifications suggests some overlap between the fear emotion and others. The happy emotion category showcased exceptional performance, with 214 true positives and 1 false positive which is misclassified as happy but actually it is angry and have no any false negative.

In the neutral category, the model demonstrated outstanding effectiveness with 213 true positives. However, it had 7 false positive, where an instance predicted as neutral were 3 of them actually fear and 4 of them were angry. There were also eight false negatives, indicating that the 7 actual neutral instances were misclassified as fear and on neutral instance classified as sad emotions.

Finally, for the sad category, the model recorded 162 true positives and achieved perfect precision, with no false negatives. There was 6 false positive where an actual 1 neutral, 4 fear, and 1 angry instance was misclassified as sad.

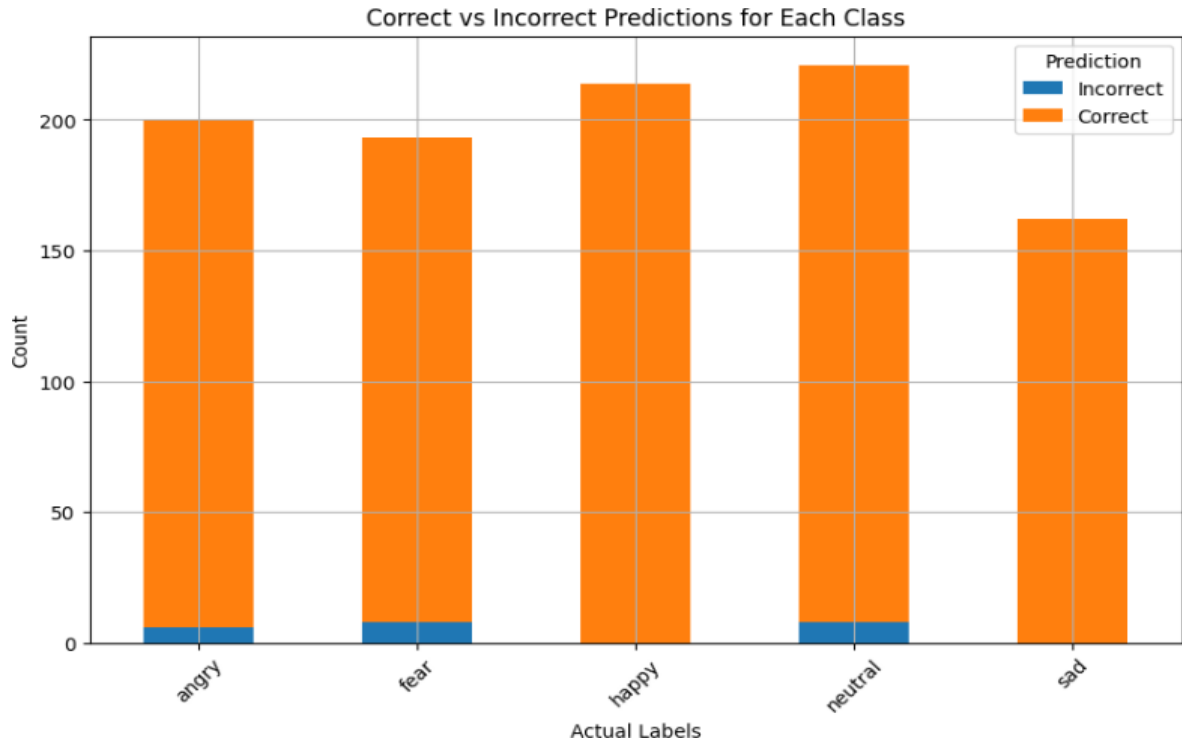


Figure 4.5: The model performance on test data correctly and incorrectly classified on each emotion class

4.4.2. Dimensional Amharic SER model using merely acoustic feature with LSTM and BiLSTM

In this subsection, we present the results of dimensional SER using only acoustic features. We begin by displaying the total loss across the three dimensions simultaneously, followed by the results for each individual dimension. The performance of the dimensional Amharic SER model, which utilizes acoustic features and an LSTM algorithm, is shown on the following graph Figure 4.6. The training loss shows a general decreasing trend throughout the training process, indicating effective learning from the training data. The validation loss also exhibits a decreasing trend, although with some variability. This fluctuation indicates that while the model is learning, it encounters some challenges in generalizing to unseen data. However, as the training progresses, the validation loss stabilizes at a lower value, which is a positive sign of the model ability to generalize effectively.

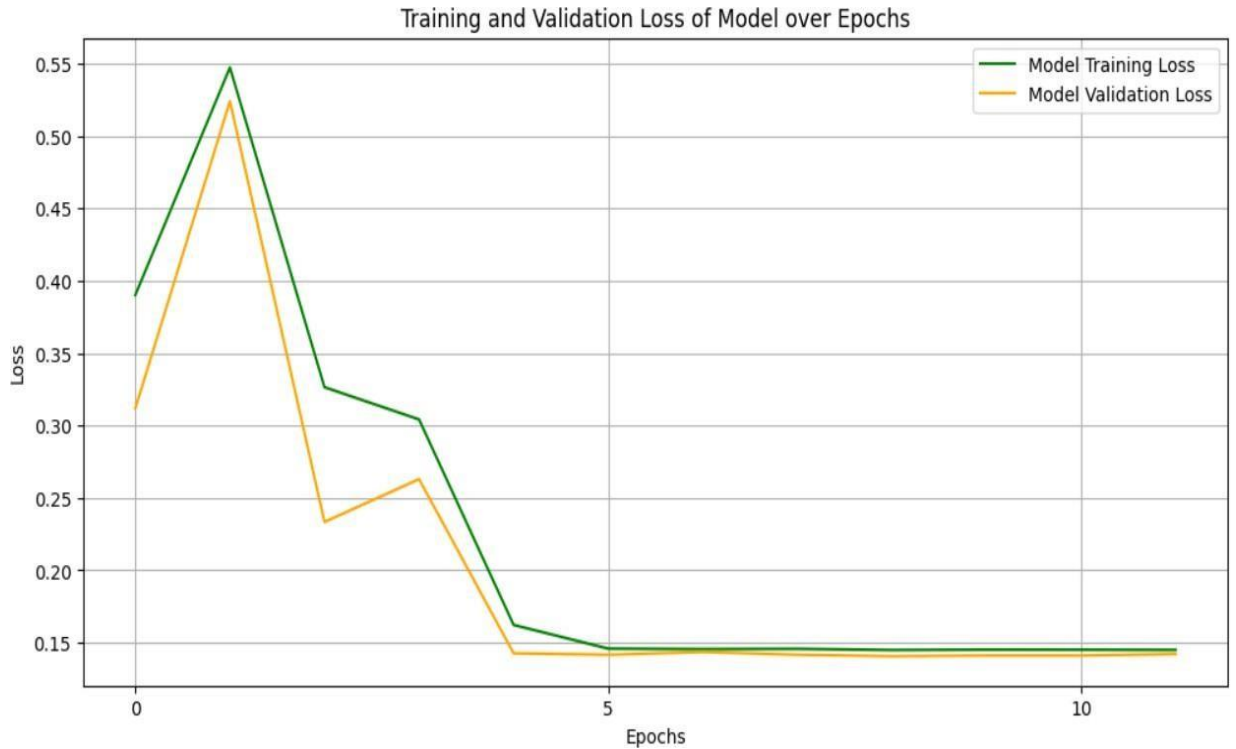


Figure 4.6: Training and testing loss of dimensional Amharic SER using acoustic feature and LSTM algorithm

Figure 4.6 shows the total loss of the model performance for all three dimensions valence, arousal, and dominance.

