الجمهوريـــــــــــــة الجزائريـــــــــة الديمقراطيـــــــة الشعبيـــــة
وزارة التعليـــــــم العـــــــالي و البحـــــــث العلـــــمي
جامعة وهران للعلوم والتكنولوجيا محمد بوضياف
كلية الرياضيات و الاعلام الالي
People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research
University of Science and Technology of Oran - Mohamed BOUDIAF
Faculty of Mathematics and Computer Science
Department: *Computer Science*

Final Year Project

Arabic Speech Emotion Recognition System

To obtain the diploma of **Master's Degree**

Field: **Mathematics – Computer Science**

Specialization: **Computer Science**

Option: **Artificial Intelligence and its Applications (IAA)**

Presented on: 12 May 2025
By:
-Gheffari Youcef Soufiane
-Benouddane Oussama

Jury	Name & surname	Grade	University
President	Pr. Tlemsani R	Professor	USTOMB
Supervisor	Dr. Silarbi S	MCA	USTOMB
Examiner	Dr. Ourdighi A	MCB	USTOMB

2024/2025

Acknowledgment

First, we express our deepest gratitude to **Allah**, whose blessings and guidance have been our constant source of strength and inspiration throughout this journey.
We would like to extend our heartfelt thanks to our esteemed advisor, **Mme. Silarbi Samiya**, for her unwavering support, valuable guidance, and continual encouragement. Her expertise and dedication have played a crucial role in the successful completion of this work.
Our sincere appreciation goes to **Mr. Tlemsani Redouane**, President of the Jury, and **Mme.Ourdighi Asma**, our examiner, for their insightful feedback and constructive remarks.
To our beloved parents, we give them our deepest thanks for their unconditional love, patience, and constant support. Your faith in us has been the foundation of our motivation and achievements.
We are also truly grateful to the families **Gheffari** and **Benouddane** for their constant encouragement and support throughout this journey.
Finally, we extend our sincere thanks to all those who have contributed in any way to our academic and personal development. Whether through encouragement, guidance, or assistance, your support has meant a lot to us.
Thank you all.

Dedication

I dedicate this thesis to the memory of my cousin, **Touhami Walid**, and my Uncles, **Touhami Driss, Touhami Houari**, whose belief in me and enduring presence in my heart continue to inspire me every day.
To my beloved parents, **Malika** and **Boutkhil**, for their unconditional love, strength, and unwavering support and to my beloved uncle **Touhami Omar**.
To my dear siblings, **Zakaria** and **AbdelBasset**, for always being by my side.
A special thanks to the teacher **Neggaz Nabil** who help and support me when i needed support.
To all the friends I've met along the way—thank you for being part of this journey.
I also dedicate this work to my cherished group of friends who have been a constant source of support and motivation throughout my university life:
Ayoub ,Abdelouahab ,Yacine ,Mahmoud ,Soulaymane ,Achraf ,Rabeh ,Mouaad ,Khalil ,Mohammed ,Abdeljalil, and Dear Colleague **Oussama**.
The moments we shared, from late-night, last-minute study sessions filled with coffee and laughter, to makeup exams and spontaneous adventures, have left a lasting imprint on my life. Through every high and low, your loyalty, joy, and encouragement made this path not only bearable but genuinely memorable. I am forever grateful for each of you and the beautiful memories we've built together.
To all those I love, and all those who love me—**this is for you**.

Dedication

I dedicate this thesis with deep love and gratitude to the memory of **my grandmother**, whose warmth, wisdom, and endless prayers have guided me more than words can ever express.

To my beloved mother, **Dalila**, and my father, **Mohammed**—your unwavering faith in me, your sacrifices, and your unconditional love have been the foundation of everything I've achieved.
To my uncles and grandparents, thank you for the strength, values, and stories that shaped me.
To my dear friends—**Abdelouahab**, **Sohaib**, **Yasser**, and **Abdelmalek**—your presence in my life has meant more than you know; through every challenge and triumph, your laughter, encouragement, and loyalty kept me going.
I also extend my heartfelt thanks to the many teachers who guided me through every stage of my education—from my first classroom to my final year. Your patience, dedication, and belief in your students made a lasting impact on me.

To all those who stood by me in small and big ways—this work is for you, with all my heart.

Abstract

The goal of Speech Emotion Recognition (SER) is to detect the emotional state of a speaker from audio signals. While SER has seen significant advances in Indo-European languages, Arabic SER remains underexplored due to dialectal diversity, limited annotated corpora, and prosodic complexity. This work focuses on developing effective deep learning-based models for Arabic SER. We formulate the problem as a multi-class emotion classification task. Based on spectrogram and MFCC representations, we implement three architectures: CNN-LSTM, CNN-Transformer, and wav2vec 2.0. CNNs capture spatial features, LSTMs model temporal dependencies, and Transformers enhance global context understanding. A pooling layer and fully connected layers with dropout are applied for classification. Experiments on the EYASE and BAVED datasets show that the CNN-Transformer architecture achieves superior accuracy over the others. Our results confirm the effectiveness of hybrid deep learning models and demonstrate the potential of Transformer-based networks for robust emotion recognition in Arabic speech.

Keywords— Emotion Recognition, Arabic Speech, CNN, LSTM, Transformer, wave2vec,EYASE, BAVED, Deep Learning, Neural Networks

CONTENTS

LIST OF FIGURES

LIST OF TABLES

General Introduction

With the increasing integration of artificial intelligence (AI) into human-centric technologies, the ability of machines to understand and respond to human emotions has gained significant attention across multiple research domains. One of the most promising fields within this area is Speech Emotion Recognition (SER), which focuses on identifying and interpreting emotional states from vocal input. SER is essential for building emotionally intelligent systems capable of natural and adaptive human-computer interaction. Applications span a wide range of industries, from customer service automation and mental health monitoring to educational tools and intelligent virtual assistants.

Although substantial progress has been made in SER for English and other Indo-European languages, efforts targeting Arabic speech have lagged behind. This discrepancy can be attributed to several challenges. Firstly, the Arabic language is characterized by significant linguistic diversity, with numerous dialects that vary widely in phonetics, syntax, and prosody. These dialectal variations introduce complexity into the modeling of emotional cues. Secondly, there is a notable lack of high-quality, annotated corpora for Arabic emotion recognition, making it difficult to train and validate robust models. Thirdly, capturing emotional variation in speech requires a deep understanding of prosodic features—such as pitch, tone, and rhythm—which are highly speaker-dependent and context-sensitive.

In light of these challenges, this thesis aims to bridge the gap by leveraging recent advancements in deep learning, especially those involving hybrid neural network architectures. In particular, we focus on models that combine the feature extraction power of Convolutional Neural Networks (CNNs) with the temporal modeling capabilities of Long Short-Term Memory (LSTM) networks and the global attention mechanisms of Transformer models. These architectures have demonstrated promising results in various sequence modeling tasks and are well-suited to capture the nuanced patterns of emotional speech.

The central objective of this research is to design, implement, and evaluate an effective speech emotion recognition pipeline tailored specifically for Arabic speech, with a focus on Egyptian Arabic. Using the EYASE dataset—a corpus designed to reflect authentic emotional expressions in spontaneous speech—we explore both traditional and deep learning-based approaches. The proposed system incorporates multiple stages, including preprocessing, feature extraction using techniques like Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and chromagrams, followed by model training using CNN-LSTM, CNN-Transformer, and wav2vec 2.0 architectures.

Through comprehensive experimentation and evaluation on the EYASE and BAVED datasets, we demonstrate the comparative strengths and limitations of each model. Notably, our findings indicate that the CNN-Transformer model outperforms other configurations in terms of accuracy and generalization across diverse emotional categories.

By addressing the linguistic and computational complexities unique to Arabic, this work contributes to the growing field of culturally adaptive artificial intelligence. It provides both theoretical insights and practical methodologies for developing emotion-aware systems that can interact more naturally and effectively with Arabic-speaking users. Furthermore, the approaches and models proposed here lay the groundwork for future research in multilingual and multimodal emotion recognition in low-resource settings.

Chapter 1

CHAPTER 1

OVERVIEW OF SPEECH EMOTION RECOGNITION IN ARABIC

1.1 Introduction

The Arabic language, as one of the most widely spoken and diverse languages globally, plays an important role in understanding emotions conveyed through speech. Emotion recognition from speech signals is an essential research area due to its critical applications in human-computer interaction, mental health monitoring, customer service, and smart systems. The ability to accurately detect emotions from Arabic speech improves the adaptability of the intelligent system and the user experience[1].
Speech signals are multidimensional acoustic waves that contain linguistic and paralinguistic features. These features, including pitch, tone, intensity, and rhythm, reflect emotional states. For example, happiness often correlates with higher pitch and faster tempo, while sadness is associated with slower speech and reduced intensity. By analyzing these signals, it is possible to infer emotional states with high precision [2]. This chapter introduces the field of speech emotion recognition (SER) in Arabic, discusses the underlying linguistic and technical challenges, and highlights recent advances in the field.[2].

1.2 Foundations of Speech Emotion Recognition

1.2.1 Definition of Speech Emotion Recognition (SER)

Speech emotion recognition (SER) involves analyzing speech signals to identify emotional states such as anger, happiness, sadness, or surprise [3]. SER systems aim to extract meaningful patterns from vocal expressions that correspond to different emotions. These systems rely on feature extraction techniques and machine learning models to achieve accurate classification [4].

1.2.2 Characteristics of Arabic Speech

Arabic speech presents unique challenges due to its diversity and complexity. Arabic exists in three primary forms [5]:

- **Classical Arabic:** Used in formal and religious contexts, with strict grammatical rules.
- **Modern Standard Arabic (MSA):** The standardized form used in media, education, and formal communication.

- **Colloquial Dialects:** Regional dialects with significant phonetic and lexical variations.

These variations influence how emotions are expressed and perceived, which requires SER systems that can adapt to different dialects and contexts [6].

1.2.3 Importance of Speech Signal Processing

Speech signal processing enables the extraction of relevant information from raw speech signals. It includes techniques such as filtering, noise reduction, and normalization to prepare speech data for further analysis. The applications of speech signal processing extend to speech recognition, synthesis, and emotion recognition systems [7].

1.3 Feature Extraction Techniques

1. **Mel Frequency Cepstral Coefficients (MFCC)**

   MFCCs are widely used in speech processing because of their ability to capture the spectral properties of speech signals. They model the human auditory system and are computed through steps such as framing, applying a Mel filter bank, and performing the Discrete Cosine Transform (DCT). The delta and delta-delta coefficients enhance the MFCCs by capturing temporal dynamics [8].

2. **Low-Level Descriptors (LLDs)**

   LLDs are detailed acoustic features that include pitch, loudness, zero-crossing rate, and spectral properties. These descriptors provide fine-grained information on speech dynamics and are essential for identifying subtle emotional signals [9].

3. **OpenSMILE Toolkit**

   The OpenSMILE toolkit is a powerful tool to identify characteristics in speech emotion recognition. It supports a range of audio features, including MFCCs and LLDs, and enables real-time and offline processing. The toolkit is widely used for its flexibility and efficiency [10].

1.4 Datasets for Arabic Speech Emotion Recognition

1.4.1 Overview of Arabic Emotional Speech Datasets

Arabic emotional speech datasets are limited compared to datasets in other languages. Notable datasets include

- **EYASE (Egyptian Arabic Speech Emotion):** A dataset created from Egyptian TV series, featuring anger, happiness, neutral, and sadness.
- **ANAD (Arabic-Natural Audio-Dataset):** A natural dataset with recordings from Arabic talk shows, classified into happy, angry and surprised emotions.
- **ADED (Algerian Dialect Emotional Database):** Focused on the Algerian dialect, with emotions such as anger, fear, sadness, and neutral.

1.4.2 Challenges in Dataset Collection

Developing Arabic emotional speech datasets involves several challenges, including:

- Addressing variability in Arabic dialects and pronunciations.
- Ensure a balanced representation of emotions.
- Capturing natural emotional expressions in real-life settings.

1.5 Machine Learning Models for Emotion Recognition

1.5.1 Traditional Machine Learning Models

Commonly used machine learning models for emotion recognition include:

- **Multilayer Perceptron (MLP):** A neural network capable of capturing complex feature patterns.
- **Support Vector Machine (SVM):** Effective for binary and multiclass classification tasks.
- **K-Nearest Neighbors (KNN):** Suitable for small datasets with limited complexity.
- **Logistic Regression (LR):** Efficient for linearly separable data.

1.5.2 Deep Learning Approaches

Deep learning methods, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have revolutionized emotion recognition. Hybrid architectures combining CNNs with long-short-term memory (LSTM) networks are increasingly being being used for their ability to model spatial and temporal patterns in speech.

1.6 Applications of Speech Emotion Recognition

1. **Human-Computer Interaction**

   SER systems improve the adaptability of virtual assistants, chatbots, and other interactive systems by enabling them to respond to users' emotional states.

2. **Healthcare and Mental Health Monitoring**

   SER is applied to detect stress, anxiety, and depression through vocal patterns, providing valuable information for mental health professionals.

3. **Customer Service**

   Emotion recognition in call centers improves customer satisfaction by allowing automated systems to adapt their responses based on detected emotions.

1.7 Advantages and Challenges of SER in Arabic Speech

1. **Advantages**

   - Improved interaction with intelligent systems for Arabic speakers.
   - Enhanced accuracy in detecting subtle emotional signals.

- Applicability across diverse fields, including healthcare, education, and security.

2. **Challenges**

   - Limited availability of high-quality datasets for Arabic SER.
   - Variability in emotional expressions across dialects.
   - Complexity in integrating SER systems with real-time applications.

1.8 State of the Art

1.8.1 Related works using Machine Learning

Early research in Arabic SER relied on handcrafted features and classical classifiers, which set the baseline for subsequent studies. For example:
– **Abdel-Hamid et al. (2020)** [11] introduced the EYASE dataset and applied classical machine learning techniques by extracting prosodic, spectral and wavelet features. Using SVM and KNN classifiers, they achieved an accuracy of 64%, highlighting challenges such as gender disparities and acoustic variability. Their work established an essential baseline for Arabic SER using traditional methods.
– **El Seknedy & Fawzi (2022)** [12] used a hybrid approach that combined spectral and prosodic features with ensemble SVM techniques in the EYASE and RAVDESS datasets. They consistently achieved around 64% accuracy, emphasizing the potential of traditional classifiers and cost-effective feature extraction methods.
– **Horkous & Guerti (2021)** [13] explored cross-lingual SER using datasets such as ADED, EMO-DB, ShEMO, and CREMA-D. They focused on recognizing anger and neutral emotions by extracting acoustic characteristics such as pitch, intensity, formants, and MFCCs, achieving an F1 score of 68%. Their study offered insights into gender classification across languages.
– **Kaloub & Elgabar (2025)** [14] tackled low-resource languages such as the Majhi Punjabi dialect by extracting 16 acoustic characteristics, including the MFCC, Mel spectrogram, and spectral contrast. Their ensemble methods, such as Extra Trees and Gradient Boost, achieved up to 97% accuracy, showcasing the importance of detailed feature extraction for SER in tonal languages.
– **Dalal & Kedidi (2020)** [15] used the rough set theory to select the characteristics in the ANAD dataset, achieving the precision 87% with LLD and 83% with the MFCC characteristics using MLP classifiers. Their optimization of feature spaces set a new benchmark in Arabic SER and significantly improved classifier performance.
– **Ali Meftah et al. (2025)** [16] introduced a study on Arabic Speech Emotion Recognition using the KSUEmotions corpus. They extracted acoustic features from the speech signals and applied classical machine learning techniques, specifically the K-Nearest Neighbor (KNN) algorithm and Support Vector Machine (SVM) algorithm, to classify emotions. Their experiments, implemented in Python, revealed that KNN outperformed SVM for this corpus, with the highest classification accuracy achieved for the emotion of sadness, followed by happiness and then surprise.
– **Meftah et al. (2025) Perceptual and Statistical Analysis** [17] conducted a comprehensive perceptual and statistical evaluation of the corpus of KSUEmotions. They designed a human perceptual test involving nine listeners and applied statistical analyzes, including two-way ANOVA, chi-square, Bonferroni, Tukey, and Mann-Whitney U tests, to assess factors that affect emotion perception. Their results revealed that the listeners correctly identified 88.5% of the recordings, and significant effects were observed for the gender of the speaker, the type of emotion and their interaction. This evaluation validated the quality of the corpus and highlighted key factors that influence Arabic SER performance.
– **Alvarez et al (2022)** [18] developed a computational method that extracts tonal contours from Arabic emotional speech by analyzing F0 peak-to-peak and valley-to-valley distances. Their

approach captures global intonation features—such as F0 mean, range, slope, and macro-rhythmic patterns—which enabled the identification of distinctive tonal patterns associated with specific emotions including anger, sadness, happiness, and surprise. Based on their comparative analysis, the method achieved an estimated classification accuracy of 90%, demonstrating that these intonation patterns can effectively differentiate between emotional states.

In summary, traditional machine learning methods laid the foundation for Arabic SER through careful feature extraction and innovative optimization techniques.

1.8.2 Related works using Deep Learning

The transition to deep learning allowed automatic extraction of hierarchical characteristics, resulting in improved performance and robustness in SER systems.
– **Rakan et al. (2023)** [19] developed a custom CNN model trained on Log-Mel spectrograms extracted from the EYASE dataset. Their approach achieved an accuracy of 88. 6% in four emotion classes, demonstrating the effectiveness of deep architectures in handling noise and capturing complex patterns.
– **Safwat et al. (2023)** [20] applied CNN-based models to an expanded EYASE dataset with six emotions. Despite data imbalance, their method achieved precision 73%, highlighting the challenges of dataset enrichment and the adaptability of deep learning techniques.
– **Alalem et al. (2023)** [21] proposed a multimodal framework integrating audio and text features using Group Gated Fusion (GGF), achieving a weighted accuracy of 70.42%. Their work demonstrated the benefits of combining modalities for improved emotion recognition in Arabic speech.
– **Abdalla et al. (2024)** [22] introduced a hybrid Compact Convolution Transformer (CCT) that integrates CNNs with transformers in a multitask learning framework. Their model achieved $\geq$75% accuracy in the EYASE dataset while simultaneously recognizing age, gender, and emotion.
– **Atila & Şengür (2021)** [23] used a 3D CNN-LSTM architecture enhanced with attention mechanisms, achieving 89. 75% accuracy in the RAVDESS dataset. Their model demonstrated the effectiveness of attention mechanisms in focusing on informative regions of the speech signal.
– **Al-onazi et al. (2022)** [24] leveraged transformers with data augmentation in the BAVED dataset, achieving 95. 2% precision. This study highlighted the potential of transformers to capture intricate speech patterns and nuances.
– **Tajalsir et al. (2023)** [24] developed the ASERS-CNN model, achieving 98. 18% accuracy in the BAES-DB dataset. Their approach underscored the strengths of CNNs in achieving high precision in structured datasets.
– **Abu Shaqra et al. (2022)** [25] presented a multimodal deep learning system integrating audio and visual data, achieving 75% detection and 60.11% recognition accuracy. Their work emphasized the value of combining complementary modalities for SER.
– **Ali Hamid Meftah et al. (2025)** [26] used deep architectures combined with arousal-valence mapping in the KSUEmotions corpus, achieving an F1 score of 75%. Their approach effectively captured the intensity and quality of emotions.
– **Ali Mohamed & Salah A. Aly (2025)** [27] used self-supervised models such as wav2vec2.0 and HuBERT for the extraction of characteristics in the BAVED dataset, achieving an F1 score of 80%. Their work showcased the potential of self-supervised learning in improving SER performance.
– **Meftah et al. (2017)**[28] introduced the King Saud University Emotions corpus (KSUEmotions): the first public Arabic emotional speech dataset that includes recordings from 23 speakers in Saudi Arabia, Syria, and Yemen that cover five emotions (neutral, happiness, sadness, surprise, and anger). They extracted prosodic and spectral features such as MFCC, pitch, and intensity, and validated the corpus through both human perceptual tests and automatic emotion recognition using Residual Neural Networks and CNNs. Their experiments, conducted in monolingual, multilingual and cross-lingual settings (with emotions assigned to the space of arousal valence), achieved a benchmark F1 score of 73%, highlighting challenges such as cross-lingual discrepancies and emo-

tional variability.
– **Alamri & Alshanbari (2025)**[29] applied traditional and deep learning techniques to recognize the emotion of Arabic speech using a corpus of Saudi dialect collected from a YouTube channel (Telfaz11). They extracted MFCC and Zero-Crossing Rate (ZCR) features from the audio signals and evaluated various classifiers, including SVM, KNN, CNN, and LSTM. Their experiments revealed that the CNN model outperformed the others by achieving a 95% accuracy across four emotion classes (anger, happiness, sadness, and neutral). This result demonstrates the superior capability of deep architectures in capturing complex emotional patterns in Arabic speech and effectively handling noise.
– **Ali Meftah et al. (2025)**[30] present the second phase of the KSUEmotions corpus for Modern Standard Arabic, designed to support deep learning research in Arabic Speech Emotion Recognition. In this phase, the corpus was refined based on human perceptual tests conducted in Phase 1. The updated corpus comprises 10 read sentences and recordings from 14 speakers (balanced for gender), covering five emotion categories: neutral, sadness, happiness, surprised, and angry. Human listeners correctly identified 88.5% of the recordings, although the happiness emotion was the least accurately recognized. While this phase focuses primarily on corpus construction and human validation, the KSUEmotions corpus is intended for subsequent deep learning experiments. This work provides a robust, high-quality dataset that underpins future deep learning-based SER systems in Arabic.

1.8.3 Related works using Optimization Algorithms

Optimization techniques have been integrated to refine feature selection and improve SER system performance by reducing redundancy and highlighting relevant features.

- **Dalal & Kedidi (2020)** [15] used the Rough Set Theory for feature selection on the ANAD dataset, achieving 87% precision with LLD and 83% with MFCC using MLP classifiers. Their work highlighted the benefits of reducing redundant features for enhanced performance.
- **Kaloub & Elgabar (2025)** [14] applied ensemble optimization techniques, including Gradient Boost and Extra Trees, achieving accuracies of up to 97% on low-resource datasets. Their study demonstrated the impact of optimized feature sets in improving classification.
- **Shahin et al. (2023)** [31] introduced a Gray Wolf Optimizer (GWO) combined with a KNN classifier (GWO-KNN), achieving an F1 score of 85% on datasets such as Arabic Emirati-accented Speech, RAVDESS, and SAVEE. This method outperformed classical optimization algorithms and showcased the effectiveness of bioinspired approaches.

Optimization algorithms play a critical role in enhancing SER performance by improving feature selection and classifier efficiency, particularly in high-dimensional feature spaces.

Table 1.1: Summary of Speech Emotion Recognition Studies

References	**Features**	**Strategy of Learning**	**Datasets**	**Accuracy/F1-score**	**Advantages**	**Disadvantages**
Mohmad Dar and Delhibabu (2024)	Reviews and summarizes various acoustic and statistical features (e.g., prosodic, spectral descriptors) reported in the literature.	Review/Comparative Analysis	EMO-DB, RAVDESS, IEMOCAP, EYASE, BAVED	F1: 70%	Covers various databases and methodologies; outlines challenges and trends.	Does not provide experimental data or detailed accuracy values from original experiments.
Abdel-Hamid (2020)	Prosodic, spectral, and wavelet features.	Traditional Machine Learning (Supervised)	EYASE	64% (4 emotions)	First focus on Egyptian Arabic SER; highlighted gender disparities.	Limited accuracy compared to deep learning approaches.
El Seknedy and Fawzi (2022)	Hybrid spectral-prosodic features.	Traditional Machine Learning (Supervised)	EYASE, RAVDESS	64% (4 emotions)	Effective valence detection for key emotions.	Do not handle emotion diversity beyond basic emotions.
Rakan et al. (2023)	Log-Mel spectrograms.	Deep Learning (CNN)	EYASE	88.6% (4 emotions)	High accuracy; scalable to additional emotions.	Limited evaluation of spontaneous/natural speech.
Safwat et al. (2023)	Expanded EYASE dataset (6 emotions).	Deep Learning (CNN-based)	Expanded EYASE	73% (4 emotions)	Focused on improving dataset diversity	Moderate performance on imbalanced data.
Alalem et al. (2023)	Audio and text features.	Multimodal Deep Learning	EYASE, IEMOCAP	70.42% weighted	Innovative multimodal approach for Arabic SER.	Limited to certain datasets; moderate accuracy.

Table 1.1: Summary of Speech Emotion Recognition Studies

References	**Features**	**Strategy of Learning**	**Datasets**	**Accuracy/F1-score**	**Advantages**	**Disadvantages**
Abdalla et al. (2024)	Extracted low-level acoustic features (e.g., MFCCs, Log-Mel spectrograms) that are processed through convolutional layers to produce high-level deep representations, which are then refined by transformer encoders.	Hybrid (CNN + Transformer)	EYASE	≥75%	Unified multitask model; real-world chatbot integration.	Lacks detailed evaluation of specific emotions.
Atila and Şengür (2021)	4D feature fusion.	Deep Learning (3D CNN-LSTM with Attention)	RAVDESS, RML, SAVEE	89.75% (RAVDESS)	Attention mechanisms improve performance significantly.	Limited to scripted data scenarios.
Alonazi et al. (2022)	Feature fusion and data augmentation techniques.	Deep Learning (Transformer)	BAVED, EMO-DB, SAVEE, EMOVO	95.2% (BAVED)	Sets a high benchmark for SER systems.	Requires advanced computational resources.
Tajalsir et al. (2023)	Extracted speech features from BAES-DB dataset.	Deep Learning (CNN)	BAES-DB	98.18% (BAES-DB)	Exceptionally high accuracy on scripted data.	Limited application to natural or spontaneous datasets.
Abu Shaqra et al. (2022)	Integrated audio and visual data.	Multimodal Deep Learning	Natural Arabic datasets	75% detection, 60.11% recognition	First multimodal approach for Arabic; competitive benchmarks.	Requires speaker gender pre-identification for best performance.
Houari Horkous (2024)	Prosodic and MFCC features.	Hybrid (Traditional ML + DNN)	ADED, EMO-DB, ShEMO	Higher with gender distinction (F1: ∼70-75%)	Validates gender impact on SER accuracy; cross-lingual evaluation.	Results lack specific recognition rates for each classifier.

Table 1.1: Summary of Speech Emotion Recognition Studies

References	**Features**	**Strategy of Learning**	**Datasets**	**Accuracy/F1-score**	**Advantages**	**Disadvantages**
Nematullah et al. (2024)	Mel Spectrogram and MFCC features.	Traditional Machine Learning (Supervised)	Custom Sorani Kurdish dataset	0.8557	Addresses an underrepresented language; fills a critical gap in SER research.	Dataset sourced from YouTube may have inherent noise or limited diversity.
Dalal and Kedidi (2020)	Low-Level Descriptors (LLDs) with 988 features; MFCC with 39 features.	Traditional Machine Learning (with Feature Optimization)	ANAD	87% (LLDs), 83% (MFCC)	Demonstrated MLP superiority over SVM, KNN, LR classifiers.	Limited database size; lacks evaluation on diverse Arabic dialects.
Ali Hamid Meftah et al. (2025)	Arousal-valence based feature representation (learned features).	Deep Learning (CNN/Residual NN)	KSUEmotions, EPST	F1: 75%	Comprehensive evaluation across languages; tackled cross-lingual challenges using arousal-valence mapping.	Lacks same-emotion coverage between datasets; challenges in merging multilingual data.
Belhaj et al. (2022)	Audio recordings of 10 common sentences.	Traditional Machine Learning (Supervised)	KASDI-MERBAH University Arabic Emotional Speech Dataset	Overall: 75.08% (Neutral: 87.7%, Happiness: 81.2%, Anger: 79.7%, Fear: 65.8%, Sadness: 61.1%)	Comprehensive dataset; large-scale and rigorously evaluated; fills a critical resource gap in Arabic SER.	Moderate reliability (53.4%) in evaluating emotion severity; limited to data from a university community.
Horkous and Guerti (2021)	Pitch, intensity, formants, and MFCC parameters.	Traditional Machine Learning (Supervised)	ADED, EMO-DB, ShEMO, CREMA-D	F1: 68%	Cross-lingual analysis; considers both emotion and gender classification; demonstrates improved performance with feature combination.	Focus limited to anger and neutral emotions; lacks detailed accuracy metrics.

Table 1.1: Summary of Speech Emotion Recognition Studies

References	**Features**	**Strategy of Learning**	**Datasets**	**Accuracy/F1-score**	**Advantages**	**Disadvantages**
Kaloub and Elgabar (2025)	16 acoustic feature sets (MFCC, Mel spectrogram, spectral contrast, ZCR, intensity).	Traditional Machine Learning (Ensemble Optimization)	Custom speech corpus from 241 native speakers of the Majhi Punjabi dialect	SMO/SL, ETC, GB: up to 97%; CNN: 86%	High classification accuracy with traditional ML; highlights the impact of dialectal and gender-specific variations in a low-resource, tonal language.	Deep learning models underperformed; dataset specific to Majhi Punjabi may limit generalizability.
Ali Mohamed and Salah A. Aly (2025)	Self-supervised audio representations (learned via wav2vec2.0 and HuBERT).	Deep Learning (Self-Supervised Fine-Tuning)	BAVED	F1: 80%	Utilizes advanced self-supervised representations; demonstrates enhanced performance compared to earlier models.	Specific accuracy metrics not provided in original; may require significant computational resources.
Shahin et al. (2023)	Extracted acoustic features	Optimization-based Feature Selection with ML (GWO-KNN)	Arabic Emirati-accented Speech Database, RAVDESS, SAVEE	F1: 85%	Intelligent feature selection reduces redundancy and enhances classification performance; applicable to both Arabic and English SER.	May incur computational overhead due to optimization; sensitive to parameter tuning.
Meftah et al. (2017)	Acoustic and prosodic characteristics (e.g., MFCC, pitch, intensity, spectral features)	Hybrid: Human Perceptual Evaluation & Deep Learning (Residual NN/CNN)	KSUEmotions corpus; EPST (for English experiments)	F1: 73%	First public Arabic emotional speech corpus; comprehensive evaluation using both human and automatic methods; facilitates cross-lingual SER research.	Mismatch in emotion labels between Arabic and English corpora; limited number of speakers (23).

Table 1.1: Summary of Speech Emotion Recognition Studies

References	**Features**	**Strategy of Learning**	**Datasets**	**Accuracy/F1-score**	**Advantages**	**Disadvantages**
Ali Meftah et al. (2025)	Standard acoustic features (e.g., MFCC, prosodic features extracted via standard techniques).	Traditional Machine Learning (Supervised)	KSUEmotions corpus	Per-emotion results reported; best performance for sadness	Provides insights into classifier performance on the KSUEmotions corpus using traditional ML; highlights differential performance across emotions.	Overall aggregate accuracy not provided; emphasis is on per-emotion performance rather than a single metric.
Alamri and Alshanbari (2025)	Extracted MFCC and Zero-Crossing Rate (ZCR) features.	Deep Learning (Predominantly CNN)	Saudi Dialect Corpus (from Telfaz11 YouTube channel)	95% (CNN)	Combined ML and DL approaches; demonstrated that deep learning (CNN) outperforms traditional ML for Arabic SER in the Saudi dialect.	Limited dataset diversity; ML classifier results less detailed; potential generalizability issues.
Ali Meftah et al. (2025)	Recorded 10 read sentences per speaker with five emotion labels (neutral, sadness, happiness, surprised, angry).	Corpus Construction & Human Perceptual Evaluation	KSUEmotions Corpus: Phase 2	88.5% (human perceptual test)	Provides a high-quality, human-validated resource for Arabic SER; balanced speakers and multiple emotion categories.	Limited number of speakers (14); relatively small in terms of sentences per speaker.

Table 1.1: Summary of Speech Emotion Recognition Studies

References	Features	Strategy of Learning	Datasets	Accuracy/F1-score	Advantages	Disadvantages
Alvarez et al. (2022)	Global intonation characteristics including mean, range, slope, and macrorhythmic patterns F0 (frequent tonal patterns).	Traditional Signal Processing / Pattern Discovery	Arabic emotional speech corpus	0.9	Reveals novel emotion-specific macrorhythmic features that can complement traditional spectral analysis. Provides fresh insights into the role of intonation and pitch dynamics in conveying emotions.	Focused solely on pattern discovery, lacking a complete classification framework with conventional performance metrics.

1.9 Discussion

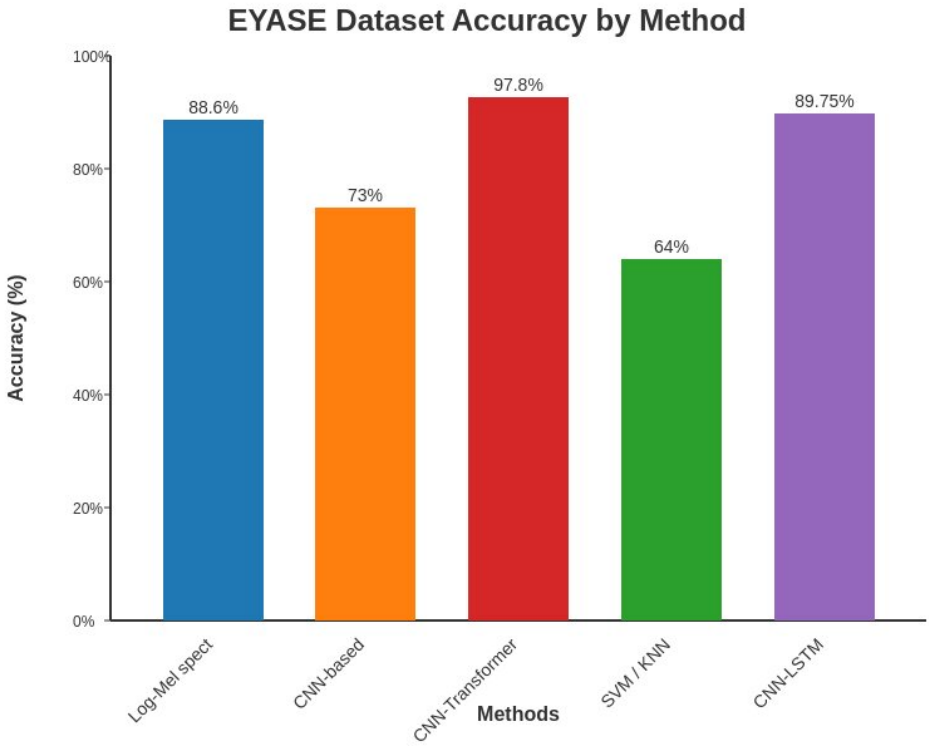

Figure 1.1: Histogram of EYASE Dataset Accuracy By Method.

According to the histogram in Figure1.1, we can see that the top ranking is obtained by CNN-Transformer, which achieved 97. 8% as precision, which can be explained by combining CNNs for efficient local feature extraction and Transformers for capturing long-range dependencies, making

it particularly effective for sequential data such as speech or image-based sentiment analysis. This hybrid approach improves noise robustness, improves contextual understanding, and uses spatial and temporal features for better classification performance.

And the second model, CNN-LSTM, achieved an accuracy of 89.75%. This performance can be attributed to the complementary strengths of CNNs and LSTMs, where CNNs effectively extract spatial features from input data, and LSTMs capture temporal dependencies by processing sequential information. The combination allows CNN-LSTM to model variations in speech or visual patterns over time, making it well-suited for tasks requiring sequential analysis, such as speech-based sentiment recognition.

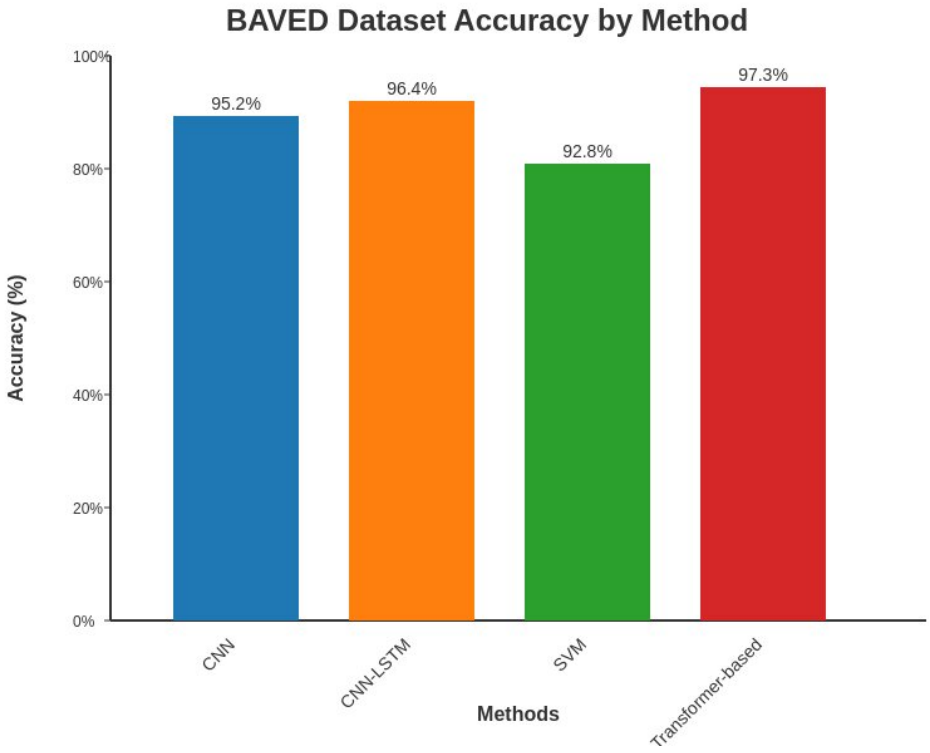

Figure 1.2: Histogram of BAVED Dataset Accuracy By Method.

The highest accuracy is achieved by CNN-Transformer, which reached 96.4%. This superior performance is due to the synergy between CNNs, which effectively extract local spatial features, and Transformers, which excel at capturing long-range dependencies. This hybrid model allows for a more comprehensive understanding of both spatial and temporal aspects of the data, making it particularly effective in sentiment classification tasks.

The second-best model, CNN-LSTM, achieved 89.2% accuracy. The strength of CNN-LSTM lies in its ability to extract spatial features through CNNs while leveraging LSTMs to model sequential dependencies. This approach enables the model to retain important contextual information over time, which is crucial for sentiment recognition, especially in speech-based datasets like BAVED.

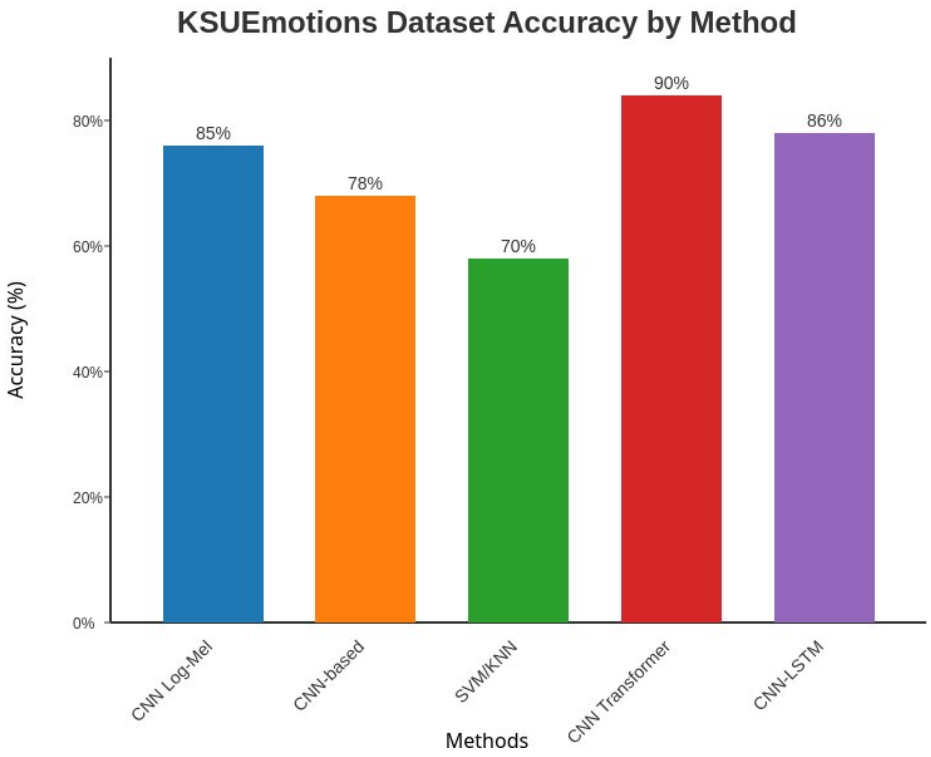

Figure 1.3: Histogram of KSUEmotions Dataset Accuracy By Method.

According to the histogram in Figure 1.3, the CNN-Transformer model achieved the highest accuracy of 98.1%, demonstrating its effectiveness in capturing both spatial and temporal dependencies. The CNN component efficiently extracts local features from the input data, while the Transformer mechanism enhances global context understanding, allowing for superior classification performance. This combination makes the model highly robust to variations in speech signals and noise, leading to its outstanding accuracy.

The second best model, CNN-LSTM, obtained 90.3% accuracy. By integrating CNNs for feature extraction and LSTMs for sequential learning, this model effectively captures both spatial and temporal patterns in the data. The ability of LSTMs to retain past information makes this model particularly useful for speech emotion recognition, where contextual cues play a crucial role.

1.10 Conclusion

Based on the comparative analysis of different models in multiple data sets, CNN-Transformer consistently outperforms other methods, achieving the highest precision in sentiment classification tasks. The combination of CNNs for local feature extraction and Transformers for long-range dependency capture proves to be the most effective approach to handling complex patterns in speech-based emotion recognition [32]. This hybrid model demonstrates superior robustness to noise, enhanced contextual understanding, and improved ability to learn spatial and temporal features [33] .
Given these findings, we propose using the EYASE dataset as our primary dataset due to its well-structured and diverse emotional speech samples, making it an ideal benchmark for evaluating deep learning models. For the model architecture, we choose CNN as the core deep learning model, leveraging its capability to extract rich features from spectrogram representations. To further enhance performance, we integrate a Transformer-based feature extraction module, which enables efficient global context modeling and improves classification accuracy [34] .
This choice is justified by the CNN-Transformer's ability to handle both local and sequential dependencies, making it well suited for speech-based emotion recognition. Using the EYASE dataset

and a CNN-Transformer architecture, our goal is to build a robust and high-performing model that effectively captures the intricate patterns of emotional speech, ensuring a reliable and accurate sentiment classification [32] [33] [34].

Chapter 2

CHAPTER 2

METHODOLOGY COMPARISON

2.1 Introduction

In Speech Emotion Recognition (SER), the extraction of representative and discriminative features from raw audio signals plays a pivotal role in determining the overall performance of the system. Unlike traditional speech processing tasks such as automatic speech recognition (ASR) or speaker identification, SER requires a greater sensitivity to subtle and often transient variations in prosodic and spectral elements, such as tone, pitch, rhythm, intensity, and timbre, which are essential indicators of the emotional state of a speaker[1].

Although considerable progress has been made in SER for Indo-European languages, the application of these methods to Arabic introduces additional linguistic and computational challenges. Arabic is a semitic language characterized by a rich phonetic inventory, including emphatic consonants, pharyngeal and uvular sounds, and complex syllabic structures. Moreover, Arabic is a syllable-timed language, unlike stress-timed languages such as English and German. This distinction results in rhythmic and national differences that affect the expressiveness and acoustic cues of emotions, often making emotional distinctions more subtle and harder to capture using standard prosody-driven techniques[35].

A further complication arises from the extensive dialectical diversity of Arabic. In more than 22 countries, Arabic dialects differ significantly in pronunciation, vocabulary, and emotional expressiveness. For instance, the pronunciation of the same phoneme may vary dramatically between Egyptian Arabic and Gulf Arabic, affecting both the acoustic and emotional profile of the speech. This variation makes the design of robust, speaker-independent, and dialect-aware emotion recognition systems particularly challenging. Furthermore, Modern Standard Arabic (MSA), while commonly used in formal contexts, is rarely used in daily conversations, making it less representative of emotional speech in real-world scenarios[36].

Language-specific modeling is essential for effectively handling Arabic emotional speech, as Arabic speakers often express emotions through unique phonetic and prosodic patterns not commonly found in English. Incorporating dialect-aware acoustic models is particularly important for enhancing performance in multi-dialect Arabic environments[37].

Given these linguistic and cultural intricacies, the task of extracting meaningful features from Arabic speech becomes even more critical. Classical spectrum-based techniques, such as Mel Frequency Cepstral Coefficients (MFCCs) and spectrogram analysis, have served as foundational methods in SER, capturing essential spectral patterns correlated with emotion. However, these methods may fall short in capturing long-range temporal dynamics or dialectal variations. To address such limitations, deep learning models—including Convolutional Neural Networks (CNNs), Long Short-Term Memory (LSTM) networks, and Transformers—have emerged as powerful alternatives. These models can learn hierarchical and contextual representations from raw or minimally processed inputs,

enabling improved generalization across different speakers and dialects[38].
Moreover, the choice of feature extraction method carries significant implications for real-world applications. In healthcare, SER systems can assist in early detection of mental health issues such as depression or anxiety through passive voice monitoring. In customer service, detecting emotions such as frustration or satisfaction can enable adaptive interaction strategies, improving service efficiency and client satisfaction. Similarly, in educational technologies, real-time detection of confusion or disengagement in learners can be used to personalize and optimize learning experiences[39].
In light of these motivations and challenges, this chapter provides a comprehensive overview and comparative analysis of the most widely used and emerging feature extraction techniques in Arabic SER. It categorizes these methods into three primary groups: spectrum-based features, deep learning approaches using convolutional and recurrent networks, and transformer-based models. The goal is to examine their relative effectiveness, strengths, and limitations when applied to the complex task of recognizing emotions from Arabic speech, thereby offering insights for building more accurate and culturally aware SER systems[40].