In terms of evaluation metrics, the model demonstrates strong performance in predicting valence, with a Mean Squared Error (MSE) of 0.03399 and a Mean Absolute Error (MAE) of 0.00278, resulting in a high Concordance Correlation Coefficient (CCC) of 0.9974. For arousal, the model shows moderate performance, with an MSE of 0.2470 and an MAE of 0.1101, leading to a CCC of 0.3513. The dominance predictions yield good results, with an MSE of 0.1022, an MAE of 0.0336, and a CCC of 0.8574.

The arousal result shows low performance which indicates LSTM algorithm is challenged to capture the arousal feature emotion recognition. The detailed predictions for valence, arousal, and dominance across various instances looks like the following.

25/25  2s 65ms/step

Evaluation Metrics:

Valence - MSE: 0.033997952980818126, MAE: 0.002781825201923766, CCC: 0.9974153903625993

Arousal - MSE: 0.2470386133914415, MAE: 0.11000596127328853, CCC: 0.3510630593924452

Dominance - MSE: 0.10226009767640873, MAE: 0.03360311739376688, CCC: 0.8574003890982022

	Valence PL	Valence AL	Arousal PL	Arousal AL	Dominance PL	Dominance AL
0	-0.641707	-0.64	0.310177	0.60	-0.416402	-0.43
1	-0.580761	-0.63	0.207604	-0.27	-0.214670	-0.33
2	0.750671	0.76	0.474622	0.48	0.347274	0.35
3	0.746574	0.76	0.473825	0.48	0.342692	0.35
4	-0.580696	-0.63	0.207494	-0.27	-0.214210	-0.33
5	0.750511	0.76	0.473583	0.48	0.346185	0.35
6	-0.580937	-0.43	0.207902	0.67	-0.215923	0.34
7	-0.451246	-0.43	0.676952	0.67	0.283854	0.34
8	-0.451246	-0.43	0.676952	0.67	0.283854	0.34
9	0.746616	0.76	0.473824	0.48	0.342700	0.35

Figure 4.7: Parallel comparison of the model performance with prediction score and actual score for three dimensions

The performance of the BiLSTM model using only acoustic features is illustrated in the provided on the following loss graph Figure 4.8. The training loss shows a significant initial decrease, indicating effective learning during the early stages of training. As the epochs progress, the training loss stabilizes at a low value, suggesting that the model is successfully capturing the underlying patterns in the training data. The validation loss also exhibits a downward trend. This indicates that the model is generalizing well to unseen data, although the variability suggests occasional challenges in maintaining consistency. Both training and validation losses converge towards the end of the training process, reflecting a strong performance of the BiLSTM model in recognizing emotions from Amharic speech based solely on acoustic features. The results indicate that the model is capable of effectively learning and generalizing, making it a promising approach for emotion recognition tasks.

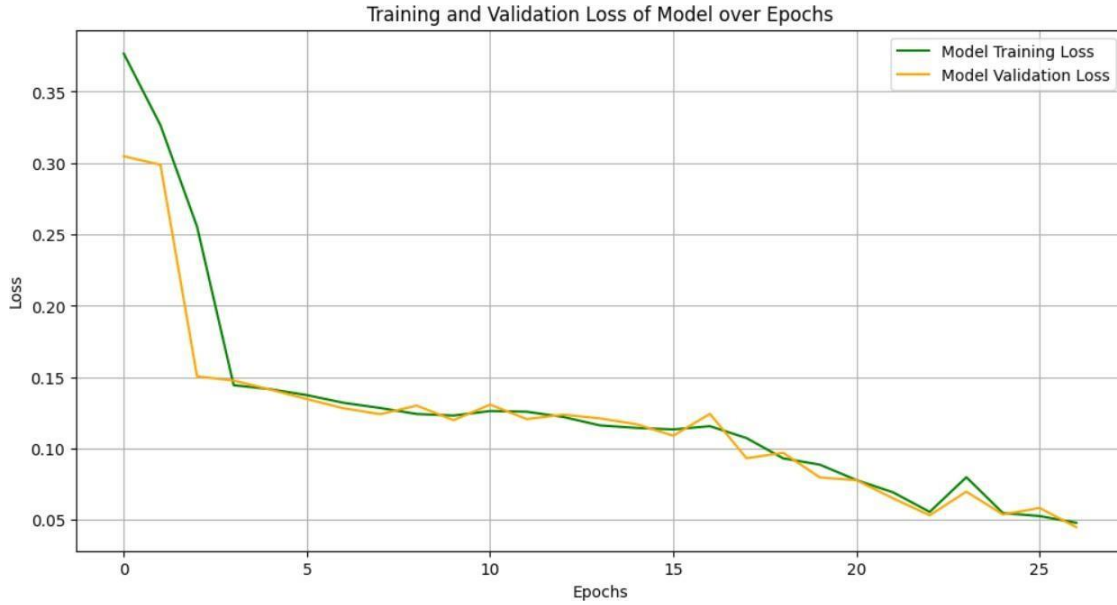


Figure 4.8: Dimensional Amharic SER model performance using BiLSTM algorithm along with acoustic feature

The evaluation metrics for the BiLSTM model using only acoustic features, shows promising performance across the emotional dimensions of valence, arousal, and dominance as shown on Figure 4.9. For valence, the model achieved MSE of 0.02374, MAE of 0.00128, and CCC of 0.9994, indicating excellent predictive accuracy. In terms of arousal, the model recorded an MSE of 0.06735, MAE of 0.02737, and CCC of 0.8957, suggesting good performance though slightly less accurate than for valence. For dominance, the model produced an MSE of 0.05177, MAE of 0.01199, along with CCC of 0.9534, indicating strong performance in this dimension as well. These evaluation metrics demonstrate that the BiLSTM model effectively recognizes emotions in Amharic speech.

25/25 ————— 4s 152ms/step

Evaluation Metrics:

Valence - MSE: 0.023745632986711028, MAE: 0.001277653006794801, CCC: 0.9994440269634097

Arousal - MSE: 0.06735178434360349, MAE: 0.027374571889059313, CCC: 0.8956652030975218

Dominance - MSE: 0.051772702736219575, MAE: 0.011989858022380324, CCC: 0.9535792858018319

	Valence PL	Valence AL	Arousal PL	Arousal AL	Dominance PL	Dominance AL
0	-0.610096	-0.63	-0.273890	-0.27	-0.300766	-0.33
1	-0.462280	-0.43	0.638415	0.67	0.295455	0.34
2	0.751146	0.76	0.479935	0.48	0.363304	0.35
3	-0.670433	-0.63	-0.292109	-0.27	-0.311836	-0.33
4	-0.409208	-0.43	0.642595	0.67	0.316390	0.34
5	-0.646023	-0.63	-0.294253	-0.27	-0.291574	-0.33
6	-0.663799	-0.63	-0.140858	-0.27	-0.381657	-0.33
7	-0.601269	-0.63	-0.118949	-0.27	-0.330206	-0.33
8	-0.620956	-0.64	0.441793	0.60	-0.419105	-0.43
9	-0.522705	-0.64	0.641748	0.60	-0.024284	-0.43

Figure 4.9: Parallel comparison of the model performance with prediction score and actual score for three dimensions

4.4.3. Dimensional Amharic SER model using bimodal feature and BiLSTM algorithm

In our study, we developed a dimensional Amharic SER model employing both acoustic and linguistic features, implemented with LSTM algorithm. The results of the training indicate an improvement in performance over the epochs. The training loss began at approximately 0.40 and rapidly decreased, stabilizing around 0.05 after 15 epochs. Similarly, the validation loss started at around 0.15 and also converged to approximately 0.05, demonstrating that the model effectively learned to generalize from the training data. This consistent decline in both training and validation loss suggests that the model can recognizing emotions in Amharic speech.