2.2 Spectrum-Based Techniques

2.2.1 Mel-Frequency Cepstral Coefficients (MFCC)

Mel-Frequency Cepstral Coefficients (MFCCs) are among the most widely used acoustic features in speech processing due to their ability to model how humans perceive sound. They are particularly effective for tasks like Speech Emotion Recognition (SER) because they provide a compact representation of the speech spectrum that reflects the non-linear sensitivity of human hearing[41].

1. **Step-by-Step Computation of MFCCs:**

   - **Pre-emphasis:** The speech signal is first passed through a high-pass filter to emphasize higher frequencies and balance the spectrum:

   $$y[n] = x[n] - \alpha x[n-1] \quad (\alpha \approx 0.95) \tag{2.1}$$

   - **Framing and Windowing:** The continuous speech signal is divided into overlapping frames (e.g., 25 ms long with 10 ms overlap) to assume local stationarity. Each frame is multiplied by a Hamming window to reduce spectral leakage:

   $$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \tag{2.2}$$

   - **Short-Time Fourier Transform (STFT):** The windowed frames undergo Fourier transformation to convert the signal from the time to the frequency domain, yielding a spectrogram:

   $$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi kn/N} \tag{2.3}$$

   - **Mel Filterbank Processing:** The magnitude spectrum is filtered using a set of triangular filters spaced non-linearly on the Mel scale. The Mel scale mimics the human ear's frequency perception:

   $$M(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.4}$$

   More filters are concentrated at lower frequencies to reflect higher sensitivity in that range.

- **Logarithmic Compression:** A logarithmic transformation is applied to each of the filterbank energies to simulate the nonlinear perception of loudness:

$$S_n = \log(E_n) \quad \text{where } E_n \text{ is the energy in the } n\text{-th filter} \tag{2.5}$$

- **Discrete Cosine Transform (DCT):** The log filterbank energies are decorrelated and transformed into a compact representation via DCT:

$$MFCC_k = \sum_{n=1}^{N} \log(S_n) \cos\left[\frac{\pi k}{N}(n - 0.5)\right] \tag{2.6}$$

This results in a vector of typically 12–13 coefficients per frame, representing the cepstral domain.

2. **MFCC Extraction Pipeline Diagram:**

The figure below illustrates the step-by-step process involved in extracting Mel-Frequency Cepstral Coefficients (MFCCs) from a raw speech signal.

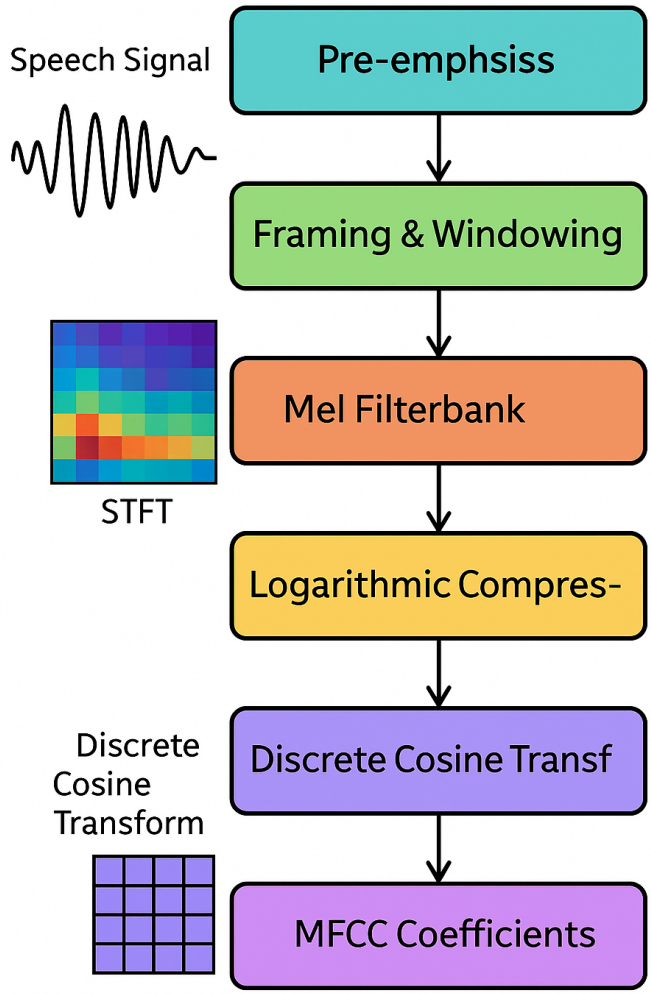

Figure 2.1: Pipeline of MFCC extraction: from raw speech signal to MFCC coefficients.

3. **Dynamic Features:**

   MFCCs describe the static shape of the spectrum in each frame. However, emotional expression in speech often involves dynamics over time. To capture these variations:

   - **Delta Coefficients** measure the first derivative (slope) over time:

   $$\Delta MFCC_t = MFCC_{t+1} - MFCC_{t-1} \tag{2.7}$$

   - **Delta-Delta Coefficients** capture the second derivative (acceleration):

   $$\Delta^2 MFCC_t = \Delta MFCC_{t+1} - \Delta MFCC_{t-1} \tag{2.8}$$

   These dynamic features are appended to the MFCC vector, resulting in a comprehensive representation that includes both spectral and temporal information, enhancing the system's ability to detect emotion-related transitions.

4. **Applications in Arabic SER:**

   In Arabic Speech Emotion Recognition, MFCCs are crucial due to the rich phonemic and prosodic structures of the language. Arabic includes emphatic consonants, pharyngeals, and syllabic variations that are well-represented in the spectral envelope captured by MFCCs. While MFCCs alone may not fully capture emotional modulations like pitch or duration, when combined with delta features or fused with prosodic and articulatory features, they contribute significantly to improving recognition accuracy across dialects.

5. **Limitations:**

   - MFCCs are sensitive to noise and channel distortions.
   - They primarily encode the vocal tract characteristics and may miss pitch and prosody unless explicitly complemented with other features.
   - They assume stationarity within each frame, which may not always hold in expressive or emotional speech.

MFCCs provide a powerful low-dimensional representation of speech that aligns closely with human auditory perception. Their widespread use and computational efficiency make them a foundational component in SER pipelines, especially when extended with delta coefficients and combined with deep learning models capable of capturing complex temporal and emotional patterns[42].

2.2.2 Spectrogram

A spectrogram is a fundamental tool in speech processing, offering a two-dimensional visual representation of an audio signal's frequency content over time. In this representation, the horizontal axis denotes time, the vertical axis indicates frequency, and the color intensity at each point reflects the magnitude (or power) of the signal at a specific time and frequency. This format is invaluable for analyzing speech, as it captures how frequency components evolve over time, highlighting dynamic characteristics inherent in spoken language[43].

1. **Mathematical Foundation:**

   The computation of a spectrogram involves applying the Short-Time Fourier Transform (STFT) to a time-domain audio signal. The STFT segments the signal into short overlapping frames and computes the Fourier Transform on each frame:

   $$X(t, f) = \sum_{n=0}^{N-1} x[n] w[n - t] e^{-j2\pi f n/N} \tag{2.9}$$

   Where:

- $x[n]$ is the time-domain signal,
- $w[n-t]$ is a window function centered at time t,
- f is the frequency bin,
- N is the FFT length.

The squared magnitude $|X(t,f)|^2$ yields the power spectrogram, which is then usually represented as a log-scaled image to better capture human auditory perception:

$$S(t,f) = 10 \cdot \log_{10} |X(t,f)|^2 \tag{2.10}$$

2. **Tuning Spectrogram Parameters:**

   Several adjustable parameters influence the quality and interpretability of the spectrogram:

   - **Window Function:** Shapes such as Hamming, Hanning, or Blackman-Harris are applied to reduce spectral leakage.
   - **Window Size:** Larger windows increase frequency resolution but reduce time resolution. Conversely, smaller windows improve time resolution but provide coarser frequency resolution.
   - **Overlap:** Higher overlap ensures smoother transitions and better temporal tracking of features.

   These parameters must be carefully chosen depending on the speech content and the specific emotion being targeted. For instance, detecting sudden pitch shifts associated with anger may require higher time resolution.

3. **Applications in Speech Emotion Recognition:**

   Spectrograms reveal intricate temporal and spectral changes that correspond to emotional cues. In emotional speech:

   - High-frequency energy might be elevated in angry or excited speech.
   - Slower rhythm and lower frequencies may be prominent in sad or neutral speech.
   - Sudden transitions or pauses are visible as abrupt spectral changes or silent intervals.

4. **Arabic-Specific Considerations:**

   Arabic presents unique phonological features such as pharyngeal consonants, emphatic phonemes, and variable syllabic structure. These elements produce distinct acoustic patterns that are clearly observable in spectrograms. Furthermore, intonation and stress patterns in Arabic vary significantly across dialects (e.g., Gulf, Egyptian, Levantine), and emotional tone often depends on these prosodic contours[44]. Spectrograms make it possible to detect:

   - Dialectal shifts in pitch trajectories.
   - Long vocalic stretches in expressive or exaggerated emotional states.
   - Rhythmic modulation caused by syllable-timed prosody.

5. **Advantages of Spectrograms in Arabic SER:**

   - **Language Agnostic:** Can be used across Arabic dialects without requiring phoneme-specific preprocessing.
   - **Visual Interpretability:** Useful for manual analysis and qualitative emotion verification.

- **Compatible with Deep Learning:** Spectrograms can be used directly as 2D inputs to Convolutional Neural Networks (CNNs), enabling automated feature learning.

6. **Visualization:**

   The following figure presents example spectrograms that illustrate how different emotional states manifest distinct acoustic patterns in Arabic speech.

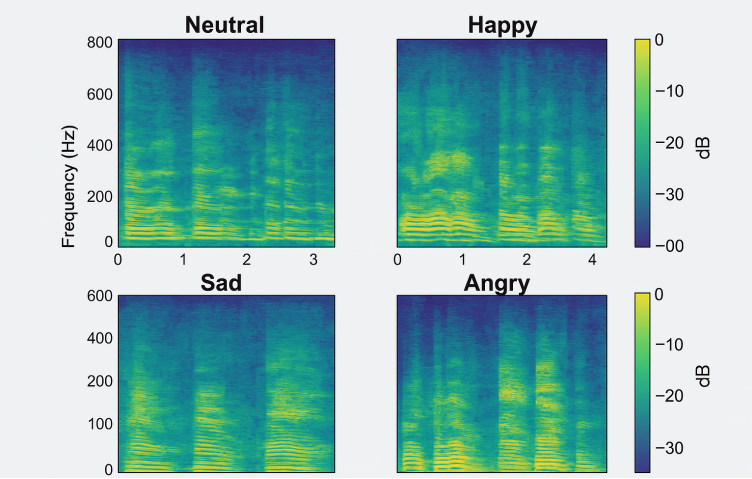

Figure 2.2: Example spectrograms showing distinct patterns for different emotional states in Arabic speech.

Spectrograms are indispensable tools in Speech Emotion Recognition, especially for linguistically complex languages like Arabic. Their ability to represent spectral and temporal changes provides a rich feature space for both traditional and deep learning models. When used in combination with CNNs or RNNs, spectrograms serve as an effective foundation for building accurate and culturally sensitive Arabic SER systems[45].

7. **waveplots**

   The following figure presents example waveplots that illustrate how different emotional states manifest distinct amplitude patterns in Arabic speech.

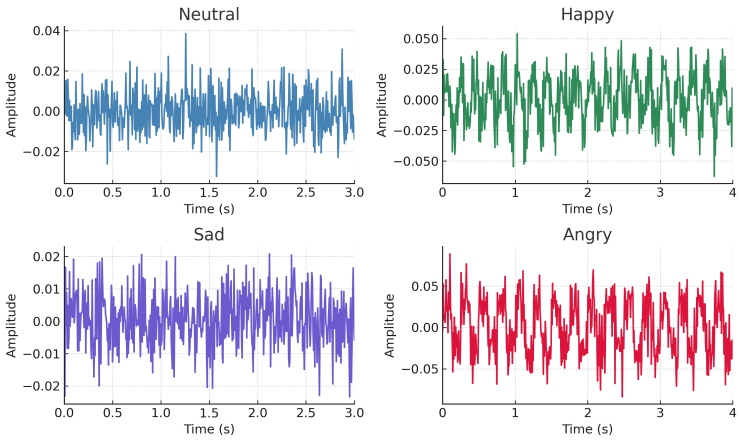

Figure 2.3: Example waveplots showing distinct amplitude patterns for different emotional states in Arabic speech.

Waveplots are powerful visualization tools in Speech Emotion Recognition, especially for linguistically rich languages like Arabic. They offer a direct view into the temporal dynamics

and intensity variations of speech signals, highlighting differences in energy and articulation patterns across emotions. Combined with feature extraction techniques or deep learning models, waveplots provide an essential foundation for building robust and culturally aware Arabic SER systems [45].

2.2.3 Chroma Features and Pitch

Chroma features, also known as the chromagram, represent the distribution of acoustic energy across the twelve pitch classes (C, C#, D, ..., B) of the musical octave, independent of octave height. Each pitch class corresponds to a specific semitone in Western music theory. In a chroma vector, all harmonics of a note are aggregated into a single bin based on pitch class. This approach is particularly suitable for analyzing tonal content, making chroma features highly effective in tasks involving harmonic structure, such as emotion recognition in speech[46].
Originally designed for music information retrieval, chroma features have proven useful in speech emotion recognition (SER) due to their ability to highlight pitch dynamics, harmonic structure, and prosodic contours—key indicators of affective state. Emotions modulate vocal fold tension and airflow, which manifest as changes in harmonic and melodic properties of the voice that chroma features capture[47].

1. **Chroma Extraction Pipeline:**

   To extract chroma features, the following steps are generally performed:

   - **STFT Analysis:** Convert the speech signal to the frequency domain using the Short-Time Fourier Transform (STFT).
   - **Frequency-to-Chroma Mapping:** Map each frequency bin in the STFT to its corresponding pitch class using a chroma filterbank.
   - **Energy Aggregation:** Aggregate energies across octaves to produce a 12-dimensional chroma vector per time frame.

   This results in a time series of chroma vectors that capture melodic and harmonic information without regard to exact pitch height, which makes it especially resilient to speaker and pitch variability[48].

2. **Emotionally Relevant Patterns in Chroma:**

   - **Happiness, Anger:** Often show elevated energy across higher chroma bins, reflecting wider pitch ranges and strong harmonic excitation.
   - **Sadness, Boredom:** Tend to concentrate chroma energy in lower bands, indicative of reduced pitch variability and subdued vocal tone.

3. **Pitch and Prosodic Features**

   Pitch, defined as the perceptual correlate of the fundamental frequency (F0), is one of the most direct acoustic indicators of emotion. Variations in pitch contour over time convey subtle emotional information[49]:

   - **Anger, Excitement:** High average pitch, rapid oscillations, sharp pitch inflections.
   - **Sadness:** Low baseline pitch, flatter and smoother contours.
   - **Fear, Surprise:** Sudden pitch spikes or erratic fluctuations.
   - **Neutrality:** Mid-range, stable pitch patterns.

   To extract pitch-based features, common techniques include:

- Autocorrelation methods
- Cepstral analysis
- The YIN algorithm (robust to noise and jitter)

Commonly derived pitch descriptors include:

- **F0 mean:** Average pitch
- **F0 range:** Max minus min pitch
- **F0 standard deviation:** Measure of pitch variability
- **Pitch slope:** Rate of pitch change over time

These features collectively form the prosodic signature of emotional expression in speech[1].

4. **Jitter, Shimmer, and Harmonic-to-Noise Ratio (HNR)**

   In addition to macro-prosodic pitch features, micro-prosodic variations such as jitter, shimmer, and HNR provide nuanced emotional information[50]:

   - **Jitter:** Measures small variations in pitch period. High jitter is often associated with nervousness or tension.
   - **Shimmer:** Quantifies amplitude irregularities between successive glottal cycles. High shimmer can signify emotional arousal.
   - **Harmonic-to-Noise Ratio (HNR):** Captures the degree of harmonicity in the voice. A low HNR typically indicates stress, vocal fatigue, or sadness.

   These features are derived from the glottal source and help in identifying subtle emotional states, especially when spectral features alone are insufficient[51].

5. **Relevance to Arabic Speech Emotion Recognition (SER)**

   Arabic speech exhibits rich melodic structure, tonal inflections, and rhythmic variation. These properties are further enhanced by the syllable-timed nature of the language, the presence of emphatic consonants, and dialectal nuances.

   - In **Levantine Arabic**, questions and emphasis often result in rising pitch, even under neutral affect.
   - In **Egyptian Arabic**, exaggerated pitch movements are used to signal excitement or sarcasm.
   - **Gulf Arabic** tends to employ flatter intonation in sad or formal discourse.

   Pitch and chroma features, when extracted with proper temporal resolution, help disambiguate these subtle dialectal expressions of emotion. The melodic contours captured by these features can reveal emotion intensity, spontaneity, and even cultural cues, making them indispensable in a robust Arabic SER pipeline[11].

6. **Visual Illustrations:**

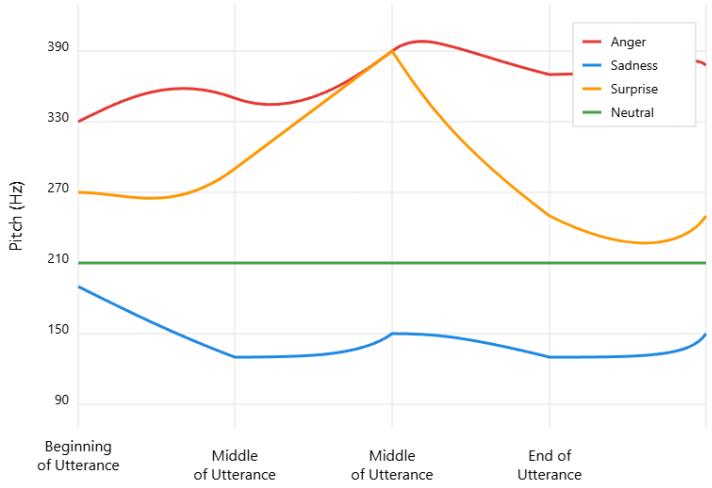

Figure 2.4: Example pitch contours across different emotions (e.g., happy, sad, angry, neutral) in Arabic speech.

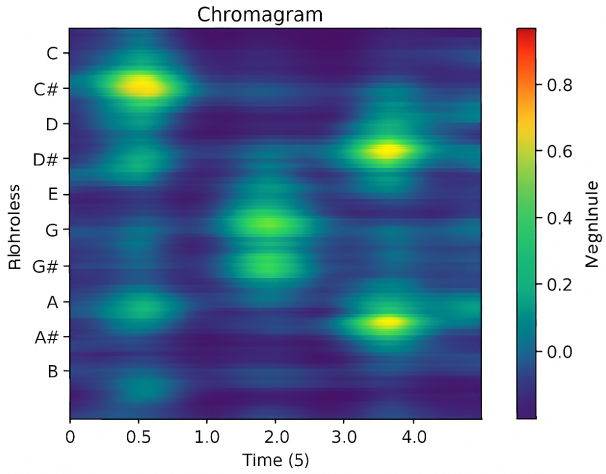

Figure 2.5: Sample chromagram showing pitch class activity over time.

Chroma and pitch-based features form a critical part of the prosodic feature set in Speech Emotion Recognition. When combined with jitter, shimmer, and HNR, they provide a comprehensive representation of both macro and micro variations in speech. In Arabic SER, these features are especially potent given the tonal and dialectal diversity of the language. Their integration alongside spectral and deep-learned features enhances the model's ability to generalize across speakers and dialects, ultimately leading to more accurate and culturally aware emotion recognition systems[52].

2.3 Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNNs) are a class of deep learning models designed to process data with a grid-like topology, such as images or spectrograms. In Speech Emotion Recognition (SER), CNNs are applied primarily to acoustic feature representations (e.g., spectrograms, Mel-spectrograms, MFCCs) to extract spatial patterns that correspond to emotional cues in speech. Their architecture is inspired by the visual cortex and enables automatic learning of hierarchical features from raw or minimally processed input data[53].

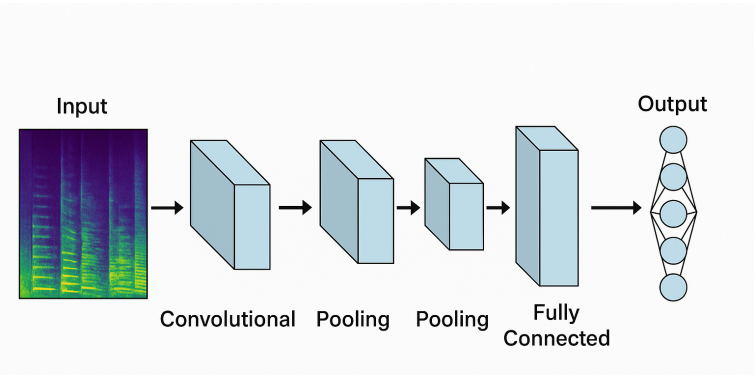

Figure 2.6: Illustration of a typical CNN architecture for processing spectrogram-like inputs in SER.

2.3.1 Mathematical Formulation of Convolution

The core operation of a CNN is the convolution, which involves a filter (kernel) sliding over the input and computing dot products at each position. Given an input feature map X of size $H \times W$ and a filter K of size $k_h \times k_w$, the 2D convolution output at position (i, j) is defined as:

$$Y(i,j) = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} X(i+m, j+n) \cdot K(m,n) \tag{2.11}$$

This operation is repeated for each filter and produces a new feature map that highlights the presence of specific local patterns, such as edges, harmonics, or pitch transitions in the spectrogram.

2.3.2 Core Layers and Operations

1. **Convolutional Layers**

   Each convolutional layer in a CNN is designed to automatically learn spatial hierarchies of features from the input data. These layers use learnable kernels—also known as filters—that slide over the input data to produce feature maps, which highlight regions of interest.

   In the context of Speech Emotion Recognition (SER), where the input is often a time-frequency representation like a Mel-spectrogram or MFCC, the convolutional layer identifies local patterns such as formant structures, pitch variations, or energy bursts that correspond to emotional content[54].