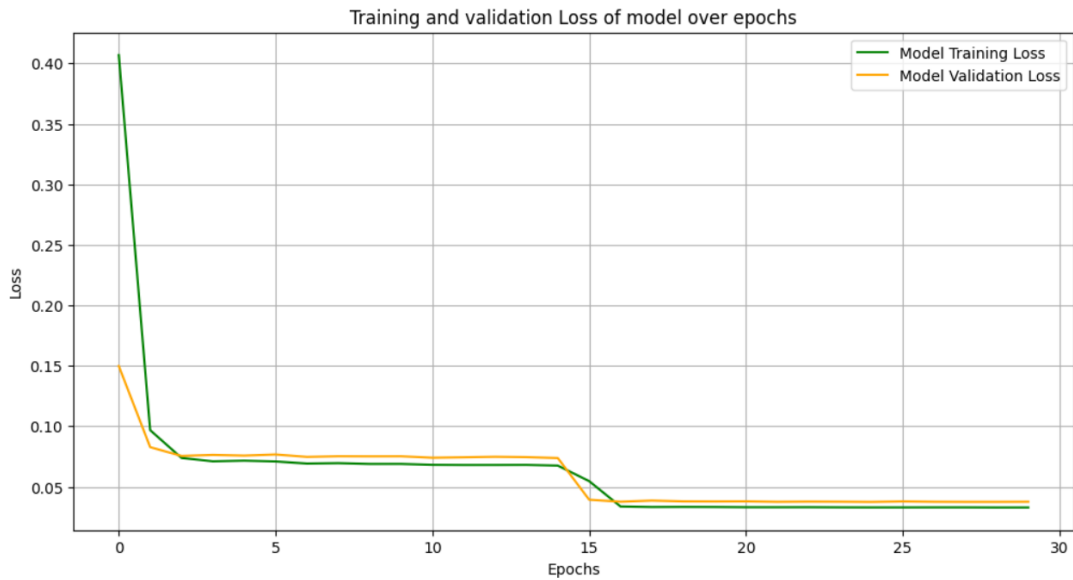


Figure 4.10: Training and validation loss of dimensional Amharic SER model using bimodal feature and LSTM algorithm

The evaluation results of the model use metrics for valence, arousal, and dominance in the Amharic SER task. MSE for valence is 0.058, with a MAE of 0.014, and a CCC of 0.978, indicating a strong correlation between predicted and actual values. For arousal, the MSE is 0.126, with an MAE of 0.050 and a CCC of 0.792, suggesting moderate performance. The dominance metrics show an MSE of 0.109, an MAE of 0.032, and a CCC of 0.862, reflecting a good level of agreement.

25/25 ————— 6s 192ms/step
 Evaluation Metrics:
 Valence - MSE: 0.058007281307374314, MAE: 0.014084213684616488, CCC: 0.9780804918322815
 Arousal - MSE: 0.12653119022672982, MAE: 0.049994293172002956, CCC: 0.7922151838080408
 Dominance - MSE: 0.10891567363209363, MAE: 0.032371169035661077, CCC: 0.8617875566082825

	Valence PL	Valence AL	Arousal PL	Arousal AL	Dominance PL	Dominance AL
0	-0.476611	-0.43	0.653980	0.67	0.241390	0.34
1	-0.607116	-0.63	-0.165951	-0.27	-0.357993	-0.33
2	-0.693170	-0.63	-0.156312	-0.27	-0.377686	-0.33
3	-0.607891	-0.63	-0.115784	-0.27	-0.275725	-0.33
4	-0.537655	-0.64	0.535151	0.60	-0.151082	-0.43
5	-0.588176	-0.63	0.192102	-0.27	-0.174890	-0.33
6	-0.669700	-0.63	-0.256221	-0.27	-0.318264	-0.33
7	-0.534748	-0.64	0.587071	0.60	-0.121550	-0.43
8	-0.614289	-0.63	0.060838	-0.27	-0.181862	-0.33
9	-0.506490	-0.43	0.624246	0.67	0.189452	0.34

Figure 4.11: Evaluation metrics for three dimensions

The results of the Dimensional Amharic SER model, which employs a BiLSTM algorithm and integrates bimodal features (acoustic and linguistic) are shown on the following loss graphs Figure 4.12. The training and validation loss across 40 epochs begins at relatively high values, approximately 0.5 for both metrics. However, within the first few epochs, there is a notable decrease in loss, demonstrating that the model rapidly learns to reduce its error. As training progresses, the training loss continues to decrease gradually, stabilizing at low loss by the end of the training period. The validation loss follows a similar way, demonstrating a consistent decline and stabilizing at a comparable level. This close alignment between the training and validation losses suggests that the model is effectively generalizing to unseen data without overfitting.

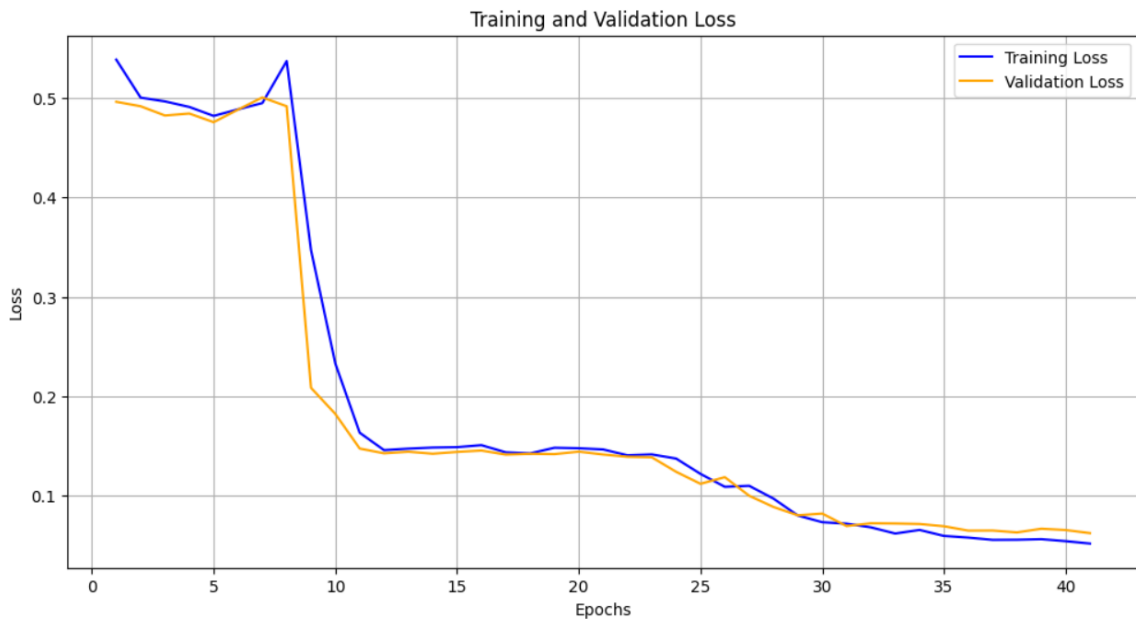


Figure 4.12: Training and validation loss of the model using bimodal feature with BiLSTM

Figure 4.12 shows the overall model performance of three dimensions. We evaluate the model performance on individual emotional dimensions: valence, arousal, and dominance. The following Figure 4.13 shows the model performance on individual dimensions such as dominance, arousal, and valence.

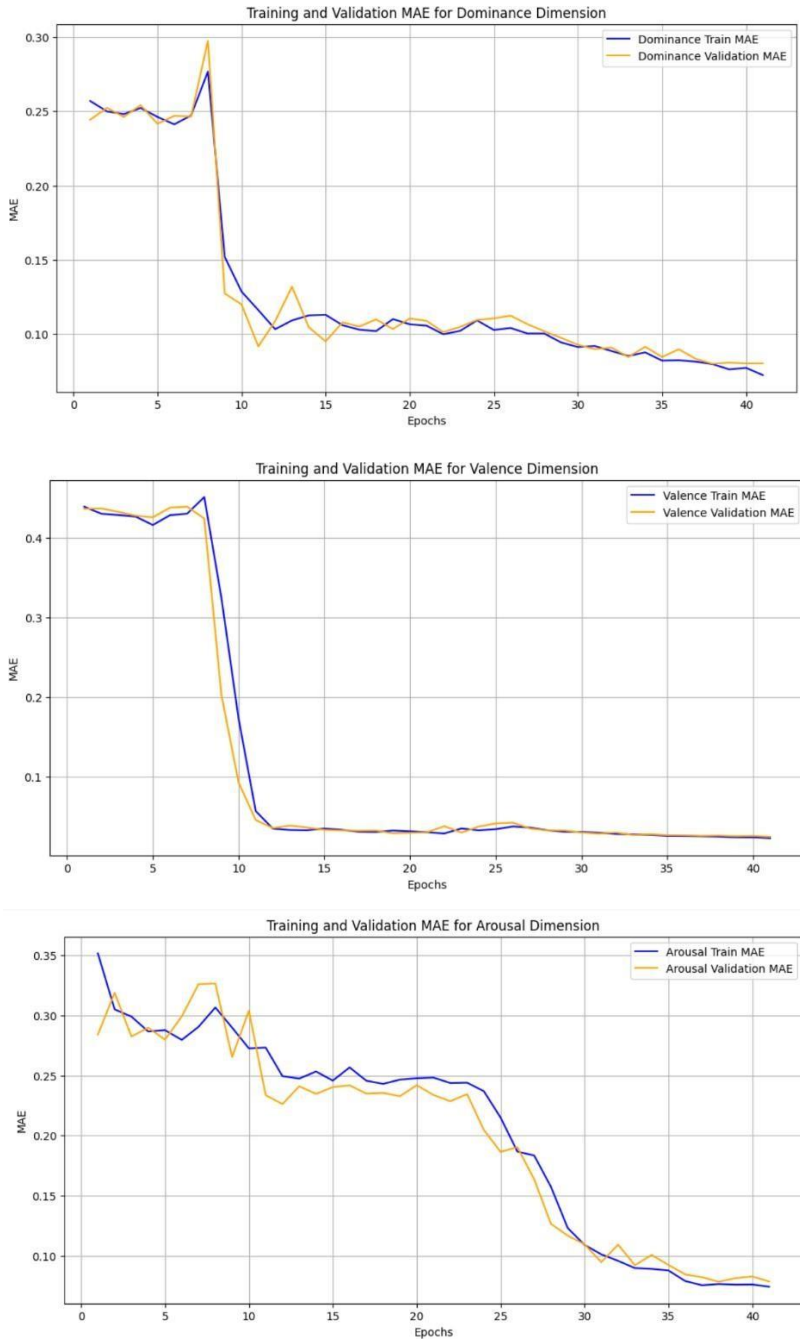


Figure 4.13: Training and validation loss on separate dimension of Amharic SER using acoustic and linguistic feature

The above Figure 4.13 shows the evaluation results for the model utilizing both linguistic and acoustic features. We presented through the training and testing loss graphs for the three emotional dimensions: dominance, arousal, and valence.

Dominance: The result shows a significant initial drop in training loss, indicating effective learning. The validation loss also decreases and stabilizes at a low value, suggesting that the model generalizes well to unseen data. This reflects the model capability to accurately capture the nuances of dominance in speech.

Valence: Similar to dominance, the training loss for valence exhibits a sharp decline, while the validation loss follows a comparable trend. Both losses stabilize at low levels, indicating that the model effectively learns and generalizes the emotional dimension of valence.

Arousal: The training loss for arousal shows a notable decrease, with the validation loss also declining significantly. It stabilizes at a low value, demonstrating the model ability to recognize arousal effectively.

The evaluation metrics shows the model performance across three emotional dimensions: Arousal, Valence, and Dominance. For Valence, the MSE is notably low at 0.0081, with an MAE of 0.00049, indicating high accuracy in predictions, as reflected by a perfect CCC of 1.000. The Arousal dimension shows a higher MSE of 0.0655 and MAE of 0.0321, with a CCC of 0.873, suggesting good predictive capability. Dominance dimension shows MSE of 0.0239 and an MAE of 0.0061, accompanied by a CCC of 0.978, indicating strong performance. The individual rows detail specific predictions and their corresponding values, showcasing the model ability to capture emotional nuances effectively as shown on Figure 4.14.

25/25 ————— 3s 92ms/step
 Evaluation Metrics:
 Valence - MSE: 0.008108576656761608, MAE: 0.0004900496170537422, CCC: 1.0006101043586246
 Arousal - MSE: 0.06552207300666046, MAE: 0.03215299155979922, CCC: 0.8738496949146903
 Dominance - MSE: 0.023963399675225178, MAE: 0.006115536711107993, CCC: 0.9775758284326713

	Valence PL	Valence AL	Arousal PL	Arousal AL	Dominance PL	Dominance AL
0	-0.636662	-0.64	0.595214	0.60	-0.429954	-0.43
1	-0.630002	-0.63	-0.270776	-0.27	-0.332526	-0.33
2	0.761353	0.76	0.479471	0.48	0.349870	0.35
3	0.762484	0.76	0.483629	0.48	0.352614	0.35
4	-0.630850	-0.63	-0.270581	-0.27	-0.332507	-0.33
5	0.768745	0.76	0.474175	0.48	0.351055	0.35
6	-0.428654	-0.43	0.665701	0.67	0.334185	0.34
7	-0.431841	-0.43	0.675249	0.67	0.332502	0.34
8	-0.427629	-0.43	0.660032	0.67	0.345059	0.34
9	0.758967	0.76	0.479702	0.48	0.353443	0.35

Figure 4.14: Evaluation metrics on each dimensions using bimodal feature with BiLSTM

From the above figure 4.12, 4.13, and 4.14 we observe that applying linguistic and acoustic feature highly improve the model performance in the task of predicting emotion with three dimensions. This linguistic feature shows promising improvement on those dimensions’ prediction.

4.4.4. Overall Model Performance

In this subsection, we present the overall model performance with both categorical and dimensional SER. Table 4.3 shows the overall performance summarization of two approaches to SER in Amharic: categorical and dimensional.