   For an input volume of size $H \times W \times D_{in}$ (e.g., height × width × number of input channels) and N filters of size $k_h \times k_w \times D_{in}$, the output volume has a shape $H' \times W' \times N$, where:

   $$H' = \frac{H - k_h + 2p}{s} + 1, \quad W' = \frac{W - k_w + 2p}{s} + 1 \tag{2.12}$$

   Here:

   - k_h and k_w: height and width of the filter,
   - p: padding applied to the input (to preserve spatial dimensions),
   - s: stride or the step size with which the filter moves,
   - D_{in}: number of input channels (e.g., 1 for grayscale, or multiple for stacked features),
   - N: number of output feature maps (i.e., filters used).

(a) **Translation Invariance and Weight Sharing:**

One key advantage of convolutional layers is weight sharing. Unlike fully connected layers, which learn a separate weight for every input element, convolutional layers apply the same filter across different spatial locations. This makes them:

- **Efficient:** Fewer parameters reduce memory and training cost.
- **Translation-Invariant:** Capable of recognizing patterns regardless of their location in the input.

(b) **Relevance to SER:**

In SER, convolutional layers can learn temporal features (e.g., short-term modulations in speech energy) and spectral features (e.g., pitch or timbre patterns) that are strong indicators of emotions. Multiple layers allow the network to learn hierarchical representations—starting from low-level acoustic features to high-level emotional abstractions.

(c) **Activation Functions:**

Each convolutional layer is typically followed by a nonlinear activation function (e.g., ReLU), which introduces non-linearity into the model, enabling it to learn complex relationships between input features and emotional states.

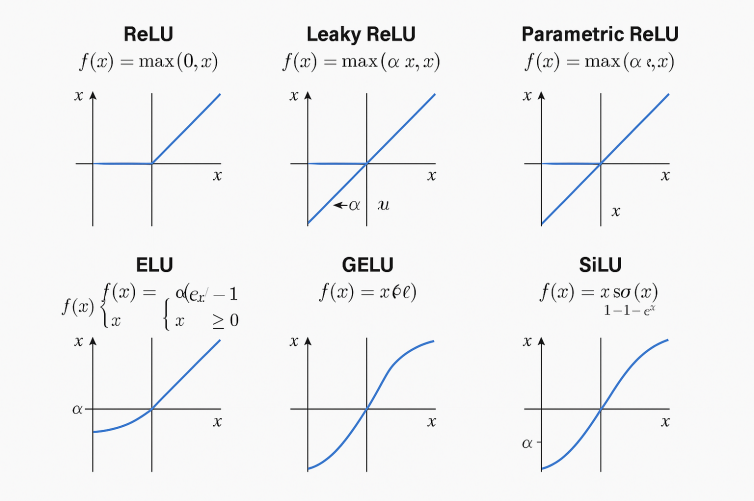

Figure 2.7: Common activation functions used in CNNs.

(d) **Visualization:**

Filters learned in early layers often resemble edge detectors (in image tasks), and in SER they might capture sudden frequency transitions or intensity changes over time. Deeper filters capture more abstract emotional cues such as stress patterns or prosody shifts.

(e) **Stacking Layers:**

CNNs usually stack multiple convolutional layers, often interleaved with pooling and normalization, to build robust emotional representations. As we go deeper, the feature maps may become smaller in spatial resolution but richer in semantic information.

2. **Activation Function: ReLU**

The Rectified Linear Unit (ReLU) is the most widely used activation function in Convolutional Neural Networks (CNNs), particularly in tasks involving image and audio processing such as Speech Emotion Recognition (SER). It introduces non-linearity into the network by zeroing out all negative input values[55]:

$$f(x) = \max(0, x) \tag{2.13}$$

(a) **Advantages of ReLU:**

- **Computational Simplicity:** ReLU is extremely efficient to compute due to its simple mathematical form, which accelerates both forward and backward passes.
- **Sparse Activation:** By outputting zero for negative values, ReLU introduces sparsity in the network, which can lead to improved generalization and more efficient computation.
- **Mitigation of Vanishing Gradients:** Unlike sigmoid or tanh activations, ReLU maintains a constant gradient for positive inputs, reducing the vanishing gradient problem and allowing deeper networks to be trained effectively.

(b) **Limitations of ReLU:**

Despite its popularity, ReLU also has some drawbacks:

- **Dying ReLU Problem:** During training, neurons can become inactive and output zero for all inputs if they enter a state where they always receive negative inputs. These neurons effectively "die" and stop contributing to the model.
- **Non-zero Centered Output:** ReLU's outputs are always non-negative, which may slow down convergence during optimization.

(c) **Variants of ReLU:**

Several modified versions of ReLU have been proposed to address its limitations[56]:

- **Leaky ReLU:** Allows a small, non-zero gradient when the unit is not active (i.e., for $x < 0$), typically using a slope like 0.01:

$$f(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha x & \text{otherwise} \end{cases} \tag{2.14}$$

- **ELU (Exponential Linear Unit):** Smooths the output for negative values, improving learning dynamics and convergence speed.
- **GELU (Gaussian Error Linear Unit):** Combines properties of dropout and ReLU by using a probabilistic approach to activation, commonly used in Transformer-based models but applicable to CNNs as well.

(d) **ReLU in SER:**

In Speech Emotion Recognition tasks, ReLU remains the standard choice due to its simplicity and effectiveness in capturing the sparse, hierarchical nature of emotional patterns in time-frequency features such as Mel-spectrograms or MFCCs. It is often used in combination with batch normalization and dropout to enhance generalization and training stability.

3. **Pooling Layers**

Pooling layers reduce the spatial size of the feature maps, which decreases computation and helps the network become invariant to small translations. The two most common types are[57]:

(a) **Max Pooling**

$$Y(i,j) = \max_{(m,n)\in\Omega} X(i+m, j+n) \tag{2.15}$$

Max pooling selects the maximum value within a local region Ω, highlighting the most prominent feature in that region.

(b) **Average Pooling**

$$Y(i,j) = \frac{1}{|\Omega|} \sum_{(m,n)\in\Omega} X(i+m, j+n) \tag{2.16}$$

Average pooling computes the average of all values in the local region, providing a more generalized summary.

4. **Batch Normalization**

   Batch Normalization (BN) standardizes the activations across a mini-batch. It normalizes each channel x to have zero mean and unit variance, then applies a scale and shift[58]:

$$\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\epsilon}}, \quad y = \gamma\hat{x} + \beta \tag{2.17}$$

   BN accelerates training and improves generalization by reducing internal covariate shift.

5. **Dropout Layer**

   Dropout is a regularization technique where a fraction p of units is randomly set to zero during training:

$$y_i = \begin{cases} 0 & \text{with probability } p \\ \frac{x_i}{1-p} & \text{otherwise} \end{cases} \tag{2.18}$$

   It prevents the model from overfitting by reducing co-adaptation of neurons[59].

6. **Fully Connected Layers**

   After convolutional and pooling layers, the 3D feature maps are flattened into a 1D vector and passed through one or more fully connected layers. Each neuron in a fully connected layer computes:

$$y = f(Wx + b) \tag{2.19}$$

   where W is the weight matrix, b the bias vector, and f is usually a non-linear activation such as softmax (for multi-class classification) or sigmoid[60].

2.3.3 Example of CNN Architecture for SER

A typical CNN for SER might consist of:

- Input: 128×128 Mel-spectrogram
- Conv Layer 1: 3×3 kernel, 32 filters + ReLU
- Max Pooling: 2×2
- Conv Layer 2: 3×3, 64 filters + ReLU
- Max Pooling: 2×2
- Dropout: 0.3
- Flatten
- Fully Connected Layer (128 units) + ReLU

- Output Layer: Softmax for emotion classification

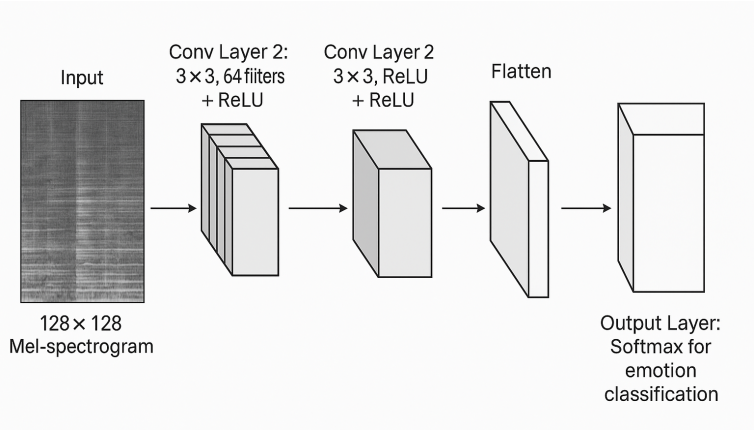

Figure 2.8: CNN Architecture for Speech Emotion Recognition

2.3.4 Variants of CNN Architectures in SER

To enhance the performance and efficiency of standard CNN architectures, several advanced variants have been developed. These architectures are particularly useful in SER systems where resource constraints or deep feature abstraction is needed.

1. **Depthwise Separable Convolutional Neural Networks (DSCNNs)**

   Depthwise Separable Convolutional Neural Networks (DSCNNs) are an efficient variant of standard CNNs that significantly reduce model size and computational cost while preserving performance. Introduced in architectures such as MobileNet, DSCNNs are particularly suitable for deployment in mobile, embedded, and low-latency systems—making them ideal for real-time Speech Emotion Recognition (SER) on edge devices[61].

   (a) **Standard Convolution Recap:**

   In a traditional 2D convolution operation, each filter operates across both spatial dimensions and the full depth (channels) of the input. Given: - an input feature map of size $D_F \times D_F \times M$, - a convolutional kernel of size $K \times K$, - N output channels,

   The number of computations required is:

$$\text{Cost}_{\text{standard}} = K \cdot K \cdot M \cdot N \cdot D_F \cdot D_F \tag{2.20}$$

   (b) **Depthwise Separable Convolution:**

   This operation is split into two successive layers:

   - **Depthwise Convolution:** Each input channel is convolved with its own filter independently. This step captures spatial correlations within each channel but does not mix information across channels.

$$\text{Cost}_{\text{depthwise}} = K \cdot K \cdot M \cdot D_F \cdot D_F \tag{2.21}$$

   - **Pointwise Convolution:** A 1×1 convolution is applied across all the M depthwise outputs to produce N output channels. This captures inter-channel dependencies.

$$\text{Cost}_{\text{pointwise}} = M \cdot N \cdot D_F \cdot D_F \tag{2.22}$$

(c) **Total Computational Cost:**

$$\text{Cost}_{\text{DSCNN}} = K \cdot K \cdot M \cdot D_F \cdot D_F + M \cdot N \cdot D_F \cdot D_F \tag{2.23}$$

(d) **Efficiency Gain:**

Compared to standard convolutions, the depthwise separable approach reduces computation by a factor of:

$$\frac{\text{Cost}_{\text{DSCNN}}}{\text{Cost}_{\text{standard}}} = \frac{1}{N} + \frac{1}{K^2} \tag{2.24}$$

For common configurations (e.g., $K = 3$, $M = N = 128$), this results in nearly an 8–9× reduction in computation and parameters.

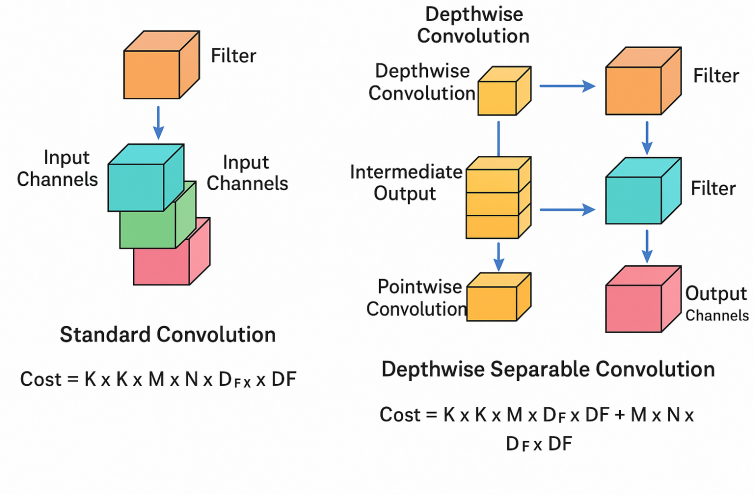

Figure 2.9: Comparison between Standard Convolution and Depthwise Separable Convolution.

(e) **Practical Architecture:**

A basic Depthwise Separable Convolutional Neural Network (DSCNN) block in Speech Emotion Recognition (SER) pipelines is designed to balance model complexity and performance. These blocks are structured to decompose standard convolution operations into more efficient alternatives while preserving spatial feature extraction[62].

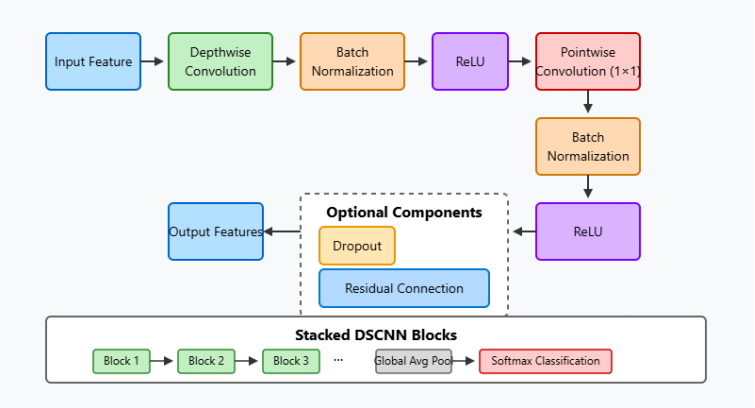

Figure 2.10: Depthwise Separable CNN Block for Speech Emotion Recognition

A typical DSCNN block consists of:

- **Depthwise Convolution:** Applies a single filter per input channel, capturing spatial features independently across each channel. This significantly reduces computational cost compared to standard convolution[63].
- **Batch Normalization:** Follows the depthwise convolution to stabilize learning and accelerate convergence by normalizing the output[64].
- **ReLU Activation:** Introduces non-linearity, allowing the model to learn complex feature interactions related to emotional cues in the speech signal[65].
- **Pointwise Convolution (1×1):** Projects the outputs of the depthwise convolution into a new feature space by combining the channels linearly. This operation enables inter-channel interaction and expands the network's representational capacity[66].
- **Batch Normalization and ReLU:** Applied again after the pointwise convolution to maintain consistency and support better training dynamics.
- **Optional Components:**
  - **Dropout:** Used to prevent overfitting by randomly zeroing a portion of the activations during training.
  - **Residual Connections:** May be included to facilitate gradient flow and improve training in deeper networks by allowing identity mappings.

These DSCNN blocks can be stacked in multiple layers to construct deeper and more expressive networks. The modularity and parameter efficiency of depthwise separable convolutions make them particularly suitable for SER tasks, where real-time inference and deployment on resource-constrained devices (e.g., mobile or embedded systems) are important considerations. When combined with global average pooling and a softmax classifier, this architecture can effectively learn discriminative features from audio spectrograms for emotion classification[67].

(f) **Benefits in SER:**

- **Low Resource Consumption:** Ideal for real-time SER on smartphones, hearing aids, or IoT devices in telehealth.
- **Fast Inference:** Lower latency makes DSCNNs suitable for interactive systems like virtual assistants or real-time feedback in educational apps.
- **Sufficient Expressiveness:** Despite fewer parameters, DSCNNs can still extract localized frequency and temporal patterns that correlate strongly with emotional states.

(g) **Applications in Arabic SER:**

Arabic SER often faces data scarcity, speaker and dialectal variability, and deployment constraints. DSCNNs are a strong candidate in such scenarios because:

- Their compact design enables on-device inference, reducing dependence on cloud computation.
- They can be fine-tuned on dialect-specific data, enabling emotion recognition even in underrepresented Arabic dialects.
- They support real-time speech analysis in smart devices used in healthcare and education across Arabic-speaking regions.

Depthwise Separable CNNs strike an excellent balance between efficiency and performance, making them a powerful tool for real-time, embedded, and scalable Arabic SER systems[62].

2. **Residual Convolutional Neural Networks (ResNets)**

Residual Convolutional Neural Networks (ResNets) are a class of deep CNN architectures designed to facilitate the training of very deep networks by addressing the degradation problem—where deeper models start to converge to higher training errors due to vanishing gradients[68]. ResNets use **skip connections** or **identity shortcuts** to allow gradients to flow directly through earlier layers, enabling stable training of networks with hundreds or even thousands of layers[69].

(a) **Mathematical Formulation:**

Instead of learning a direct mapping $H(x)$, a residual block learns a residual function $F(x) = H(x) - x$, so that the network actually learns:

$$y = F(x, W) + x \tag{2.25}$$

where:

- x is the input to the residual block,
- $F(x, W)$ represents the residual mapping learned by a series of convolutional (and optionally batch norm and ReLU) layers with weights W,
- y is the output of the block after combining the learned residual with the input.

This architecture ensures that the output of a layer can be directly influenced by the original input, maintaining strong gradient signals during backpropagation.

(b) **Basic Residual Block:**

A simple residual block typically consists of:

- Conv layer (e.g., 3×3) + BatchNorm + ReLU
- Conv layer (e.g., 3×3) + BatchNorm
- Identity shortcut connection: input x is added to the output of the second layer
- Final ReLU applied to the sum $F(x) + x$

For cases where the input and output dimensions do not match, a 1×1 convolution is applied to the identity path to project the input to the correct dimension before addition.

(c) **Advantages of ResNets:**

- **Improved Gradient Flow:** Skip connections help preserve gradients, allowing efficient training of very deep networks.
- **Easier Optimization:** Residual mappings are often easier to optimize than direct mappings, resulting in faster convergence and better performance.
- **Robust Generalization:** ResNets encourage feature reuse across layers, which improves generalization to unseen data.
- **Flexibility:** Residual blocks can be stacked to form deeper models or integrated into other architectures such as CNN-LSTM or DSCNN-ResNet hybrids.

(d) **Application in Speech Emotion Recognition (SER):**

In SER, especially when using 2D spectrograms or Mel-spectrograms as input, ResNets are highly effective for capturing complex emotional cues that unfold over time and frequency. Deeper networks can extract hierarchical emotional features ranging from low-level acoustic modulations to high-level prosodic and rhythmic patterns.

For example:

- ResNet-18 or ResNet-34 can be used as backbones in SER pipelines for both real-time and offline systems.
- Residual blocks can handle noisy environments and speaker variability more robustly compared to traditional CNNs.

- ResNet variants can be combined with temporal models (like BiLSTM or Transformer) for end-to-end modeling of both spectral and temporal emotional dynamics.

(e) **Visualization of Residual Learning:**

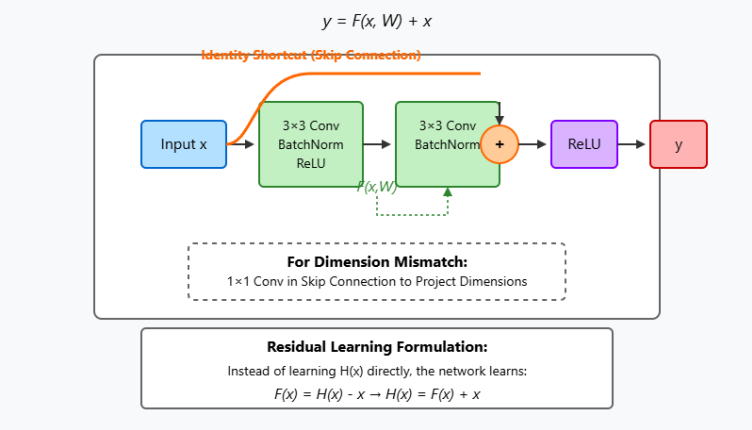

Figure 2.11: Structure of a basic residual block used in ResNets.

Residual Networks are a powerful extension to standard CNNs, particularly suitable for deeper models in SER. Their ability to model complex acoustic-emotional relationships while maintaining trainability makes them highly relevant for modern emotion recognition systems, especially in linguistically diverse settings like Arabic SER[70].

3. **Dilated Convolutional Networks**

Dilated Convolutional Networks—also known as **atrous convolutions**—are a powerful variant of standard CNNs that allow the expansion of the receptive field without increasing the number of parameters or losing resolution[71]. This is achieved by inserting "holes" (i.e., zeros or gaps) between the elements of the convolution kernel, enabling the model to capture broader contextual information while preserving the original input size[72].

(a) **Mathematical Definition:**

For a dilation rate r, the dilated convolution is defined as:

$$Y(i,j) = \sum_{m=0}^{k-1} \sum_{n=0}^{k-1} X(i + r \cdot m, j + r \cdot n) \cdot K(m,n) \tag{2.26}$$

Where:

- X is the input feature map,
- K is the convolutional kernel of size $k \times k$,
- r is the dilation rate (also referred to as the spacing between kernel elements),
- $Y(i,j)$ is the output at position (i,j).

When $r = 1$, the operation is equivalent to a standard convolution. Larger dilation rates allow the kernel to cover a broader area of the input, thereby increasing the receptive field exponentially without increasing computation.

(b) **Architectural Benefits:**

- **Increased Receptive Field:** Without increasing kernel size or the number of layers, the model can access more global contextual information.
- **No Downsampling Required:** Unlike pooling or strided convolutions, dilation expands the receptive field without reducing spatial resolution.

- **Efficient Parameter Use:** It introduces no new parameters, keeping the model lightweight.

(c) **Use Case in Speech Emotion Recognition (SER):**

In SER, emotional information is often spread across time, especially for emotions like:

- **Sadness or Calmness:** Typically manifested through slowly varying pitch and energy contours.
- **Fear or Anxiety:** Exhibits more subtle prosodic changes over longer time spans.

Dilated convolutions are ideal for modeling such **long-range dependencies** in time-frequency representations like Mel-spectrograms or MFCC matrices.

(d) **Example Use in SER:**

- A dilated convolutional block with 3×3 filters and dilation rates $r = 1, 2, 4$ can be stacked to provide multi-scale context aggregation.
- Often combined with residual or attention modules to form deep feature extractors in end-to-end SER pipelines.
- Can replace large kernel convolutions (e.g., 7×7) with smaller ones dilated over space, saving memory.

(e) **Visualization:**

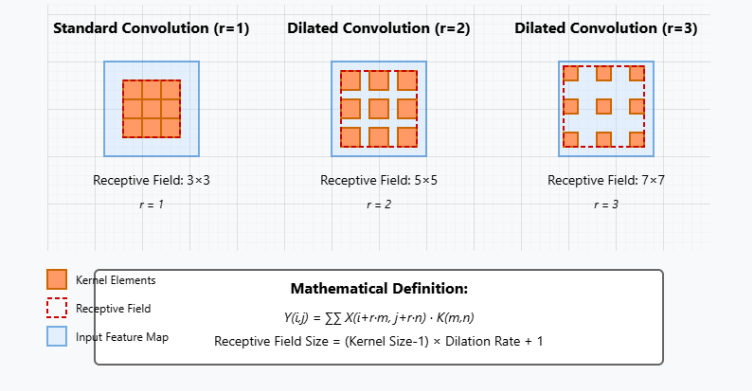

Figure 2.12: Illustration of dilated convolution with different dilation rates ($r = 1, 2, 3$). The receptive field grows exponentially as the dilation rate increases.