4.5 Discussion

In this section, we provide a comprehensive analysis of the results generated by our proposed model for categorical and dimensional Amharic SER. The findings thoroughly examined in relation to our established research objectives, allowing us to assess the performance and effectiveness of the model. We investigate into how the combination of both acoustic and linguistic features enhances the model ability to accurately predict emotional states. Furthermore, we compare these results with the categorical SER approach, offering insights into the comparative advantages and limitations of each methodology. This detailed discussion not only highlight the strengths of our proposed models but also to contribute to the broader understanding of emotion recognition in the

Table 4.3: overview of model performance categorical and dimensional along with acoustic feature only and bimodal feature

Categorical approach				
Accuracy for BiLSTM		Training = 99.35%, validation acc = 99.14%, and Test =98%		
Dimensional Amharic SER using acoustic feature only				
Dimensions and Algorithms		MSE	MAE	CCC
Valence	LSTM	0.03399	0.00278	0.9974
	BiLSTM	0.02374	0.00128	0.9994
Arousal	LSTM	0.2470	0.1101	0.3513
	BiLSTM	0.06735	0.02737	0.8957
Dominance	LSTM	0.1022	0.0336	0.8574
	BiLSTM	0.05177	0.01199	0.9534
Dimensional Amharic SER using acoustic and linguistic feature				
Dimensions and Algorithms		MSE	MAE	CCC
Valence	LSTM	0.0580	0.0140	0.9780
	BiLSTM	0.0081	0.00049	1.0000
Arousal	LSTM	0.1265	0.0499	0.7922
	BiLSTM	0.0655	0.0321	0.8738
Dominance	LSTM	0.1089	0.0323	0.8617
	BiLSTM	0.0239	0.0061	0.9775

Amharic language, ultimately paving the way for future advancements in this area of research.

4.5.1 Experimental Result of Categorical SER

The categorical approach evaluates how effectively the model can classify emotions into predefined categories. Here, the BiLSTM model achieves an accuracy of 99.14%, underscoring its ability to capture and distinguish between different emotional states in Amharic speech. This high accuracy demonstrates the robustness of the bidirectional LSTM in processing sequential data, where it can utilize both past and future contexts to make more accurate predictions. The detailed result is explained on the Figure 4.4 confusion matrix analysis. As the above Figure 4.3 and Figure 4.4 shows, our model performs best as comparing Ephrem et al [11] work. This result is achieved by using hyperparameter tuning, use combination of ZCR, RMS, and MFCC feature as combination of feature enhance the performance[47], data augmentation, and advanced deep learning algorithm.

4.5.2. Experimental Result of Dimensional SER using Acoustic Feature

Dimensional emotion recognition involves predicting continuous values for emotional dimensions such as valence, arousal, and dominance. For the valence dimension, which represents the pleasantness of the emotion, BiLSTM outperforms LSTM across all metrics. The MSE and MAE for BiLSTM are significantly lower, and the CCC is closer to 1, indicating a higher agreement between the predicted and true values. This suggests that BiLSTM is better at capturing the nuances of the pleasantness in speech signals. This is because BiLSTM algorithm capture in both backward and forward direction. In the Arousal dimension, which indicates the intensity of the emotion, the performance improvement with BiLSTM is particularly outstanding. The MSE and MAE are considerably reduced with BiLSTM, and the CCC increases dramatically from 0.3513 to 0.8957. This indicates that BiLSTM is far superior in capturing the variations in emotional intensity, making it a more reliable model for recognizing the energy or excitement levels in speech. Moreover, for Dominance, which reflects the level of control in the emotion, BiLSTM again shows

superior performance with lower MSE and MAE, and higher CCC. This demonstrates that BiLSTM is more effective in understanding the control dynamics in emotional speech.

4.5.3. Experimental Result of Dimensional SER using Acoustic and Linguistic Feature

The outcomes of the dimensional SER experiment demonstrate a significant improvement not only in valence dimension but also arousal and dominance dimension prediction when utilizing both acoustic and linguistic features compared to using merely acoustic features. When using both acoustic and linguistic features, the performance in the Valence dimension improves significantly for both models, but especially for BiLSTM, which achieves perfect prediction accuracy with a CCC of 1.0. This indicates that the additional linguistic information helps the model better understand and predict the pleasantness of emotions. The addition of linguistic features also enhances the model ability to predict arousal. The BiLSTM model again outperforms the LSTM model with lower MSE and MAE, and higher CCC. This suggests that combining acoustic and linguistic features allows the model to capture the emotional intensity more accurately. In addition to arousal and valence, the BiLSTM model achieves significantly better performance with lower MSE and MAE and higher CCC for dominance dimension. The combined features enable the model to better understand and predict the level of control or influence in the emotion.

Our experiment shows that across both categorical and dimensional approaches, the BiLSTM model consistently outperforms the LSTM model. This highlights the importance of bidirectional processing in capturing the full context of the speech signal, which is crucial for accurate emotion recognition. The results clearly show that integrating acoustic and linguistic features enhances the model performance. This is particularly obvious in the dimensional approach, where the addition of linguistic features results in substantial improvements in prediction accuracy and agreement with true values. This finding emphasizes the importance of incorporating multiple feature types in emotion recognition tasks. The significant enhancement in those dimensions' prediction highlights the potential for improved emotional understanding in applications such as sentiment analysis, mental health, human-computer interaction, and other SER applications.

4.6 Error Analysis

4.6.1 Categorical Amharic SER model performance error analysis

As shown on Figure 4.4, the confusion matrix clearly shows some emotions are misclassified especially angry, fear, and neutral emotions. This error happened because of overlapping the emotional states.

Ambiguity in Emotions: Emotions often overlap, and a single speech sample express multiple emotion simultaneously. For example, a speech that is classified as "angry" might also convey elements of "frustration" or "disappointment". Some happy and angry emotions have similar sound. This overlap can lead to misclassification as the model may struggle to pinpoint the dominant emotion.

Subjectivity of Emotions: Emotions are subjective experiences, and different listeners may interpret the same speech differently. This subjectivity can lead to inconsistencies in labeling during training, which in turn affects the model ability to generalize.

4.6.2 Dimensional Amharic SER using merely acoustic feature and bimodal features model performance error analysis

Relying solely on acoustic features is not capture the full complexity of emotional expression. This limitation can lead to misprediction, especially for emotions that share similar acoustic characteristics, such as "sadness" and "fear" for valence dimension and "happy" and "angry" for arousal dimension. In case of bimodal feature, the metrics shows outstanding performance but can improve by maximizing data size especially diverse linguistic data that can hold more emotion nuance of semantics words.

4.7 Answering the research questions

- To what extent does the deep learning approach improve SER?

The deep learning approach significantly improves SER by using complex neural network architectures, such as LSTM and BiLSTM, which can effectively model sequential data. In the case of the categorical Amharic SER model, the use of deep learning algorithm resulted in improved accuracy. The BiLSTM model achieved an impressive accuracy of 98% in categorical SER which shows 8.013% improvement on the bench mark. This is because we use combination of three feature such as MFCC, ZCR, and RMS that able to

enhance accuracy [46]. Furthermore, this study incorporates data augmentation that can simulate real time speech, and use advanced deep learning algorithm such as BiLSTM that can process in both backward and forward direction. Dimensional Amharic SER model, the use of deep learning techniques resulted in a marked high predictive performance demonstrated low error rates in dimensional predictions, with MSE values for valence, arousal, and dominance being notably low. In acoustic features, the BiLSTM consistently outperforms the LSTM model across all emotional dimensions' arousal, valence, and dominance. The CCC metrics shows the BiLSTM algorithm outperform the LSTM algorithm for valence, arousal, and dominance with 0.20%, 154.95%, and 11.21% respectively for merely acoustic feature. In addition to acoustic feature, bimodal feature also shows a promising improvement. BiLSTM algorithm improves 2.25%, 10.30%, and 13.43% for valence, arousal, and dominance respectively. Moreover, the MSE and MAE also shows that the BiLSTM algorithm superior from LSTM algorithm.

➤ To what extent does the dimensional SER vary from the categorical SER?

As illustrated in Figure 4.14, dimensional SER provides a more nuanced representation of emotions compared to categorical SER. For instance, the first sample, classified as "fear" in the categorical SER model, is categorized as a "distressed" emotion in the dimensional framework. Because the predicted valence score for this sample is -0.6367, which is slightly higher than the categorical ground truth of -0.64, paired with a relatively low arousal value. Furthermore, for samples 2, 3, 5, and 9, which are all categorized as "happy" emotions in the categorical SER model, the predicted valence, arousal, and dominance values range from 0.7589 to 0.7687 for valence, 0.4742 to 0.4836 for arousal, and 0.3499 to 0.3534 for dominance. These values show a slight fluctuation around the categorical ground truth benchmarks of 0.76, 0.48, and 0.35, respectively. This variability illustrates that these samples represent a spectrum of happiness, ranging from lower levels of happiness to delighted. Similarly, instance number 8, classified as "anger" in the categorical SER model, presents a slightly lower arousal prediction along with a higher valence score in the dimensional analysis. This score suggests this sample ranges between anger and annoyed emotion.

To sum up, the findings highlight that dimensional SER can more effectively capture the complexities of emotional expressions than categorical models. By representing emotions as a variety of valence, arousal, and dominance, the dimensional approach allows a richer interpretation of emotional states than categorical classifications. This nuanced understanding enhances applications in fields requiring emotional intelligence

- What effect does the linguistic feature provide to the performance of valence prediction in dimensional Amharic SER?

The inclusion of linguistic features significantly enhances the performance of not only valence prediction but also arousal and dominance prediction in dimensional Amharic SER. Linguistic features provide contextual information that complements acoustic features, allowing the model to better understand the semantics of speech. In the case of BiLSTM algorithm, linguistic feature shows the improvement of valence prediction on MSE, MAE, and CCC of 65.86%, 61.72%, and 0.06% respectively. The dominance dimension shows 53.83%, 49.04%, and 2.53% improvement on MSE, MAE, and CCC respectively. The LSTM algorithm shows improvement on linguistic feature of 48.78%, 54.67%, and 125.56% on MSE, MAE, and CCC for arousal dimension. These improvement shows model enhancement when moving from using only acoustic features to bimodal features (acoustic and linguistic).

CHAPTER FIVE: CONCLUSION AND RECOMMENDATION

5.1 Conclusion

In this thesis work, we showed the effectiveness of a dimensional Amharic SER model developed using deep learning techniques. Our research focused on recognizing emotional dimensions such as valence, arousal, and dominance within Amharic speech, addressing a significant gap in the existing literature on emotion recognition for underrepresented languages. We demonstrated that the integration of both acoustic and linguistic features significantly enhances the model performance, allowing for a more nuanced understanding of emotional expressions. We conducted a thorough validation of our approach, which demonstrated that the dimensional model surpasses traditional categorical methods in its performance to accurately capture the complexities of human emotions.

Our categorical SER model, primarily designed for comparison with the dimensional SER model, achieved an impressive performance, surpassing the baseline of our study with an accuracy of 99.14% on validation data and 98% on test data. The dimensional Amharic SER model demonstrates the best performance, exhibiting low error rates across all three dimensions. The MSE values for valence, arousal, and dominance are 0.0081, 0.0655, and 0.0239, respectively. Additionally, the MAE values are 0.00049 for valence, 0.0321 for arousal, and 0.0061 for dominance, all achieved using bimodal features. The CCC metrics indicate that the BiLSTM algorithm surpasses the LSTM algorithm in valence, arousal, and dominance, with improvements of 0.20%, 154.95%, and 11.21% respectively when only acoustic features are considered. When incorporating bimodal features, the BiLSTM algorithm demonstrates further enhancements, achieving increases of 2.25%, 10.30%, and 13.43% for valence, arousal, and dominance, respectively. Additionally, both the MSE and MAE metrics confirm that the BiLSTM algorithm is superior to the LSTM algorithm. The development of the dimensional SER (SER) model facilitates the practical application including customer service, education, healthcare, and public presentation skill coaching. By utilizing emotion-aware systems, organizations can enhance user engagement and create more personalized experiences that address the emotional desires of individuals.

Contribution

In this research, we explore the development of a Dimensional SER model for the Amharic language using Deep Learning techniques. Our work enhances the understanding of emotional expression in speech and its practical applications across various domains. The key contributions of this study are as follows:

1. We annotated a publicly available Amharic SER dataset in a dimensional space, focusing on critical dimensions such as valence, arousal, and dominance, thereby enriching the dataset for future research endeavors.
2. We develop user-friendly annotation tools to easy and help the annotation team.
3. We developed a categorical Amharic SER model to make a comparison from baseline and the dimensional model, providing appreciated insights into the effectiveness of different approaches.
4. We develop dimensional Amharic SER model utilizing both acoustic features alone and a combination of acoustic and linguistic features, allowing for analysis of emotional recognition factors.
5. We conducted a thorough comparison of the performance between categorical and dimensional SER models, and the improvements in valence prediction achieved through the integration of linguistic features.

The study contributes in answer to the challenges faced by Amharic speakers in communication and emotional understanding, our research offers significant contributions at development a more inclusive technological landscape. We developed a dimensional SER model specifically tailored for Amharic speakers, which enhances communication technologies and facilitates a better understanding of emotional content in Amharic speech.

Limitations

This study has the following potential limitations. To begin with, the size of dataset is small and not included two basic emotions disgust and surprise [18], and challenged to classify and predict those emotions. In addition, the subjectivity in emotion annotation[12] the process of assigning precise dimensional labels (valence, arousal, and dominance) to

speech recordings is difficult. Moreover, the study can't classify or predict non-speech-based emotion like music which is important to analyze the emotion perceived in music and other non-verbal speech.

Challenges

The challenges faced in developing the dimensional Amharic SER model include the following:

- Scarcity of annotated datasets for Amharic speech, especially in a dimensional emotional context, poses a significant challenge, as there are currently no annotated datasets available in dimensional space.
- The annotation team were challenged to provide the accurate VAD score as emotion is subjective [12].
- We faced challenges in transcribing Amharic speech due to the absence of a pretrained model that can accurately transcribe the Amharic language.

5.2 Future Work and Recommendation

For the future research in the area of dimensional Amharic SER, the following works can enhance and contribute SER.

Increase dataset: including additional samples, particularly targeting emotions such as disgust and surprise. Consider annotating the dataset by measuring the energy of speech, which can provide valuable cues for recognizing these emotions effectively.

Multimodal Emotion Recognition for Amharic: combining speech, text, and image data significantly improve the robustness and effectiveness of recognizing emotion, as different modalities can complement each other.