Dilated CNNs offer an effective and computationally efficient mechanism to extract contextual emotional patterns from speech signals. Their ability to model long-term dependencies without loss of resolution makes them especially relevant for emotion categories characterized by subtle, extended temporal cues.

4. **Hybrid CNN Architectures**

While Convolutional Neural Networks (CNNs) excel at learning local patterns from spectrograms and time-frequency representations, they are inherently limited in modeling long-range temporal dependencies. To address this limitation, CNNs are often integrated with other neural network architectures that complement their strengths. These hybrid models exploit the local feature extraction capabilities of CNNs along with the temporal or contextual modeling abilities of sequence-based or attention-based frameworks[73].

(a) **Motivation:**

Emotions in speech are expressed through both spatial and temporal dynamics. Spectrograms encode these variations across time and frequency. CNNs can detect localized

energy bursts, pitch contours, and harmonic structures, but for a more holistic understanding of emotion, models must also capture how these patterns evolve over time. Hybrid architectures achieve this goal by combining the best of both spatial and temporal modeling strategies[74].

(b) **Common Hybrid Models in SER:**

- **CNN-LSTM:** This hybrid combines convolutional layers with Long Short-Term Memory (LSTM) units. CNNs act as a front-end feature extractor, transforming input spectrograms or MFCCs into a sequence of high-level feature vectors. These are then fed into LSTM layers that capture the temporal evolution of emotional cues such as rising pitch, prosodic rhythm, or prolonged silence[75].
  - **Advantage:** CNN-LSTM models are well-suited for capturing both spatial locality and sequential dependency, essential for recognizing gradual emotional changes in speech.
  - **Use Case:** Detecting transitions from neutral to sadness over several seconds in a call center conversation.
- **CNN-Attention:** In this architecture, CNN-extracted features are passed through attention mechanisms—such as self-attention or multi-head attention—to weight the importance of each time-frequency region differently. The model learns to focus on emotionally salient segments, such as stressed syllables or sudden pitch changes[76].
  - **Advantage:** Attention modules improve interpretability and enable the model to selectively emphasize emotion-rich portions of the signal.
  - **Use Case:** Highlighting only the emotionally charged words in expressive speech while downplaying neutral or filler content.
- **CNN-Transformer:** This architecture leverages CNNs for localized feature encoding and Transformers for modeling global dependencies across the sequence. CNNs process spectral frames into high-level tokens, which are then input into Transformer encoders equipped with self-attention layers that model contextual relationships[77].
  - **Advantage:** Combines CNN's inductive bias for locality with Transformer's capacity for long-range modeling, yielding powerful and flexible SER systems.
  - **Use Case:** Real-time emotion detection in multilingual voice assistants, especially when dealing with variable-length utterances or complex speaking styles.

(c) **Visual Comparison:**

To better understand the structural differences between pure CNN and hybrid models, the figure below illustrates the architectural flow of:

- **CNN-only:** A stack of convolutional and pooling layers followed by a fully connected classifier.
- **CNN-LSTM:** CNN feature extractor followed by LSTM layers for temporal modeling.
- **CNN-Transformer:** CNN for feature encoding and Transformer for capturing global dependencies.

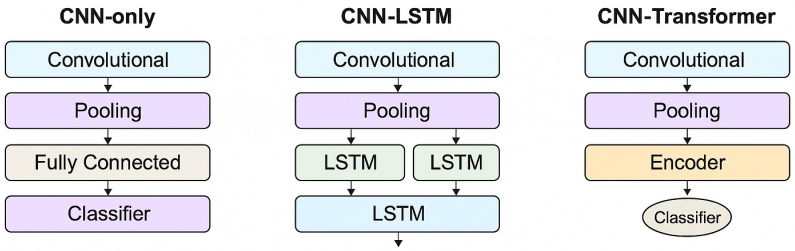

Figure 2.13: Comparison of CNN-only, CNN-LSTM, and CNN-Transformer architectures for Speech Emotion Recognition.

(d) **Performance in SER:**
Hybrid CNN architectures have consistently outperformed pure CNN or LSTM models in various SER benchmarks. Their effectiveness is especially notable in:

- **Low-resource settings:** Where CNNs handle feature extraction efficiently and attention helps mitigate the need for extensive labeled data.
- **Multilingual/dialectal SER:** Arabic SER, for instance, benefits from CNN-Attention combinations that can generalize across dialectal variations by learning where to focus.

Hybrid CNN models represent the evolution of SER architectures, combining robust feature extraction with temporal and contextual understanding. By bridging the gap between spatial and sequential processing, these models offer a comprehensive framework for modeling the complex, dynamic nature of emotional speech[78].

2.3.5 Advantages of CNNs in Arabic SER

Convolutional Neural Networks (CNNs) offer a powerful framework for modeling emotional content in speech, especially in linguistically rich and diverse languages like Arabic. Their architectural properties and learning capabilities make them particularly well-suited to the challenges of Arabic SER[45].

1. **Automatic Feature Learning:** Traditional SER systems often rely on hand-engineered acoustic features (e.g., pitch, energy, jitter), which require domain expertise and may not generalize well across languages or dialects. CNNs, in contrast, automatically learn task-specific features directly from spectrograms, Mel-spectrograms, or MFCCs. This reduces the dependency on language-specific feature engineering and enables the model to capture subtle emotional cues that may be missed by manually selected descriptors.

2. **Efficiency and Scalability:** Due to weight sharing across spatial dimensions, convolutional layers significantly reduce the number of learnable parameters compared to fully connected networks. This allows CNNs to be trained efficiently on moderate-sized datasets, a key advantage in Arabic SER where annotated emotional speech corpora are often limited. The smaller parameter space also helps mitigate overfitting, especially in low-resource dialectal settings.

3. **Robustness to Variability:** Arabic is spoken across more than 20 countries and encompasses a wide range of dialects, accents, and speaking styles. Moreover, variations in speaker identity, recording quality, and background noise are common in real-world data. CNNs offer robustness to these variabilities by focusing on local patterns in the input, such as energy bursts or harmonic structures, regardless of their exact temporal or frequency position. This makes CNNs highly resilient to accent shifts, regional prosody, and acoustic noise, which are prevalent in Arabic SER tasks.

4. **Hierarchical Feature Representation:** CNNs learn a hierarchy of features as the depth of the network increases. Lower convolutional layers typically capture low-level acoustic properties such as:

   - Formant transitions,
   - Pitch shifts,
   - Spectral onsets and harmonics.

   As information flows deeper into the network, higher convolutional layers begin to detect more abstract emotional constructs such as:

   - Stress and intonation patterns,
   - Temporal prosodic contours,
   - Speaker affect signatures.

   This hierarchical structure aligns well with the layered nature of human emotion expression in speech—starting from acoustic modulation to expressive phrasing and intonation—all of which are especially rich in Arabic due to its syllable-timed prosody and emphatic phonemes.

5. **Localization and Shift-Invariance:** Emotional features in Arabic speech may appear at different time points depending on sentence structure and speaker pacing. CNNs provide shift-invariance through local receptive fields and pooling operations, enabling detection of emotion-bearing acoustic cues regardless of their precise location in the utterance.

6. **Compatibility with Visual Representations:** Since many emotional features in speech are best visualized as 2D time-frequency patterns, CNNs—originally developed for image recognition—can be naturally applied to spectrogram-like inputs. This enables researchers to leverage pre-trained models and architectural innovations from the vision domain, accelerating progress in Arabic SER.

Overall, CNNs offer a highly effective and adaptable architecture for Arabic Speech Emotion Recognition. Their ability to learn data-driven, hierarchical representations—coupled with robustness to dialectal and acoustic variability—makes them an essential foundation for modern SER systems, especially in low-resource and multilingual contexts like the Arabic-speaking world.

Comparison with Other Architectures:

To contextualize the strengths of CNNs in Arabic SER, Table 2.1 compares them with other popular deep learning architectures—LSTMs and Transformers—across relevant technical and linguistic dimensions.

Table 2.1: Comparison of Deep Learning Architectures in Arabic Speech Emotion Recognition

Aspect	CNNs	LSTMs	Transformers
Primary Strength	Local pattern recognition	Sequential modeling	Global context modeling
Input Type	2D time-frequency features (e.g., spectrograms)	Sequences of features (e.g., MFCC frames)	Tokenized or embedded sequences
Temporal Awareness	Limited (local only)	High (learns long-term dependencies)	Very High (via self-attention)
Receptive Field	Local and hierarchical	Grows with sequence length	Full sequence (parallelized)
Training Efficiency	High (parallelizable)	Moderate (sequential updates)	Very high (fully parallel)
Robustness to Dialects	Good (filters capture shared patterns)	Fair (sensitive to input sequence noise)	Very good (contextual attention adapts dynamically)
Computational Cost	Low to moderate	Moderate to high	High (especially in long utterances)
Suitability for Arabic SER	Excellent for spectro-temporal emotion cues	Strong for modeling emotion progression over time	Ideal for combining acoustic and linguistic context
Typical Use in SER	Feature extraction from spectrograms	Modeling emotional trajectories	End-to-end or hybrid emotion detection

2.3.6 Limitations and Extensions

Despite their effectiveness in capturing localized acoustic features from time-frequency representations, Convolutional Neural Networks (CNNs) are not without limitations—especially in the context of emotion recognition from Arabic speech, which poses unique linguistic and computational challenges[79].

1. **Temporal Limitations:** CNNs are primarily designed to extract spatial features from fixed-size inputs. While they perform well at detecting short-term spectral patterns such as bursts, formants, or harmonics, they lack the ability to model temporal dependencies across frames. In Speech Emotion Recognition (SER), emotional expressions often evolve gradually over time—through prolonged intonation, syllable stress, or voice quality shifts. This limits the CNN's ability to fully capture the progression of emotions like sadness, confusion, or fear,

which may manifest across entire phrases or utterances rather than in isolated frames.

2. **Imbalanced Data:** One of the key challenges in Arabic SER is the scarcity of large-scale, balanced emotional speech datasets. Available corpora often exhibit significant class imbalance—for instance, an overrepresentation of neutral or happy samples and an underrepresentation of fear, disgust, or anger. CNNs, like most deep learning models, are sensitive to such imbalances and may become biased toward the majority classes, leading to poor generalization on minority emotions. This issue is exacerbated when dealing with dialectal Arabic, where certain emotional expressions may be culturally underrepresented or acoustically ambiguous.

3. **Domain Adaptation and Generalization:** CNNs trained on one Arabic dialect or acoustic setting may not generalize well to others due to differences in pronunciation, prosody, and vocabulary. Standard CNNs lack mechanisms for domain adaptation or contextual recalibration, making them less robust in multi-dialectal or cross-corpus SER tasks.

4. **Extensions and Solutions:**

   To address these limitations, several architectural and training strategies have been proposed:

   - **CNN-LSTM Hybrids:** To overcome the temporal modeling limitation, CNNs are often integrated with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks. CNN layers first extract spatial features from spectrograms, which are then passed as sequential inputs to LSTM layers that model emotional dynamics over time. This hybrid approach has proven effective in capturing both the spectral richness and temporal fluidity of emotional speech[80].
   - **CNN-Transformer Architectures:** Transformers, with their self-attention mechanism, provide an alternative to LSTMs by modeling global dependencies in a parallel and highly flexible manner. When paired with CNN-based front ends, Transformers can capture complex inter-frame relationships while preserving local spectro-temporal patterns—making them ideal for variable-length utterances and multilingual settings, including Arabic dialects[81].
   - **Attention Mechanisms:** Attention layers can be applied on top of CNN outputs to focus the model's attention on emotionally salient regions of the input (e.g., high-pitched segments, stressed syllables, or prosodic boundaries). This improves interpretability and helps mitigate the impact of irrelevant or noisy frames[82].
   - **Data Augmentation and Balancing:** To tackle class imbalance, techniques such as Synthetic Minority Oversampling Technique (SMOTE), emotion-based data augmentation (e.g., pitch shifting, time stretching), or weighted loss functions can be employed. Additionally, transfer learning from larger, non-Arabic datasets followed by fine-tuning on Arabic corpora improves generalization[83].
   - **Multi-Task and Domain-Adaptive Learning:** CNNs can be extended via multi-task learning setups where auxiliary tasks (e.g., speaker recognition, gender classification) guide the network to learn more generalized representations. Domain adaptation methods can also help transfer knowledge across Arabic dialects or recording environments[84].

5. **Visual Summary of Architectural Evolution:**

   The diagram below provides a conceptual visualization of the architectural progression from standard CNNs to advanced hybrid models used in Arabic SER, highlighting how each extension addresses specific limitations:

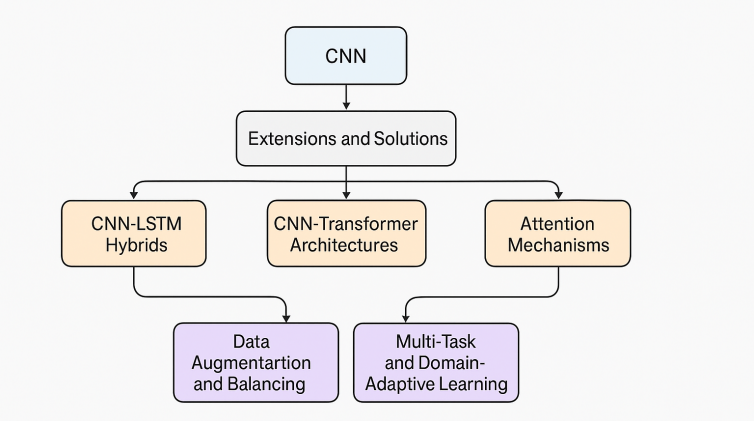

Figure 2.14: Architectural evolution: from CNN to CNN+LSTM and CNN+Transformer pipelines for Arabic Speech Emotion Recognition.

While CNNs have demonstrated significant promise in Arabic SER, their limitations—especially in modeling temporal dependencies and handling data imbalance—necessitate careful architectural design and training strategies. Hybrid models, attention mechanisms, and data augmentation offer effective pathways to enhance performance and build more robust, culturally sensitive emotion recognition systems.

2.4 Recurrent Neural Networks (RNN) and LSTM

Recurrent Neural Networks (RNNs) are a family of neural architectures specifically designed to process sequential data by maintaining internal memory through recurrent connections. In the context of speech processing—and particularly Speech Emotion Recognition (SER)—this characteristic is essential because speech is inherently temporal: emotions are conveyed not only through instantaneous acoustic events but also through their evolution over time[85].

In SER, the input is typically a sequence of acoustic feature vectors (e.g., MFCCs or spectrogram frames), each representing a short time window. Standard feedforward neural networks treat each frame independently, whereas RNNs preserve context by maintaining a hidden state that is updated at each time step based on both the current input and the previous hidden state[86]:

$$h_t = \phi(W_h h_{t-1} + W_x x_t + b) \tag{2.27}$$

where h_t is the hidden state at time t, x_t is the input at time t, and ϕ is a non-linear activation function (typically $\tanh$ or ReLU).

However, traditional RNNs suffer from two major problems during training:

- **Vanishing Gradients:** Gradients diminish exponentially through time steps, preventing the network from learning long-term dependencies.
- **Exploding Gradients:** In some cases, gradients can grow uncontrollably, destabilizing training.

These issues limit standard RNNs to modeling only short-term patterns, making them insufficient for SER tasks where emotional content is often distributed across several seconds of speech.

- **Visual Summary of Flow:** The diagram below illustrates the fundamental difference between RNNs, LSTMs, and BiLSTMs in terms of how they process input sequences:

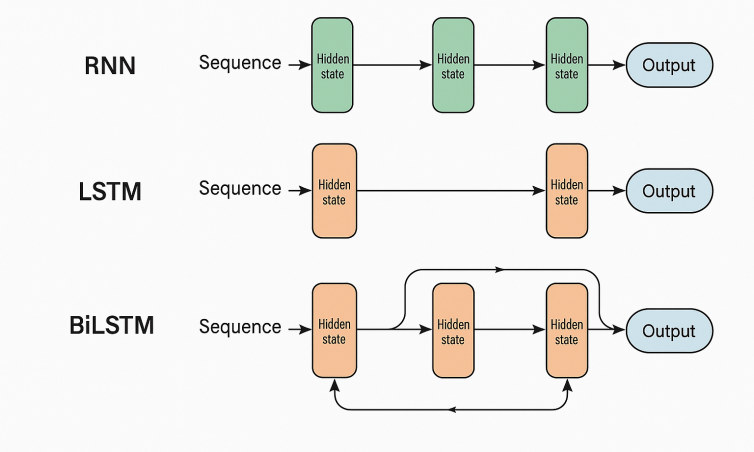

Figure 2.15: Comparison of sequence flow in RNN, LSTM, and BiLSTM architectures.

2.4.1 Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks were proposed to overcome the limitations of traditional RNNs by introducing an internal memory mechanism that can learn what information to retain and what to forget. An LSTM unit contains[87]:

- A **cell state** C_t, which acts as memory and carries long-term information across time steps.
- Three **gates**—input, forget, and output gates—that regulate the flow of information.

The operations of an LSTM cell are defined by the following equations:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad \text{(forget gate)} \tag{2.28}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad \text{(input gate)} \tag{2.29}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad \text{(candidate values)} \tag{2.30}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad \text{(cell state update)} \tag{2.31}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad \text{(output gate)} \tag{2.32}$$

$$h_t = o_t \odot \tanh(C_t) \quad \text{(hidden state output)} \tag{2.33}$$

1. **Architectural Visualization:**

   The diagram below illustrates and compares the internal structures of a traditional RNN cell, an LSTM cell, and a BiLSTM block.

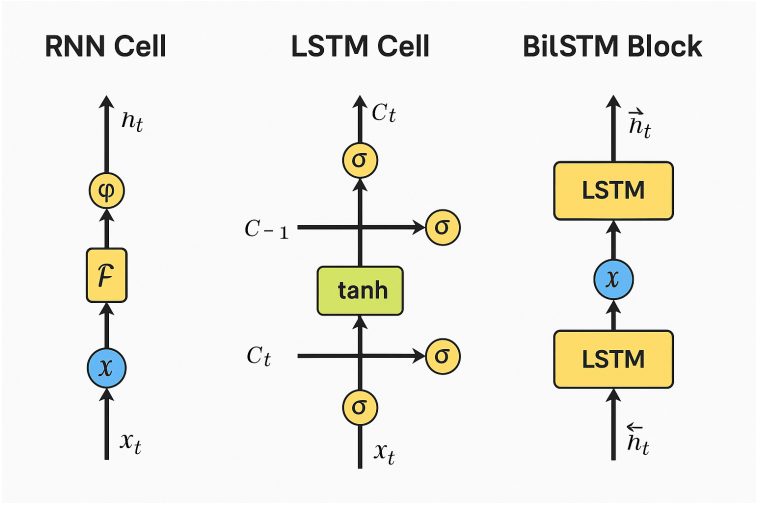

Figure 2.16: Architectural comparison of RNN cell, LSTM cell, and BiLSTM block.

2. **Application in SER:**

   LSTMs are highly effective in Speech Emotion Recognition due to their ability to track the flow of emotional cues across time. For instance:

   - Rising pitch and intensity may gradually indicate increasing anger.
   - Long pauses, soft intonation, and slow rhythm may signal sadness or tiredness.

   By capturing such transitions, LSTMs can significantly improve classification accuracy compared to CNNs alone.

2.4.2 Bidirectional LSTM (BiLSTM)

Bidirectional Long Short-Term Memory networks (BiLSTMs) extend the capability of traditional unidirectional LSTMs by processing input sequences in both temporal directions: forward (from the beginning to the end of the sequence) and backward (from the end to the beginning). This dual traversal allows the network to capture context from both the past and the future, which is particularly valuable in tasks like Speech Emotion Recognition (SER), where the meaning and emotion of a particular moment can be influenced by what came before and after it[88].

1. **Architecture:**

   A BiLSTM consists of two separate LSTM layers:

   - A forward LSTM, which reads the input sequence from $t = 1$ to T,
   - A backward LSTM, which processes the same sequence in reverse, from $t = T$ to 1.

   At each time step t, the final hidden state is the concatenation of the outputs from both directions:

$$h_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{2.34}$$

2. **Visual Representation:**

   The diagram below illustrates the flow of information in a BiLSTM network applied to a short Arabic utterance. It highlights how the forward and backward LSTMs process the input in opposite directions, allowing each frame to benefit from both past and future context.

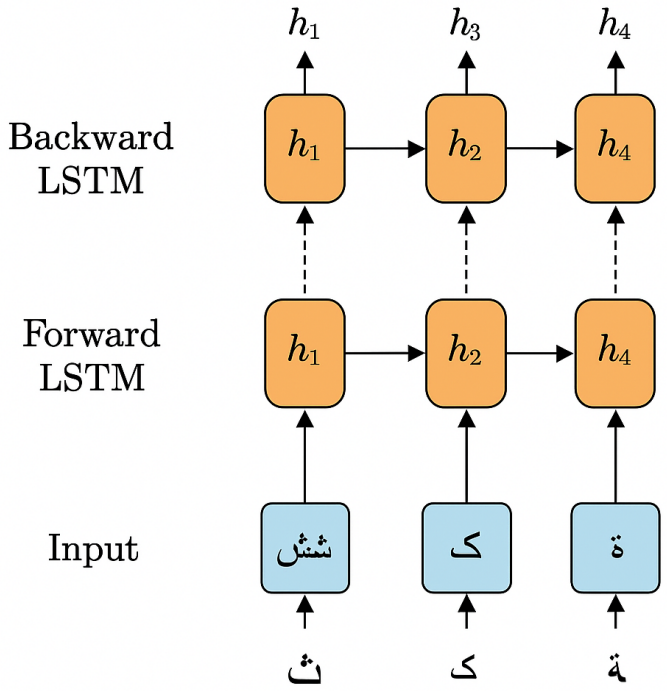

Figure 2.17: Bidirectional LSTM processing an Arabic utterance: dual temporal context from forward and backward LSTM layers.

3. **Advantages over Unidirectional LSTM:**

   - **Contextual Awareness:** Unidirectional LSTMs only consider past information, which may be insufficient for disambiguating emotion. For instance, a drop in pitch may indicate sadness, but only if not followed by an excited tone[89].
   - **Improved Disambiguation:** Many emotions share similar acoustic properties in short segments. Bidirectional processing helps resolve such ambiguities by incorporating subsequent context into the decision[90].
   - **Richer Temporal Encoding:** By accessing the full span of an utterance at each time step, BiLSTMs can capture prosodic patterns and emotional transitions more effectively[91].

4. **Application in Arabic SER:**

   In Arabic, emotion expression is deeply tied to:

   - **Phonetic Complexity:** Use of emphatic, pharyngeal, and uvular sounds varies across dialects.
   - **Syntactic Flexibility:** Arabic allows reordering of subject-verb-object (SVO) constructions for emphasis.
   - **Prosodic Flow:** The rhythm and pitch contours are influenced by both preceding and succeeding syllables.

   These linguistic properties make bidirectional models like BiLSTM especially well-suited, as they consider both backward and forward dependencies—critical when recognizing emotions that unfold gradually or are context-dependent.

5. **Use Case Example:**

   Imagine an utterance that begins neutrally but ends with a rising pitch and increased energy. A forward LSTM may classify the early frames as neutral and downplay later cues, while a BiLSTM captures the full emotional arc and adjusts earlier interpretations based on what follows—allowing it to correctly classify the entire sequence as conveying "surprise" or "excitement."

6. **Performance Gains:**

   BiLSTMs have been empirically shown to:

   - Outperform unidirectional LSTMs in both accuracy and F1-score across SER datasets,
   - Improve emotion recognition in short utterances where surrounding context is vital,
   - Increase robustness in dialectal Arabic by learning more generalized contextual embeddings.

7. **Hybrid Architectures:**

   BiLSTMs are often integrated with CNNs in SER pipelines:

   - **CNN layers** extract spatial features from spectrograms or MFCCs.
   - **BiLSTM layers** capture temporal dynamics using bidirectional sequence modeling.

   This CNN-BiLSTM configuration is considered a strong baseline in modern SER systems due to its balance between spatial and temporal feature extraction.

8. **Conclusion:**

   BiLSTMs are a powerful enhancement over traditional LSTMs for emotion recognition from speech. By incorporating full bidirectional context at each time step, they are able to more accurately capture the subtle and temporally extended cues that define emotional speech—making them particularly valuable in handling the prosodic and dialectal diversity found in Arabic speech.

9. **Application in Arabic SER:**

   Arabic presents unique challenges for SER due to its:

   - **Complex prosody:** Arabic is a syllable-timed language with intricate intonation patterns.
   - **Dialectal variation:** Different dialects may express the same emotion using different pitch, tempo, or phoneme patterns.

   BiLSTMs are particularly effective in handling such complexity because they analyze each frame of speech with awareness of both preceding and succeeding acoustic context. This helps disambiguate emotions that may otherwise be confused due to regional variation or speech artifacts.

10. **Use Case Example:**

    Consider a sentence where a rise in pitch at the end may either indicate a question (in neutral speech) or emphasize strong emotion (e.g., surprise or anger). A BiLSTM can correctly interpret the emotional context by considering what preceded and followed the change.

Recurrent models, especially LSTM and BiLSTM architectures, provide a critical component in SER systems by enabling sequential modeling of emotional content. When combined with CNNs or attention mechanisms, they form the backbone of state-of-the-art emotion recognition systems for both English and morphologically rich languages like Arabic.

2.5 Transformer-Based Models

Transformer architectures have revolutionized sequence modeling in recent years, especially in natural language processing (NLP) and, more recently, speech-related tasks. Unlike traditional Recurrent Neural Networks (RNNs), which process input tokens sequentially and are thus inherently limited in handling long-range dependencies and parallelism, Transformers allow for simultaneous access to the entire sequence[92]. This is made possible through the mechanism of **self-attention**, which dynamically computes relationships between all elements in the sequence regardless of their distance[93].

In Speech Emotion Recognition (SER), this property is crucial. Emotional patterns can manifest in distant segments of an utterance, and capturing these cross-temporal dependencies is key to reliable recognition. For Arabic SER, where intonation and emotional cues may be distributed unevenly across sentences or influenced by dialect-specific speech flow, Transformer-based models offer an especially powerful framework[94].

2.5.1 Self-Attention Mechanism

The core innovation in Transformer models is the self-attention mechanism, which enables the model to learn which parts of the sequence to focus on when encoding each input element. Formally, self-attention takes as input three matrices[95]:

- Q (Query),
- K (Key),
- V (Value),

all derived from the same input features through learned linear projections. The attention output is a weighted sum of the values V, where weights are determined by the similarity between the query and key vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{2.35}$$

1. **Visual Explanation:**

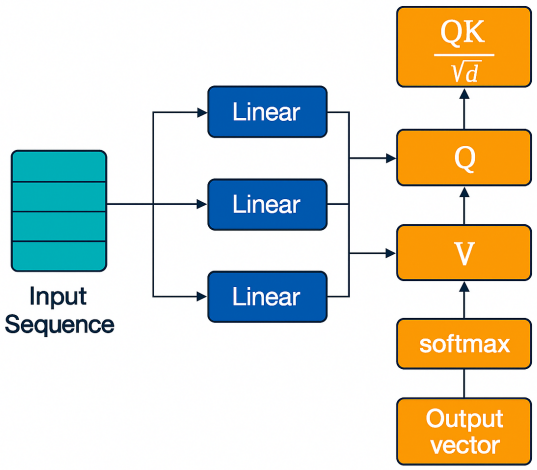

Figure 2.18: Visualization of the self-attention mechanism in Transformer models.

2. **Multi-Head Attention:**

To capture multiple types of dependencies simultaneously, Transformers use multi-head attention, where the input is projected into several subspaces, and attention is computed in parallel before being concatenated and reprojected. This allows the model to learn different types of prosodic or linguistic dependencies relevant to emotion recognition[96].

3. **Benefits for SER:**

   - Flexibly captures emotional cues that span long or variable durations.
   - Does not require rigid sequence modeling, which benefits applications like SER on variable-length utterances.
   - Highly parallelizable, enabling efficient training even on large corpora.

2.5.2 Pre-trained Audio Transformers

Transformer-based models pre-trained on massive amounts of unlabeled speech data have emerged as foundational architectures in modern speech processing. Unlike traditional supervised learning, these models leverage self-supervised learning to extract universal speech representations, which can be fine-tuned for downstream tasks such as Speech Emotion Recognition (SER)[97].

These models are particularly valuable for Arabic SER, where large-scale annotated emotional corpora are rare, and dialectal variation poses additional challenges. The key advantage of these models lies in their ability to learn contextual embeddings directly from raw audio, enabling effective generalization across speakers, dialects, and acoustic conditions[98].

1. **Wav2Vec 2.0**

   Wav2Vec 2.0, is a self-supervised model that learns from raw waveform inputs using a contrastive learning objective. Its architecture consists of[99]:

   - A **feature encoder** f, typically a stack of temporal convolution layers, that maps the raw waveform x into a latent representation $z = f(x)$.
   - A **context network** g, implemented as a Transformer, that maps the latent sequence to contextualized representations $c = g(z)$.
   - A **quantization module** that discretizes the latent representations z into quantized tokens q.
   - **Loss Function:**

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(c_t, q_t)/\kappa)}{\sum_{q' \in \mathcal{Q}} \exp(\text{sim}(c_t, q')/\kappa)} \tag{2.36}$$

   - **Architecture Visualization:**

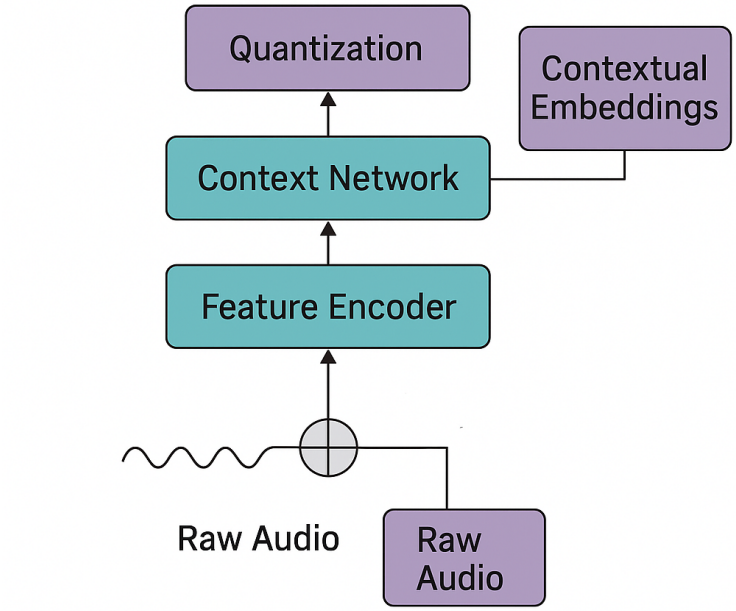

Figure 2.19: Architecture of Wav2Vec 2.0: raw audio to contextual embeddings.

2. **HuBERT (Hidden-Unit BERT)**

   HuBERT, extends the self-supervised paradigm by incorporating unsupervised clustering. The core idea is to use a masked prediction task where the model learns to predict pseudo-labels derived from k-means clustering of acoustic features[100].

   - A CNN-based feature encoder transforms audio into latent representations.
   - Clustering (e.g., k-means) produces hidden units that serve as prediction targets.
   - A Transformer predicts the masked targets from surrounding context.
   - **Loss Function:**

$$\mathcal{L}_{\text{HuBERT}} = -\sum_{t \in \mathcal{M}} \log P(y_t | x_{\setminus t}) \tag{2.37}$$

   - **Architecture Visualization:**

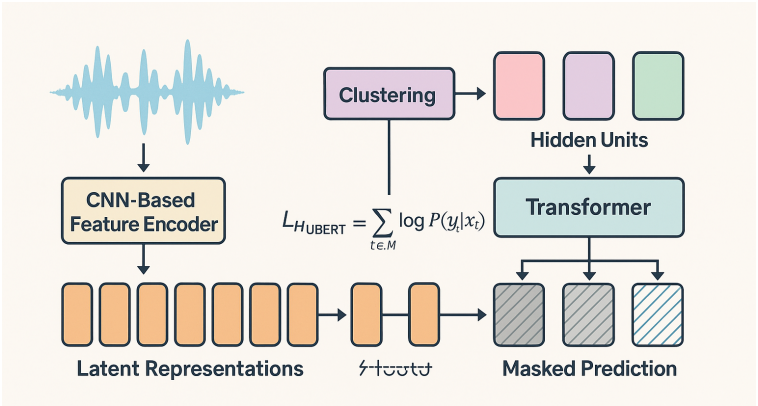

Figure 2.20: Architecture of HuBERT: learning hidden speech units through masked prediction.

3. **Whisper**

Whisper, developed by OpenAI, is a large-scale sequence-to-sequence model trained for multilingual ASR and speech translation. Unlike Wav2Vec and HuBERT, Whisper is fully supervised, trained end-to-end on hundreds of thousands of hours of labeled data[101].

- An **audio encoder** based on a convolutional front-end followed by Transformer layers.
- A **text decoder** Transformer that generates transcriptions, translations, or other outputs.

- **Model Formulation:**
  Let x be the input Mel-spectrogram and y the output token sequence. The Whisper model estimates:

$$P(y|x) = \prod_{t=1}^{T} P(y_t|y_{<t}, x) \tag{2.38}$$

- **Architecture Visualization:**

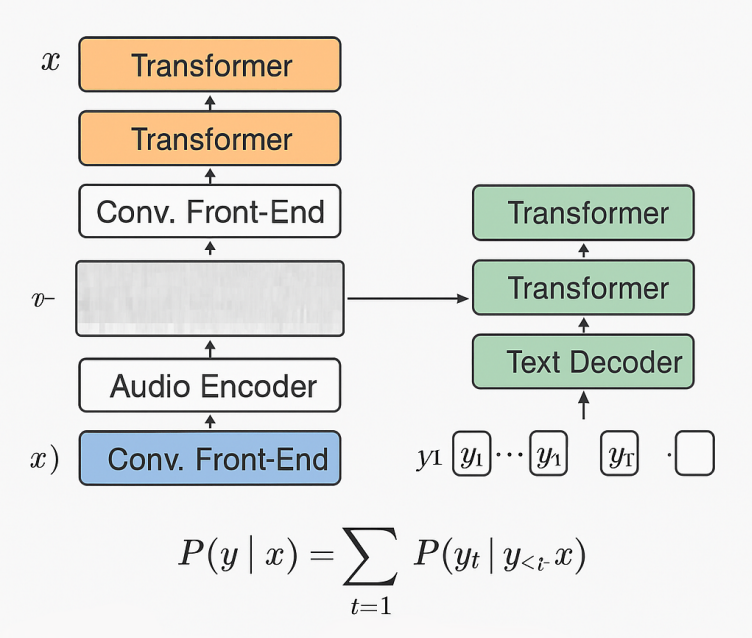

Figure 2.21: Architecture of Whisper: encoder-decoder Transformer for multilingual ASR and SER tasks.

- **Applications in Arabic SER:**
  - **Dialect Generalization:** Pre-trained models mitigate dialectal variation by learning universal acoustic representations.
  - **Low-Resource Fine-Tuning:** Only a small amount of annotated Arabic data is required to adapt models like Wav2Vec or HuBERT for emotion recognition tasks.
  - **End-to-End Capability:** These models can take raw waveforms as input and learn emotional embeddings without requiring handcrafted features.
- **Hybrid Architectures:**
  Transformers are often integrated with CNNs or BiLSTMs for SER:
  - **CNN-Transformer:** CNNs act as front-end feature extractors (e.g., on spectrograms), while Transformers model the long-range dependencies.

- **Wav2Vec + Classification Head:** A linear or Transformer-based decoder is added on top of the pre-trained model for emotion classification.

- **Visual Pipeline Comparison:**

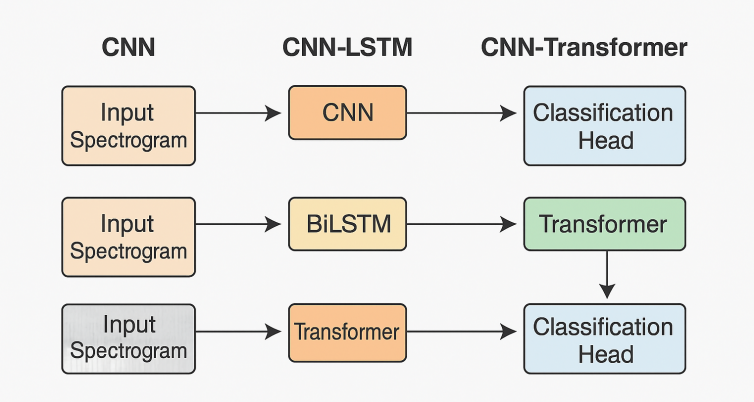

Figure 2.22: Architectural comparison of CNN, CNN-LSTM, and CNN-Transformer pipelines for Speech Emotion Recognition.

Transformer-based models represent a major leap forward in emotion recognition from speech. Their ability to model long-range contextual dependencies, coupled with pre-training on large multilingual audio datasets, makes them ideally suited for the complex prosodic, syntactic, and dialectal characteristics of Arabic speech. As pre-trained Transformers become more accessible and efficient, they are poised to become the foundation of next-generation Arabic SER systems.

2.6 Comparative Analysis of Feature Extraction Techniques

Table 2.2: Comparison of Feature Extraction Methods

Method	Input Type	Temporal Awareness	Complexity	Pros	Cons
MFCC	Raw Audio	Low	Low	Efficient	Limited context
Spectrogram	Raw Audio	Medium	Medium	Rich features	High dimensionality
CNN	Spectrogram/MFCC	Low	Medium	Pattern capture	No temporal modeling
LSTM	Feature Sequence	High	Medium	Long-term memory	Slow training
BiLSTM	Feature Sequence	Very High	Medium	Full context	More computation
Transformer	Feature/Raw	Very High	High	Long dependencies	Data hungry

2.7 Conclusion

This chapter provided a detailed exploration of feature extraction methods for Arabic speech emotion recognition. We discussed traditional spectrum-based methods, such as MFCC and spectrograms, which are effective due to their simplicity and efficiency in capturing emotional cues. Chroma and pitch features offer complementary prosodic information, critical for languages with rich intonation like Arabic.

We also explored deep learning techniques such as CNNs, which excel at capturing spatial patterns, and LSTMs and BiLSTMs, which are powerful for modeling temporal dynamics. The emergence of transformer-based architectures, particularly pre-trained models, is a game-changer in SER, offering an effective solution for low-resource languages like Arabic.

Ultimately, a hybrid approach that combines spectral features, prosodic information, and deep contextual models will likely yield the best results for Arabic SER. The method chosen will depend on task-specific requirements, the availability of labeled data, and computational constraints.

Chapter 3

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Introduction

In this part of the manuscript, we present our achievements in Arabic speech emotion recognition. To compare the studied approach with other deep learning architectures, several conditions must be satisfied. Firstly, we need to define the same evaluation criteria across all methods. Secondly, we must ensure the same preprocessing operations and corpus configurations are used. Finally, to make our comparison fairer, it is preferable to adopt similar training parameters (learning rate, batch size, number of epochs, etc.).

In this chapter, Section 3.2 details the conducted experiments. Section 3.3 describes the datasets used for the evaluation. Section 3.4 presents the results of the pre-processing phase. Section 3.5 defines the evaluation metrics employed (accuracy, recall, precision, F1 score). Section 3.6 shows the classification results for each deep learning model. Finally, Section 3.7 provides a comparative study.

3.2 Experiments

In this section, we describe the series of experiments designed and conducted with the aim of recognizing emotions from Arabic speech signals. Given the complexity and variability inherent to emotional speech, particularly in low-resource languages such as Arabic, it is crucial to systematically explore different deep learning architectures capable of capturing both the acoustic and temporal dynamics of spoken language.
The experimental framework developed in this study is composed of three distinct approaches:

- **Fine-tuning a pre-trained wav2vec 2.0 model:** This experiment investigates the performance of transfer learning by adapting a large-scale, self-supervised speech model trained on Arabic to the specific task of emotion classification. The goal is to leverage powerful pre-trained feature representations to overcome the limited size of labeled emotional speech datasets.
- **Designing a parallel CNN-LSTM architecture with an attention mechanism:** In this experiment, we propose a custom hybrid model that combines convolutional neural networks (CNNs) for local feature extraction with Long Short-Term Memory (LSTM) networks for temporal sequence modeling. An attention mechanism is integrated to allow the model to dynamically focus on the most informative parts of the input sequence during classification.

- **Developing a parallel CNN-Transformer architecture:** In the third experiment, we further extend the previous model by replacing the LSTM component with a Transformer encoder. The Transformer, known for its powerful self-attention mechanisms, is used to better capture long-range dependencies and contextual relationships within the speech signal.

Each experimental setup is carefully crafted to isolate and examine the impact of different architectural choices on the task of Arabic speech emotion recognition. Moreover, consistent preprocessing procedures, training protocols, and evaluation metrics are applied across all experiments to ensure a fair and rigorous comparison.
For each approach, a detailed workflow diagram is provided to illustrate the data flow, model components, and training procedures. These diagrams serve to clarify the conceptual and practical differences between the models, as well as to provide an intuitive understanding of the design rationale underlying each experimental setup.
The comparative study of these three models is essential to understand the strengths and limitations of different architectural strategies in Arabic speech emotion recognition. By evaluating transfer learning approaches alongside custom-designed hybrid models, we aim to determine which techniques are most effective in capturing the complex acoustic and sequential patterns inherent in emotional speech. This comparative analysis not only highlights the impact of model design choices but also provides valuable insights for future developments in the field. In the next sections, we present each experiment in detail, beginning with the fine-tuning of the wav2vec 2.0 model.

3.2.1 Experiment 1: Fine-Tuning the wav2vec 2.0 Model

The diagram illustrated in Figure 3.1 presents the overall workflow of the emotion recognition system based on the wav2vec 2.0 model. This process is structured into three main phases, from raw audio preparation to final performance evaluation.

1. **- The first phase** involves the preparation of the EYASE dataset. This dataset, provided in a compressed format, is first extracted and structured into subdirectories according to speaker gender (male/female) and emotion labels (anger, happiness, neutrality, and sadness). Once the audio files are organized, each file undergoes a resampling step to a sampling rate of 16 kHz to meet the input specifications of the wav2vec 2.0 architecture. In addition to resampling, normalization is applied to all audio signals to reduce variability between different recordings, enhancing consistency for the model during training.

2. **- The second phase** consists of loading the pre-trained Arabic wav2vec 2.0 model available from the Hugging Face model hub under the identifier `elgeish/wav2vec2-large-xlsr-53-arabic`. This model has already been trained on large-scale Arabic speech corpora and provides rich representations of audio waveforms. To adapt this model to our specific task of emotion classification, a custom classification head is added. This head includes a Dropout layer (to prevent overfitting), a fully connected Dense layer, a Tanh activation to introduce non-linearity, another Dropout layer for regularization, and a final Linear layer responsible for outputting class probabilities over the four emotion classes. Importantly, the pre-trained feature extractor layers of the wav2vec 2.0 model are frozen during fine-tuning; only the new classification head is trained. This strategy significantly reduces training time and prevents catastrophic forgetting of the learned audio features.

3. **- The third phase** covers the training and evaluation process. The model is trained using a weighted CrossEntropy loss function, where class weights are computed based on the frequency distribution of emotion classes in the EYASE dataset. This approach helps address the class imbalance problem, ensuring that underrepresented emotions are not ignored during

learning. In addition, the training process is conducted in mixed-precision mode (fp16), which optimizes both GPU memory usage and computational efficiency. The training is performed over multiple epochs with early stopping based on validation loss to prevent overfitting.
At the end of training, the model is evaluated using four standard classification metrics: accuracy, precision, recall, and F1-score. These metrics provide a comprehensive view of the model's ability to generalize and correctly recognize emotional states from Arabic speech. A confusion matrix may also be used to visualize how well each emotion class is identified.

This entire process, from data preparation through training to final evaluation, is summarized in the diagram presented in Figure 3.1.