Exploration of Other Under-resourced Languages: Conduct similar studies with other under-resourced languages such as Tigrigna and Oromifa. This is not only contributed to the area of emotion recognition but also aid in developing SER models for diverse linguistic communities.

By considering these future directions into account, researchers can further progress in the area of Amharic SER, successfully address the challenges associated with under-resourced languages, and make important contributions to the wider field of emotion recognition. This continuous ongoing improve the comprehension of emotional expression across various linguistic contexts and facilitate the development of inclusive and effective emotion recognition systems worldwide.

REFERENCE

- [1] M. M. V. Chalapathi, M. R. Kumar, N. Sharma, and S. Shitharth, "Ensemble Learning by High-Dimensional Acoustic Features for Emotion Recognition from Speech Audio Signal," *Security and Communication Networks*, vol. 2022, 2022, doi: 10.1155/2022/8777026.
- [2] B. Fernandes and K. Mannepalli, "Speech emotion recognition using deep learning lstm for tamil language," *Pertanika J Sci Technol*, vol. 29, no. 3, pp. 1915–1936, Jul. 2021, doi: 10.47836/PJST.29.3.33.
- [3] S. Chen, J. Zhao, Q. Jin, and S. Wang, "Multimodal multi-task learning for dimensional and continuous emotion recognition," *AVEC 2017 - Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge, co-located with MM 2017*, pp. 19–26, Oct. 2017, doi: 10.1145/3133944.3133949.
- [4] H. Chen, S. Wu, X. Wu, Z. Lin, and S. W. X. W. Z. L. Hanwen Chen, "Speech emotion recognition using multiple classification models based on MFCC feature values," *Applied and Computational Engineering*, vol. ACE Vol.6, no. 1, pp. 1037–1047, Jun. 2023, doi: 10.54254/2755-2721/6/20230449.
- [5] O. ' Shaughnessy *et al.*, "Cross-Corpus Multilingual Speech Emotion Recognition: Amharic vs. Other Languages," *Applied Sciences 2023, Vol. 13, Page 12587*, vol. 13, no. 23, p. 12587, Nov. 2023, doi: 10.3390/APP132312587.
- [6] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Speech Emotion Recognition with Emotion-Pair Based Framework Considering Emotion Distribution Information in Dimensional Emotion Space," *Interspeech*, vol. 2017-August, pp. 1238–1242, 2017, doi: 10.21437/INTERSPEECH.2017-619.
- [7] H.-G. ; Kim *et al.*, "Emotional Stress Recognition Using Electroencephalogram Signals Based on a Three-Dimensional Convolutional Gated Self-Attention Deep Neural Network," *Applied Sciences 2022, Vol. 12, Page 11162*, vol. 12, no. 21, p. 11162, Nov. 2022, doi: 10.3390/APP122111162.
- [8] D. Kollias and S. Zafeiriou, "Exploiting Multi-CNN Features in CNN-RNN Based Dimensional Emotion Recognition on the OMG in-the-Wild Dataset," *IEEE Trans Affect Comput*, vol. 12, no. 3, pp. 595–606, Jul. 2021, doi: 10.1109/TAFFC.2020.3014171.
- [9] K. Chauhan, K. K. Sharma, and T. Varma, "Speech Emotion Recognition Using Convolution Neural Networks," *Proceedings - International Conference on Artificial Intelligence and Smart Systems, ICAIS 2021*, pp. 1176–1181, Mar. 2021, doi: 10.1109/ICAIS50930.2021.9395844.
- [10] M. Qi and H. Zhang, "Dimensional emotion recognition based on two stream CNN fusion attention mechanism," *SPIE*, vol. 12699, p. 126990O, May 2023, doi: 10.1117/12.2678902.

- [11] E. A. Retta *et al.*, “A New Amharic Speech Emotion Dataset and Classification Benchmark,” *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 1, pp. 1–22, Jan. 2022, doi: 10.1145/3529759.
- [12] C. Barhoumi, Y. Ben, A. Multimedia, and Y. Ben Ayed, “Real-Time Speech Emotion Recognition Using Deep Learning and Data Augmentation,” 2023, doi: 10.21203/rs.3.rs-2874039/v1.
- [13] B. T. Atmaja and M. Akagi, “Improving Valence Prediction in Dimensional Speech Emotion Recognition Using Linguistic Information,” *Proceedings of 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-Ordination and Standardisation of Speech Databases and Assessment Techniques, O-COCOSDA 2020*, pp. 166–171, Nov. 2020, doi: 10.1109/O-COCOSDA50338.2020.9295032.
- [14] H. Zhang, J. Yin, and X. Zhang, “The Study of a Five-Dimensional Emotional Model for Facial Emotion Recognition,” *Mobile Information Systems*, vol. 2020, 2020, doi: 10.1155/2020/8860608.
- [15] I. J. Roseman and C. A. Smith, “Appraisal Theory Overview, Assumptions, Varieties, Controversies,” *Appraisal Processes in Emotion*, pp. 3–19, Nov. 2023, doi: 10.1093/OSO/9780195130072.003.0001.
- [16] K. Tatlılıoğlu and N. Senchylo-Tatlılıoğlu, “A Theoretical Perspective on Psycholinguistics,” *Psycholinguistics in a Modern World*, vol. 15, pp. 241–245, Dec. 2020, doi: 10.31470/2706-7904-2020-15-241-245.
- [17] “(PDF) Geneva Emotion Wheel Rating Study.” Accessed: Aug. 30, 2024. [Online]. Available: https://www.researchgate.net/publication/280880848_Geneva_Emotion_Wheel_Rating_Study
- [18] “Emotion classification - Wikipedia.” Accessed: Aug. 30, 2024. [Online]. Available: https://en.wikipedia.org/wiki/Emotion_classification
- [19] “The 2D valence-arousal model of emotion proposed by Russel [50] | Download Scientific Diagram.” Accessed: Aug. 30, 2024. [Online]. Available: https://www.researchgate.net/figure/The-2D-valence-arousal-model-of-emotion-proposed-by-Russel-50_fig1_330861905
- [20] B. W. Schuller and A. M. Batliner, “Computational paralinguistics: Emotion, affect and personality in speech and language processing,” *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, pp. 1–321, Oct. 2013, doi: 10.1002/9781118706664.
- [21] F. Yang, X. Zhao, W. Jiang, P. Gao, and G. Liu, “Multi-method Fusion of Cross-Subject Emotion Recognition Based on High-Dimensional EEG Features,” *Front Comput Neurosci*, vol. 13, p. 473322, Aug. 2019, doi: 10.3389/FNCOM.2019.00053/BIBTEX.