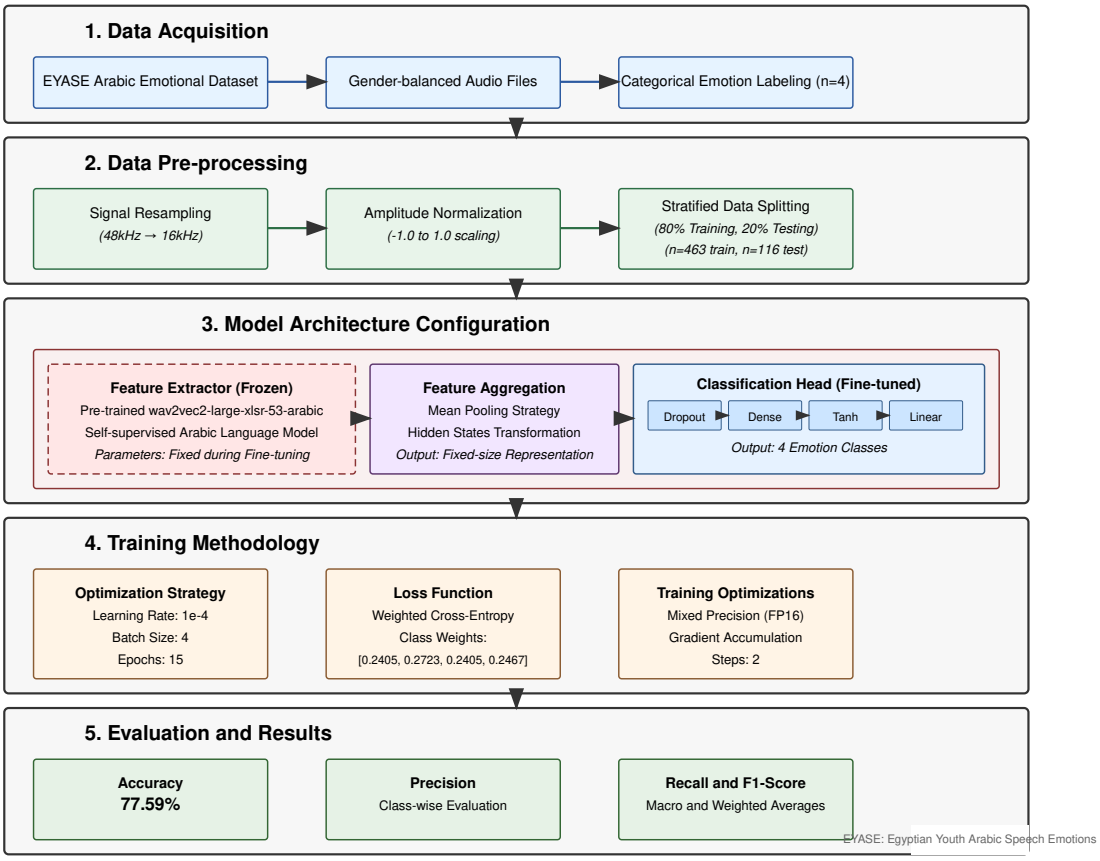

Figure 3.1: General Process of Emotion Recognition using wav2vec 2.0

3.2.2 Experiment 2: Parallel CNN-LSTM with Attention Mechanism

The diagram illustrated in Figure 3.2 presents the general workflow of the emotion recognition system based on a parallel CNN and LSTM architecture enhanced with an attention mechanism. This process is organized into three main phases: preprocessing and feature extraction, parallel feature modeling, and classification.

1. **- The first phase** consists of preprocessing the EYASE dataset. The audio files are first resampled to a uniform sampling rate of 16 kHz to ensure consistency across the dataset. Following this, normalization is applied to the audio waveforms. To further enhance the

model's robustness against real-world variations, data augmentation techniques are introduced. Specifically, Gaussian white noise is added to a portion of the training samples to simulate different recording conditions. After normalization and augmentation, Mel-Spectrograms are extracted from each audio signal. The Mel-Spectrograms serve as the main input features for the subsequent model training, offering a rich time-frequency representation of speech.

2. **- The second phase** involves feeding the extracted Mel-Spectrograms into a custom-designed parallel architecture. This architecture comprises two main branches:

   - **The CNN Branch:** This branch consists of multiple convolutional layers designed to extract spatial features from the Mel-Spectrograms, such as local time-frequency patterns that are often associated with particular emotional states.
   - **The LSTM Branch with Attention:** In parallel, the second branch uses a Long Short-Term Memory (LSTM) network to model the sequential dynamics present in speech signals. An attention mechanism is applied on top of the LSTM outputs, allowing the model to focus selectively on the most informative parts of the sequence when making predictions.

   After feature extraction in both branches, the outputs are concatenated and passed through a dense layer to merge spatial and sequential information.

3. **- The third phase** is dedicated to classification and evaluation. The merged features are fed into a fully connected layer followed by a softmax activation, producing a probability distribution over the emotion classes. Training is carried out using the Stochastic Gradient Descent (SGD) optimizer with a carefully tuned learning rate. To monitor the model's performance, the validation loss and accuracy are tracked at each epoch. The final evaluation metrics include accuracy, precision, recall, and F1-score.

This complete process, from preprocessing to classification, is illustrated in Figure 3.2.

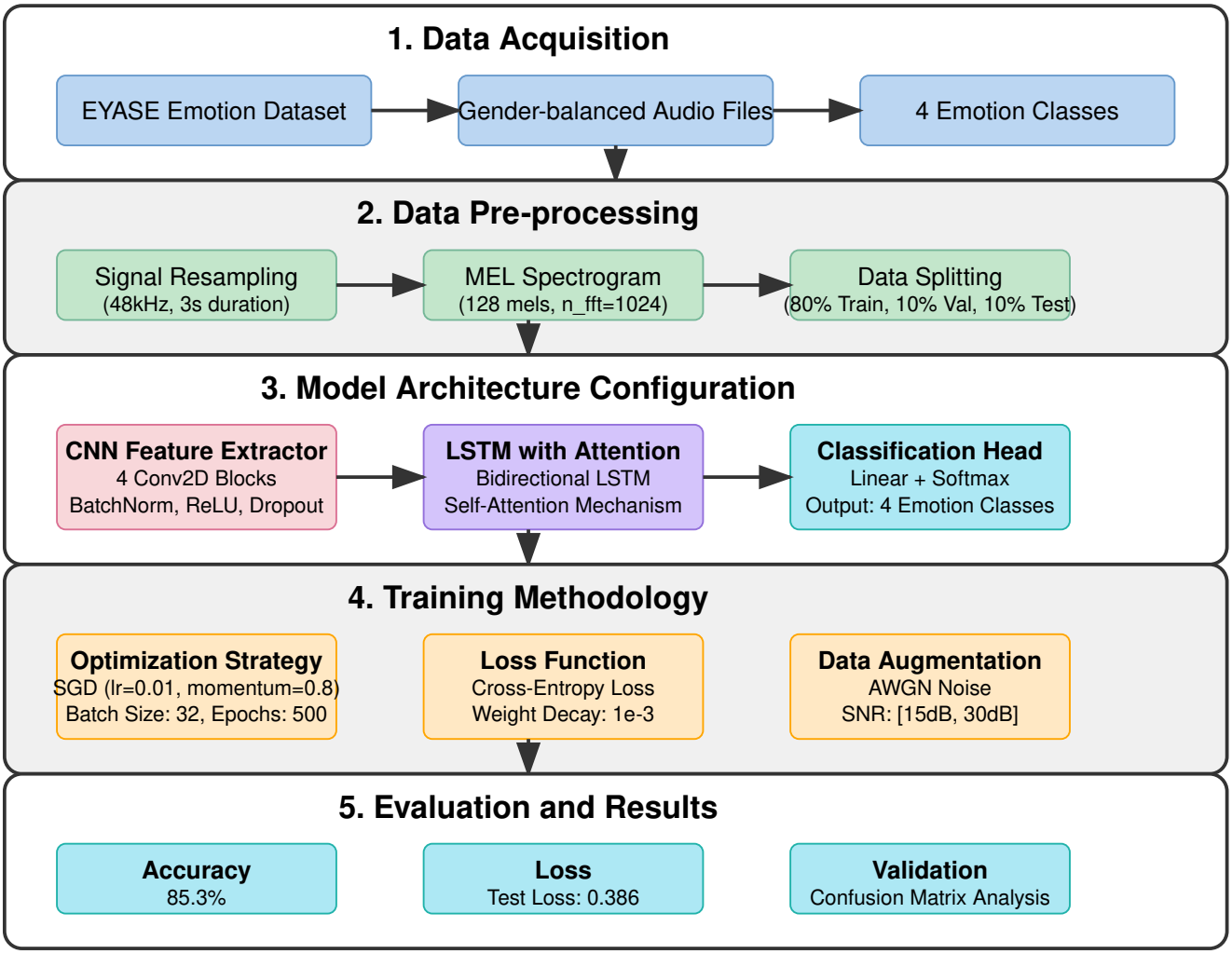

Figure 3.2: General Process of Emotion Recognition using Parallel CNN and LSTM with Attention

3.2.3 Experiment 3: Parallel CNN-Transformer Architecture

The diagram illustrated in Figure 3.3 depicts the overall workflow of the emotion recognition system based on a parallel CNN-Transformer architecture. This process follows a structure similar to the previous experiment but replaces the LSTM branch with a Transformer encoder for better sequence modeling.

1. **- The first phase** again begins with preprocessing the EYASE dataset. All audio files are resampled to 16 kHz and normalized to a standard amplitude range. Mel-Spectrograms are extracted from each audio recording, providing a rich 2D time-frequency representation. Data augmentation through Gaussian noise addition is also applied, similar to Experiment 2, to improve model robustness to variability in recording conditions.

2. **- The second phase** involves feeding the Mel-Spectrogram features into a parallel architecture composed of two independent branches:

   - **The CNN Branch:** This branch captures local spatial features from the Mel-Spectrograms using convolutional layers. These layers are particularly effective at identifying short-term patterns that correspond to phonetic and prosodic elements related to emotion.

- **The Transformer Branch:** Instead of an LSTM, the second branch uses a Transformer encoder. The Transformer, originally introduced for sequence modeling in natural language processing, allows the model to capture long-range dependencies and global contextual information within the speech signal. Multi-head self-attention mechanisms enable the model to attend to multiple parts of the input simultaneously, making it highly effective for emotional state modeling.

The outputs of the two branches are concatenated to form a combined feature vector that integrates both local and global information.

3. **- The third phase** concerns classification and evaluation. The concatenated features are passed through a fully connected layer and a softmax activation to produce class probabilities. Training is performed using the SGD optimizer. Validation performance is monitored throughout training to prevent overfitting. Evaluation metrics such as accuracy, precision, recall, and F1 score are computed to assess the effectiveness of the model.

This complete workflow, from raw data to final classification, is synthesized in Figure 3.3.

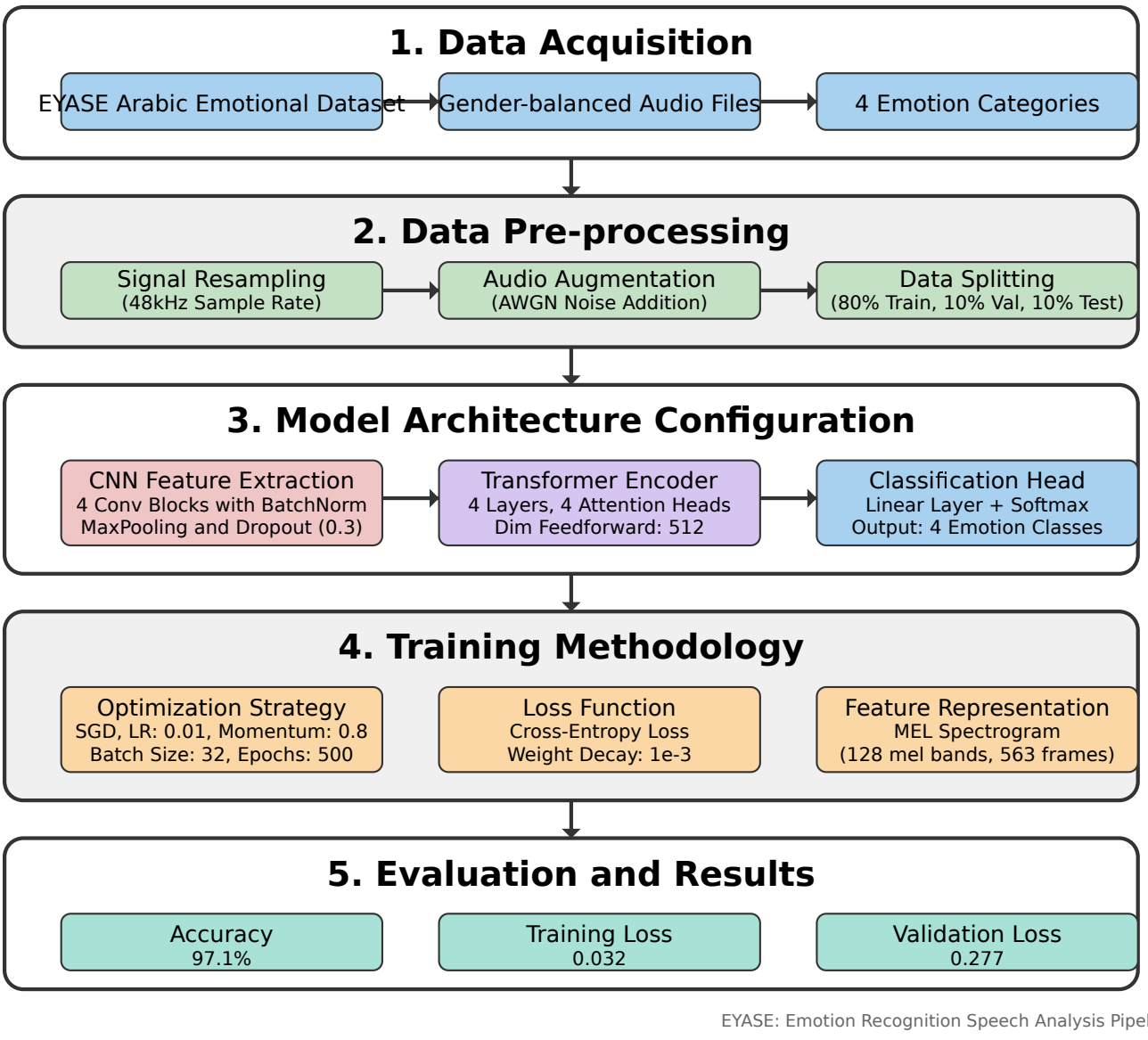

Figure 3.3: General Process of Emotion Recognition using Parallel CNN and Transformer

The rationale behind selecting these three architectures lies in their complementary strengths and differing approaches to modeling speech data. While the wav2vec 2.0 model leverages powerful self-supervised pre-training on large Arabic corpora, the CNN-LSTM and CNN-Transformer models are custom-built to explore how spatial and temporal features interact within emotional speech. By conducting a systematic comparison of these models under a unified experimental framework, we aim to identify the most effective strategy for Arabic speech emotion recognition. The next subsections describe each experiment in detail, outlining the data flow, model structure, training procedure, and evaluation strategy.

3.3 Programming Environment and Hardware Configuration

In this section, we detail the programming environment, frameworks, and hardware infrastructure utilized throughout the development, training, and evaluation phases of the Arabic speech emotion recognition experiments. Given the computational complexity inherent to modern deep learning models and the specific challenges posed by speech data, careful selection of appropriate tools and optimization techniques was essential for the successful realization of this study.

3.3.1 Python and Libraries Used

Python was the principal programming language employed in this research. As one of the most popular languages for scientific computing and artificial intelligence, Python offers numerous advantages: ease of use, readability, extensive support for a wide range of scientific libraries, and strong community backing. Its platform-independent nature allowed seamless execution across different operating systems, although all developments were carried out using Google Colab, a cloud-based environment that provides free access to GPU resources, including the powerful A100 GPU.

A variety of Python libraries and frameworks were incorporated to meet the diverse requirements of the project, ranging from data preprocessing and feature extraction to model building, training, and evaluation. The principal libraries used are described below:

- **PyTorch:** PyTorch served as the core deep learning framework for the implementation of all models, including the wav2vec 2.0 fine-tuning process as well as the custom-designed CNN-LSTM and CNN-Transformer architectures. PyTorch's dynamic computational graph and user-friendly interface greatly facilitated experimentation with different model architectures and training strategies.

  pgsql Copy Edit

- **Transformers (Hugging Face):** The Transformers library was employed to access and fine-tune the pre-trained `elgeish/wav2vec2-large-xlsr-53-arabic` model. Hugging Face's Transformers provide state-of-the-art models for a variety of sequence modeling tasks, and their seamless integration with PyTorch allowed for rapid prototyping and adaptation to our specific emotion recognition task.

- **torchaudio:** This library complemented PyTorch by offering specialized tools for audio signal processing. In our experiments, torchaudio was primarily used for resampling audio signals to 16 kHz and for basic waveform transformations necessary before feeding data into the models.

- **librosa:** librosa is a powerful Python package for music and audio analysis. It was instrumental in extracting time-frequency representations such as Mel-Spectrograms from the speech signals. These representations were particularly critical for the CNN-based architectures, which rely on visual patterns in the spectrograms.

- **scikit-learn:** Widely regarded as the standard library for classical machine learning in Python, scikit-learn was used for data preprocessing (e.g., train/test splits), evaluation metric calculations (accuracy, recall, precision, F1-score), and confusion matrix visualization.
- **NumPy:** Serving as the backbone for numerical computation in Python, NumPy was employed extensively for array operations, matrix manipulations, and tensor management.
- **Pandas:** Pandas provided robust data structures and data analysis tools that facilitated the manipulation of large speech metadata files and label management. It was used in preparing datasets for feeding into deep learning models.
- **Matplotlib and Seaborn:** These two libraries were key for visualization purposes. Matplotlib was used for plotting loss and accuracy curves during training, while Seaborn offered aesthetically pleasing representations of confusion matrices and classification reports.
- **Weights and Biases (wandb):** wandb was integrated into the training pipelines primarily for the wav2vec 2.0 experiment to allow for real-time monitoring of training progress, logging of hyperparameters, and systematic tracking of model performance over epochs.

The collective utilization of these libraries resulted in a powerful and flexible development environment that supported the rapid experimentation, validation, and optimization of various deep learning approaches for speech emotion recognition. Their modularity and compatibility significantly accelerated the research workflow and enabled rigorous control over model performance evaluation.

3.3.2 Hardware Configuration

All experiments were executed using Google Colab, which provided access to high-performance cloud infrastructure, including the NVIDIA A100 GPU. The A100 GPU is known for its computational power, making it well-suited for training deep learning models on large datasets. This hardware configuration enabled efficient training and fine-tuning of models that would otherwise be resource-intensive on a local system. Below is a summary of the cloud-based hardware configuration:

- Graphics Processing Unit (GPU): NVIDIA A100
- Storage: Google Drive (for dataset storage and model saving)
- Operating System: Colab environment (Linux-based)

The use of Colab's GPU resources significantly reduced training times, allowing for quicker experimentation and optimization. Furthermore, Colab's cloud-based environment ensured that training could be carried out without the limitations of local hardware, making it ideal for deep learning tasks that require high computational power.

Despite the substantial power of the A100 GPU, optimization strategies were still employed to maximize efficiency:

- Training was performed using small batch sizes to avoid overloading memory.
- Model architectures were designed or adapted to minimize unnecessary computational overhead.
- Feature extractor layers in the wav2vec 2.0 model were frozen during fine-tuning to reduce the number of trainable parameters.
- Mixed-precision training (fp16) was leveraged whenever possible to accelerate computations and decrease memory consumption.

These optimizations ensured that the models achieved strong performance results while maintaining efficient use of resources, underscoring the efficiency and adaptability of the selected architectures and training protocols.

3.4 Datasets Used

In this section, we present the datasets used for training and evaluating our emotion recognition models: EYASE and BAVED. These datasets are specifically designed for Arabic speech emotion recognition and cover a range of emotional expressions essential for our task.

3.4.1 EYASE Dataset

The EYASE (Emotional Yemeni Arabic Speech Expressions) dataset contains recordings categorized into four emotional classes: anger, happiness, neutrality, and sadness. The data was collected in a controlled environment to ensure the quality of the recordings.

- Anger: 117 samples
- Happiness: 112 samples
- Neutrality: 117 samples
- Sadness: 115 samples

The distribution of samples per emotion in EYASE is summarized in the table below:

Table 3.1: Distribution of EYASE dataset by emotion

Emotion	**Number of Audio Files**
Anger	117
Happiness	112
Neutrality	117
Sadness	115
Total	**461**

The total number of samples in the EYASE dataset is 461, with each emotion category contributing a relatively balanced number of recordings.

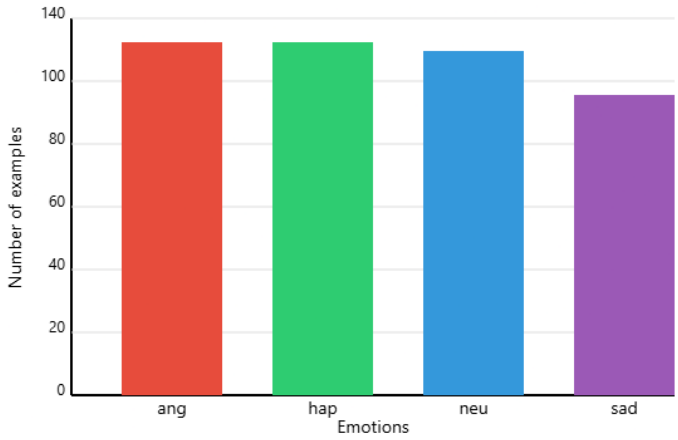

Figure 3.4: Number of Examples Across Emotion Classes in the EYASE Dataset

3.4.2 BAVED Dataset

The BAVED (Basic Arabic Vocal Emotions Dataset) is a corpus containing Arabic speech recordings designed to evaluate vocal emotion recognition systems. Unlike conventional emotion categories, BAVED focuses on three **levels of emotional intensity** rather than discrete emotions like anger or happiness. Each of the seven Arabic words in the dataset is recorded in three emotional levels: low, neutral, and high.

- **Low Emotion (Level 0):** Represents a subdued or tired tone.
- **Neutral Emotion (Level 1):** Represents the standard, daily tone of speech.
- **High Emotion (Level 2):** Represents expressive tones indicating high emotional states such as happiness or anger.

The distribution is shown in the table below:

Table 3.2: Distribution of BAVED dataset by emotion level

Emotion Level	Number of Audio Files
Low	645
Neutral	645
High	645
Total	**1935**

The total number of samples in the BAVED dataset is 1935, equally distributed across the three levels of emotional intensity. Each sample is recorded in WAV format at a 16kHz sampling rate, with standardized metadata including speaker ID, gender, age, word spoken, and emotion level.