- [22] S. A, M. H, and H. N, “Effective Feature Selection in Speech Emotion Recognition Systems using Generative Adversarial Networks,” Nov. 2022, doi: 10.21203/RS.3.RS-2244414/V1.
- [23] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, “Analysis of emotional speech—a review,” *Intelligent Systems Reference Library*, vol. 105, pp. 205–238, Mar. 2016, doi: 10.1007/978-3-319-31056-5_11.
- [24] G. Fairbanks and L. W. Hoaglin, “An experimental study of the durational characteristics of the voice during the expression of emotion,” *Speech Monographs*, vol. 8, no. 1, pp. 85–90, Dec. 1941, doi: 10.1080/03637754109374888.
- [25] H. Tang, W. Liu, W. L. Zheng, and B. L. Lu, “Multimodal Emotion Recognition Using Deep Neural Networks,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10637 LNCS, pp. 811–819, 2017, doi: 10.1007/978-3-319-70093-9_86.
- [26] “Speech Emotion Recognition Based on Fuzzy Support Vector Machine,” *Machine Learning Theory and Practice*, vol. 3, no. 4, Dec. 2022, doi: 10.38007/ML.2022.030406.
- [27] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.
- [28] Y. Bengio, P. Simard, and P. Frasconi, “Learning Long-Term Dependencies with Gradient Descent is Difficult,” *IEEE Trans Neural Netw*, vol. 5, no. 2, pp. 157–166, 1994, doi: 10.1109/72.279181.
- [29] R. Pascanu, T. Mikolov, and Y. Bengio, “On the difficulty of training Recurrent Neural Networks,” Nov. 2012, Accessed: Aug. 30, 2024. [Online]. Available: <http://arxiv.org/abs/1211.5063>
- [30] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, and J. Schmidhuber, “LSTM: A Search Space Odyssey,” *IEEE Trans Neural Netw Learn Syst*, vol. 28, no. 10, pp. 2222–2232, Oct. 2017, doi: 10.1109/TNNLS.2016.2582924.
- [31] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/NECO.1997.9.8.1735.
- [32] “Structure of LSTM. Source: Olah, C. (2015). Understanding LSTM networks. | Download Scientific Diagram.” Accessed: Aug. 30, 2024. [Online]. Available: https://www.researchgate.net/figure/Structure-of-LSTM-Source-Olah-C-2015-Understanding-LSTM-networks_fig2_367764925
- [33] F. A. Gers, J. Schmidhuber, and F. Cummins, “Learning to forget: Continual prediction with LSTM,” *Neural Comput*, vol. 12, no. 10, pp. 2451–2471, 2000, doi: 10.1162/089976600300015015.

- [34] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Networks*, vol. 18, no. 5–6, pp. 602–610, Jul. 2005, doi: 10.1016/J.NEUNET.2005.06.042.
- [35] "(PDF) Dimensional and Discrete Emotion Recognition using Multi-task Learning from Acoustic and Linguistic features extracted from Speech." Accessed: Jan. 05, 2024. [Online]. Available: https://www.researchgate.net/publication/353287610_Dimensional_and_Discrete_Emotion_Recognition_using_Multi-task_Learning_from_Acoustic_and_Linguistic_features_extracted_from_Speech
- [36] H. Chen, S. Wu, X. Wu, Z. Lin, and S. W. X. W. Z. L. Hanwen Chen, "Speech emotion recognition using multiple classification models based on MFCC feature values," *Applied and Computational Engineering*, vol. ACE Vol.6, no. 1, pp. 1037–1047, Jun. 2023, doi: 10.54254/2755-2721/6/20230449.
- [37] T. Dimitrova-Grekow, A. Klis, and M. Igras-Cybulska, "Speech emotion recognition based on voice fundamental frequency," *Archives of Acoustics*, vol. 44, no. 2, pp. 277–286, 2019, doi: 10.24425/AOA.2019.128491.
- [38] B. Schuller, G. Rigol, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine - Belief network architecture," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 1, 2004, doi: 10.1109/ICASSP.2004.1326051.
- [39] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 65–68, 2009, doi: 10.1109/ICASSP.2009.4959521.
- [40] B. Kratzwald, S. Ilić, M. Kraus, S. Feuerriegel, and H. Prendinger, "Deep learning for affective computing: Text-based emotion recognition in decision support," *Decis Support Syst*, vol. 115, pp. 24–35, Nov. 2018, doi: 10.1016/J.DSS.2018.09.002.
- [41] E. D. Emiru, S. Xiong, Y. Li, A. Fesseha, and M. Diallo, "Improving Amharic Speech Recognition System Using Connectionist Temporal Classification with Attention Model and Phoneme-Based Byte-Pair-Encodings," *Information 2021, Vol. 12, Page 62*, vol. 12, no. 2, p. 62, Feb. 2021, doi: 10.3390/INFO12020062.
- [42] A. Antoniou, A. Storkey, and H. Edwards, "Data Augmentation Generative Adversarial Networks," *J Phys A Math Theor*, vol. 44, no. 44, pp. 1–13, 2017, Accessed: Aug. 30, 2024. [Online]. Available: <https://arxiv.org/abs/1711.04340v3>
- [43] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C. H. Lee, "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression," *IEEE Signal Process Lett*, vol. 27, pp. 1485–1489, 2020, doi: 10.1109/LSP.2020.3016837.

- [44] H. M and S. M.N, “A Review on Evaluation Metrics for Data Classification Evaluations,” *International Journal of Data Mining & Knowledge Management Process*, vol. 5, no. 2, pp. 01–11, Mar. 2015, doi: 10.5121/IJDKP.2015.5201.
- [45] NCSS, “PASS Sample Size Software 812-1 Lin’s Concordance Correlation Coefficient”.
- [46] “Deep Learning.” Accessed: Oct. 08, 2024. [Online]. Available: <https://www.deeplearningbook.org/>
- [47] N. Chauhan, T. Isshiki, and D. Li, “Speaker recognition using LPC, MFCC, ZCR features with ANN and SVM classifier for large input database,” *2019 IEEE 4th International Conference on Computer and Communication Systems, ICCCS 2019*, pp. 130–133, Feb. 2019, doi: 10.1109/CCOMS.2019.8821751.

Appendix

We have designed an annotation tool specifically for the annotation of Amharic speech in the context of Dimensional Amharic Speech Emotion Recognition (SER). This tool streamlines the workflow for our annotation team by providing several key features:

User-Friendly Interface: The tool features an intuitive interface that simplifies navigation, allowing annotators to work efficiently.

Audio Playback: Users can listen to Amharic speech samples directly within the tool, which is essential for accurately assessing the emotional content of the speech and provide scores for valence, arousal, and dominance.

Error Correction: It allows for easy updates to annotations if any errors are noticed during the annotation process, ensuring the final dataset is accurate.

Automatic Positioning: The tool automatically locates the position of the annotation, streamlining the workflow and reducing the time spent searching for specific audio segments.

Missing Values Identification: It highlights missing values or unannotated speech segments, making it straightforward for annotators to identify and complete any gaps.

Export Functionality: After completing the annotation process, it allows to export the results as an Excel file.

This annotation tool significantly enhances the efficiency of our annotation team, ensuring that we accurately capture the emotional dimensions of Amharic speech.

BAHIR DAR UNIVERSITY

BAHIR DAR INSTITUTE OF TECHNOLOGY

Dimensional Amharic Speech Emotion Recognition Annotation Tool

a5-04-02-02-60.wav

⏪ ⏩ ⏴ ⏵

Transcription:

Valence:

Arousal:

Dominance:

a5-03-01-02-41.wav	-0.45	0.60	0.32	የራስህ ጉዳይ አላቅሰህም
a5-03-01-02-39.wav	-0.43	0.67	0.34	የራስህ ጉዳይ አላቅሰህም
a5-03-01-02-30.wav	-0.46	0.55	0.30	የራስህ ጉዳይ አላቅሰህም
a5-03-01-02-28.wav	-0.43	0.69	0.35	የራስህ ጉዳይ አላቅሰህም
a5-03-01-02-26.wav	-0.43	0.73	0.35	የራስህ ጉዳይ አላቅሰህም

Unannotated File Name Actions