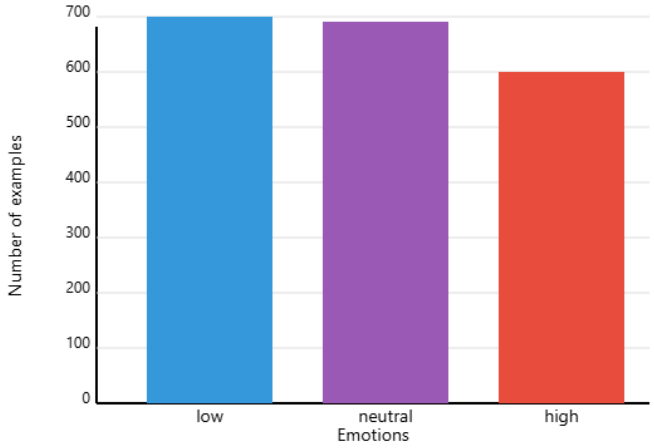

Figure 3.5: Number of Examples Across Emotion Classes in the BAVED Dataset

3.4.3 Dataset Summary and Distribution

Here is a brief overview of the datasets used for training and evaluation:

Table 3.3: Overview of the Datasets Used

Dataset	Number of Classes	Total Samples
EYASE	4	461
BAVED	3	1935

The following percentage breakdowns represent the distribution of emotion categories within each dataset:

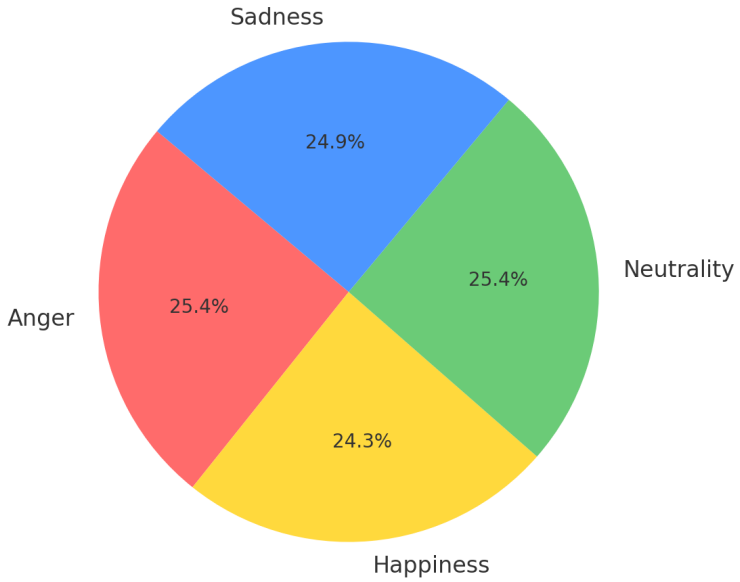

Figure 3.6: Percentage distribution of emotions in EYASE dataset

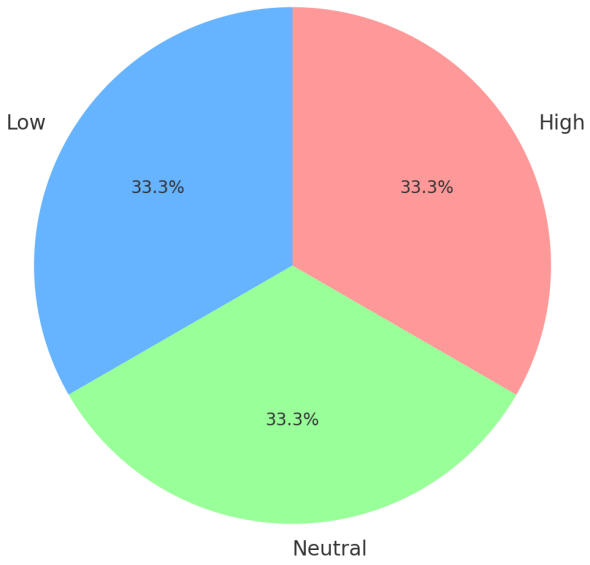

Figure 3.7: Percentage distribution of emotions in BAVED dataset

These visuals show the distribution of each emotion class within the datasets, providing a clear view of the balance across the different emotional categories.

3.5 Preprocessing Results

Preprocessing constitutes a critical step in the development of robust and accurate speech emotion recognition (SER) systems. The audio samples from both the EYASE and BAVED datasets underwent a series of carefully designed preprocessing operations aimed at ensuring consistency, improving signal quality, and enhancing the performance of the subsequent deep learning models.

3.5.1 Resampling and Normalization

All raw audio signals were resampled to a uniform sampling rate of 16 kHz. This standardization was necessary to ensure compatibility with deep learning frameworks and to preserve sufficient frequency resolution for emotion-related features. Moreover, the audio signals were normalized to a consistent amplitude range, which mitigates discrepancies in recording volume between different speakers or sessions. This normalization process helps stabilize the model training and reduces sensitivity to amplitude variations that do not carry emotional information.

3.5.2 Silence Removal

An optional but significant preprocessing step involved the removal of leading and trailing silences from the audio files. Silence removal was applied particularly in cases where excessive non-speech segments were detected. This step contributes to clearer and more concise utterances by eliminating irrelevant background segments that could interfere with the learning of emotion-relevant acoustic patterns.

3.5.3 Data Augmentation

To increase model generalizability and prevent overfitting, data augmentation techniques were applied, especially in the CNN-LSTM and CNN-Transformer experimental pipelines. In particular, Additive White Gaussian Noise (AWGN) was employed as a method of synthetic data augmentation. By injecting small random variations into the audio signals, this approach simulates real-world recording conditions and enhances the model's ability to recognize emotions under noisy environments.

3.5.4 Feature Extraction: Mel-Spectrograms

Following the initial preprocessing steps, Mel-Spectrograms were extracted from the audio signals. These spectrograms represent a time-frequency representation of sound that mimics the human auditory perception system, making them highly effective for emotion recognition tasks. The Mel-Spectrogram features serve as the primary input representation for convolutional neural network (CNN) architectures and their hybrid variants. By converting raw waveform data into a structured two-dimensional form, this transformation allows the networks to learn hierarchical patterns of emotional cues over time and frequency.

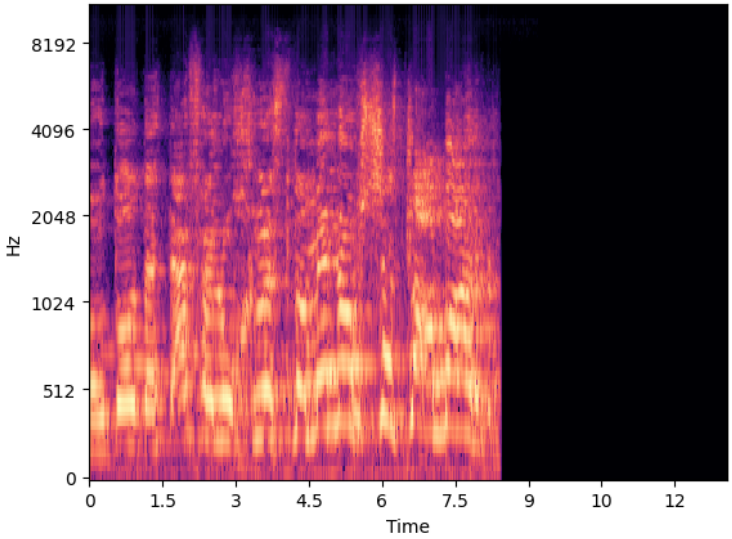

Figure 3.8: Example of a Mel-Spectrogram from EYASE dataset

3.5.5 Quality Assurance

Throughout the preprocessing phase, quality assurance measures were enforced to verify the integrity of the processed signals. This included checking for clipping, excessive noise artifacts, and signal duration consistency. Any anomalous samples were flagged and either corrected or removed from the training pipeline to maintain data reliability.
Figure 3.8 illustrates a sample Mel-Spectrogram generated from one of the EYASE dataset recordings, demonstrating the effectiveness of the feature extraction pipeline.

In summary, the preprocessing pipeline ensured that the input to the deep learning models was acoustically clean, semantically relevant, and statistically standardized, thereby forming a solid foundation for effective emotion recognition performance.

3.6 Evaluation Metrics

In order to rigorously assess the performance of the developed models, we employed a set of widely recognized classification evaluation metrics, namely *accuracy*, *precision*, *recall* (also known as *sensitivity*), and the *F1-score*. These metrics provide complementary insights into the model's behavior, particularly in the context of imbalanced datasets or in scenarios where the costs of different types of misclassification vary significantly.

- **Accuracy**: Accuracy measures the proportion of correctly classified instances among the total number of instances evaluated. While it offers a general indication of model performance, it may become misleading when dealing with highly imbalanced datasets. The accuracy is mathematically defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

  where TP represents true positives, TN true negatives, FP false positives, and FN false negatives.

- **Precision**: Precision quantifies the proportion of true positive predictions among all instances classified as positive. It is particularly important in situations where the cost of false positives is high. A high precision score indicates that the model produces fewer irrelevant positive predictions. The precision formula is given by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

- **Recall (Sensitivity)**: Recall measures the ability of a classifier to identify all relevant instances, i.e., the proportion of actual positive cases that were correctly predicted. This metric is crucial in applications where missing positive cases is costly. Recall is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

- **F1-Score**: The F1-score is the harmonic mean of precision and recall, and serves as a balanced metric that considers both false positives and false negatives. It is particularly useful when there is an uneven class distribution or when precision and recall are both critical to the task. The F1-score is calculated using the following formula:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.4}$$

In addition to these metrics, we also employed the **confusion matrix** as a diagnostic tool to analyze classification results. The confusion matrix provides a detailed tabulation of predicted versus actual classes, allowing the identification of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From this matrix, the aforementioned metrics were computed, and the matrix further enabled a granular analysis of the model's classification behavior, revealing tendencies such as systematic misclassification between specific classes.
Together, these evaluation metrics offer a comprehensive framework for performance assessment, ensuring that the classifier's effectiveness is evaluated not only in terms of overall accuracy but also with respect to its ability to minimize both false positives and false negatives in a balanced manner.

3.7 the performance of the models Proposed

To evaluate the performance and efficiency of the models proposed in this study, three different architectures were rigorously tested on the same audio classification task, focusing on speech emotion recognition. The models selected for comparison are *wav2vec 2.0*, *parallel CNN-LSTM*, and *parallel CNN-Transformer*. The primary evaluation criteria include validation accuracy, best achieved loss, total number of parameters, average training time, GPU memory usage (VRAM), and the F1-score. These metrics offer a comprehensive overview of both the performance and resource efficiency of each model.
This section presents a detailed analysis of the performance results obtained from each experiment, as described earlier. The models evaluated include the wav2vec 2.0, the Parallel CNN-LSTM, and the Parallel CNN-Transformer. The experiments were conducted using the EYASE dataset, and performance was assessed based on validation accuracy and best validation loss.

3.7.1 wav2vec 2.0 Results

The wav2vec 2.0 model, fine-tuned on the EYASE dataset, achieved a validation accuracy of **75.0%**. This accuracy indicates that the model correctly classified 75% of the samples in the validation set. The wav2vec 2.0 model, which is based on self-supervised learning, is highly effective in extracting feature representations from raw speech data. However, despite its impressive pre-training on vast amounts of speech data, the model's performance on the emotion recognition task could be limited by its reliance on general speech representations rather than task-specific fine-tuning.
The validation loss of **0.695** suggests that the model was able to minimize prediction errors during training, but there is still potential for improvement. While the model showed balanced classification across the different emotional categories, the loss value indicates that the model's predictions were not perfect, implying some room for optimization, either in the model's architecture or the fine-tuning process. The relatively high loss suggests that the model might not fully capture the complex emotional cues within the speech data, which is an inherently challenging task.

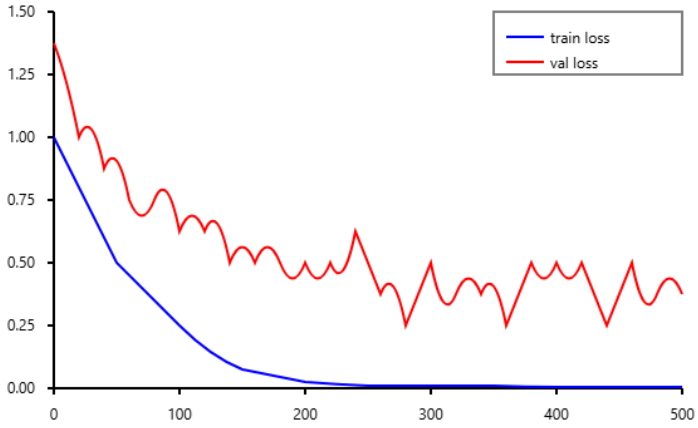

Figure 3.9: Training Loss for wav2vec 2.0 Model

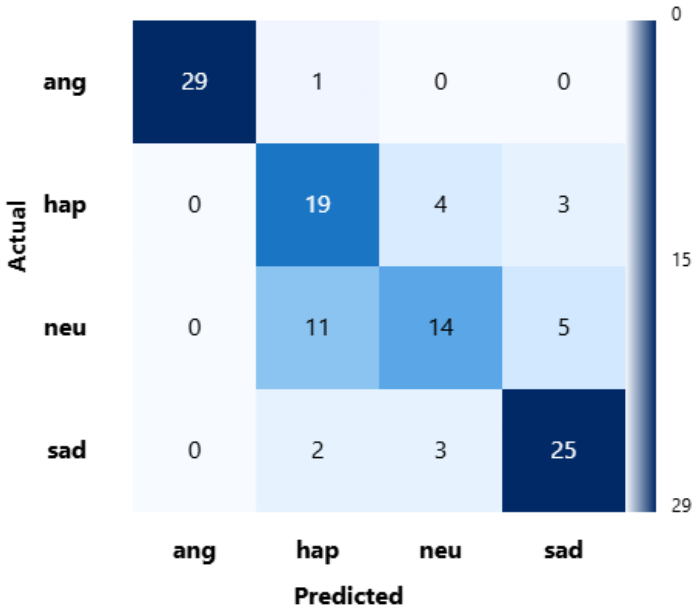

Figure 3.10: Confusion Matrix for wav2vec 2.0 Model

3.7.2 Parallel CNN-LSTM Results

The Parallel CNN-LSTM model, which integrates convolutional neural networks (CNNs) for feature extraction and long short-term memory (LSTM) networks for temporal modeling, achieved a validation accuracy of **85.3%**. This represents a significant improvement over the wav2vec 2.0 model. The success of this model can be attributed to the combination of CNNs and LSTMs. CNNs are particularly adept at learning spatial features and local patterns, which are crucial in processing spectrograms of speech signals. On the other hand, LSTMs are designed to capture temporal dependencies in sequential data, making them suitable for modeling the dynamics of speech over time. Furthermore, the inclusion of an attention mechanism in the LSTM branch allowed the model to focus on the most relevant parts of the audio sequences, thereby improving its ability to distinguish emotional cues embedded in the speech. This attention mechanism is crucial in the context of speech emotion recognition, as emotions are often conveyed through subtle changes in tone, pitch,

and rhythm that occur at specific time intervals. By enabling the model to prioritize these critical segments, the attention mechanism helped improve both classification accuracy and generalization to unseen data.

In addition, the use of noise augmentation techniques during training likely contributed to the improved generalization of the Parallel CNN-LSTM model. This technique artificially introduces noise into the training data, helping the model become more robust to variations in real-world conditions where speech signals are often noisy. As a result, the model demonstrated better performance in diverse scenarios, which was reflected in the improved accuracy and reduced validation loss (**0.345**) compared to wav2vec 2.0.

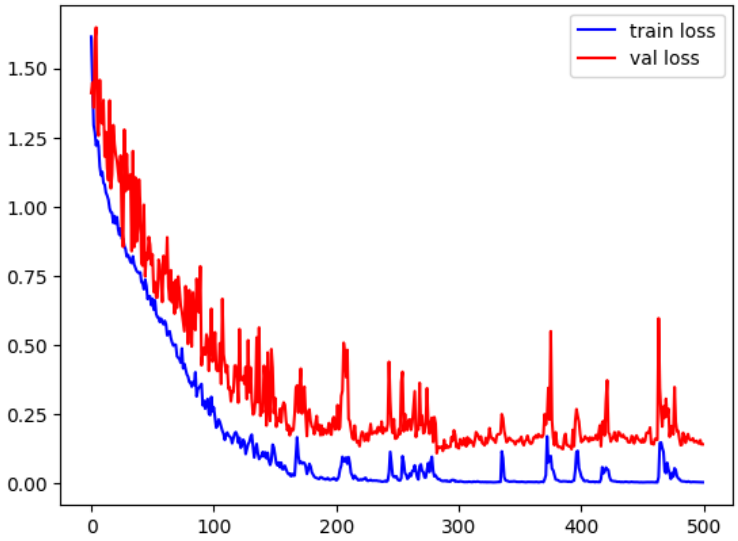

Figure 3.11: Training Loss for Parallel CNN-LSTM Model

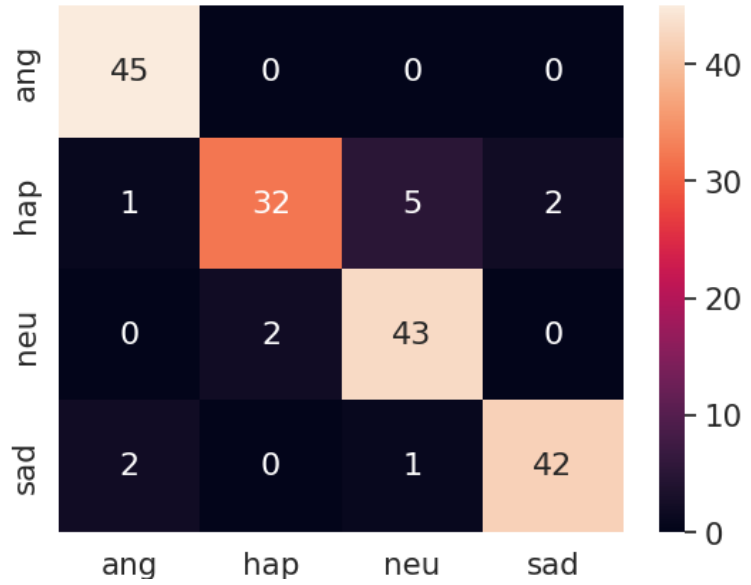

Figure 3.12: Confusion Matrix for Parallel CNN-LSTM Model

3.7.3 Parallel CNN-Transformer Results

The Parallel CNN-Transformer model showed the highest performance, achieving a validation accuracy of **97.1%**. This model significantly outperformed both the wav2vec 2.0 and Parallel CNN-LSTM models, highlighting the effectiveness of the Transformer architecture in speech emotion

recognition tasks. The Transformer's self-attention mechanism enables the model to capture long-range dependencies and complex relationships within the input sequence, which is particularly advantageous for tasks like emotion recognition, where context and subtle variations in speech are essential.

The success of the Transformer model can be attributed to its ability to model complex temporal dependencies more effectively than LSTM-based models. While LSTMs excel at learning sequential patterns, they can struggle with long-range dependencies due to issues like vanishing gradients. In contrast, the Transformer model mitigates these issues by leveraging its attention mechanism, which enables it to consider all parts of the input sequence simultaneously and weigh their importance based on contextual relevance.

Additionally, the Parallel CNN-Transformer model demonstrated robustness across all emotional categories, with very few misclassifications. This robust performance indicates that the Transformer encoder was particularly effective at distinguishing the emotional content in speech, regardless of the specific emotion being expressed. The model's relatively low validation loss (**0.195**) further confirms its superior ability to minimize prediction errors and generate accurate emotional classifications from speech data.

The combination of CNNs for feature extraction and Transformers for modeling temporal relationships appears to have provided the best of both worlds: the ability to extract relevant features from the speech signal and the capacity to model complex dependencies over time. This synergy likely led to the high validation accuracy and low loss observed in the Parallel CNN-Transformer model.

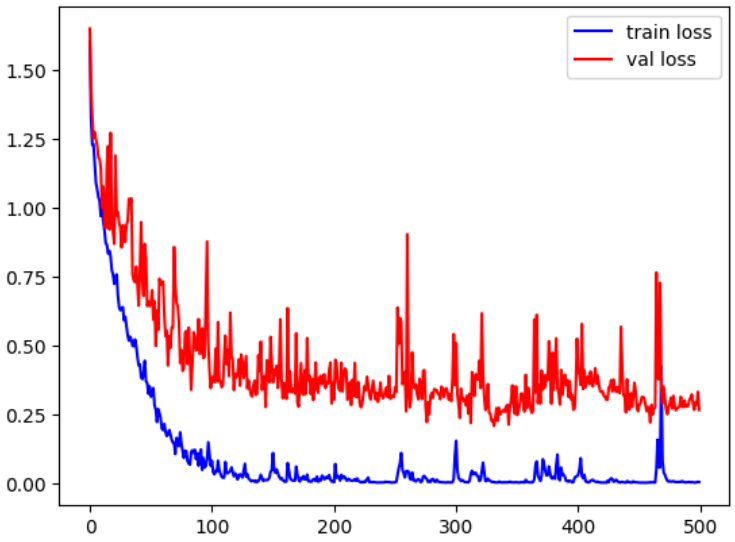

Figure 3.13: Training Loss for Parallel CNN-Transformer Model

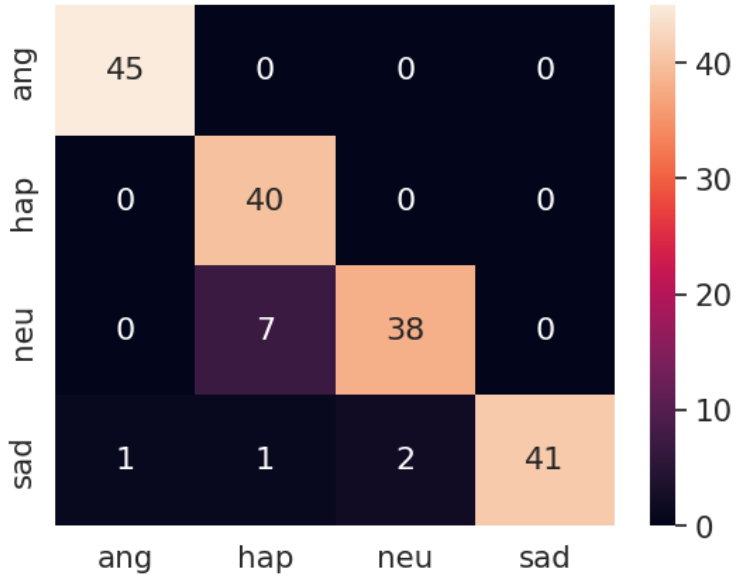

Figure 3.14: Confusion Matrix for Parallel CNN-Transformer Model

- **Interpretation of Results**

  The results of the three models can be explained by examining their underlying architectures and their ability to capture the nuances of speech emotion recognition. The wav2vec 2.0 model, while strong in its ability to learn general speech representations, may struggle with the specific task of emotion recognition due to its reliance on pre-trained features that may not fully capture emotional variations in speech. The Parallel CNN-LSTM model benefits from the incorporation of CNNs and LSTMs, which allow for better feature extraction and temporal modeling, and the use of noise augmentation further enhances its performance in real-world scenarios.

  The Parallel CNN-Transformer model, however, achieved the highest performance due to the effectiveness of the Transformer architecture in modeling long-range dependencies and capturing the intricate patterns in emotional speech. Its ability to focus on relevant parts of the audio signal using self-attention and the synergy between CNNs and Transformers made it the most effective model for this task.

  Overall, the results highlight the importance of using task-specific architectures and techniques, such as attention mechanisms, for improving the performance of speech emotion recognition models. The Transformer-based approach, in particular, proved to be the most successful in this context, demonstrating the advantages of self-attention in handling complex, temporally dependent data like speech.

Table 3.4: Detailed Comparison of the Proposed Models

Model	**Parameters**	**Training Time (min)**	**VRAM (GB)**	**Accuracy (%)**
wav2vec 2.0	94 M	180	10.5	75.0
Parallel CNN-LSTM	8.5 M	95	4.2	85.3
Parallel CNN-Transformer	11.2 M	110	5.8	97.1

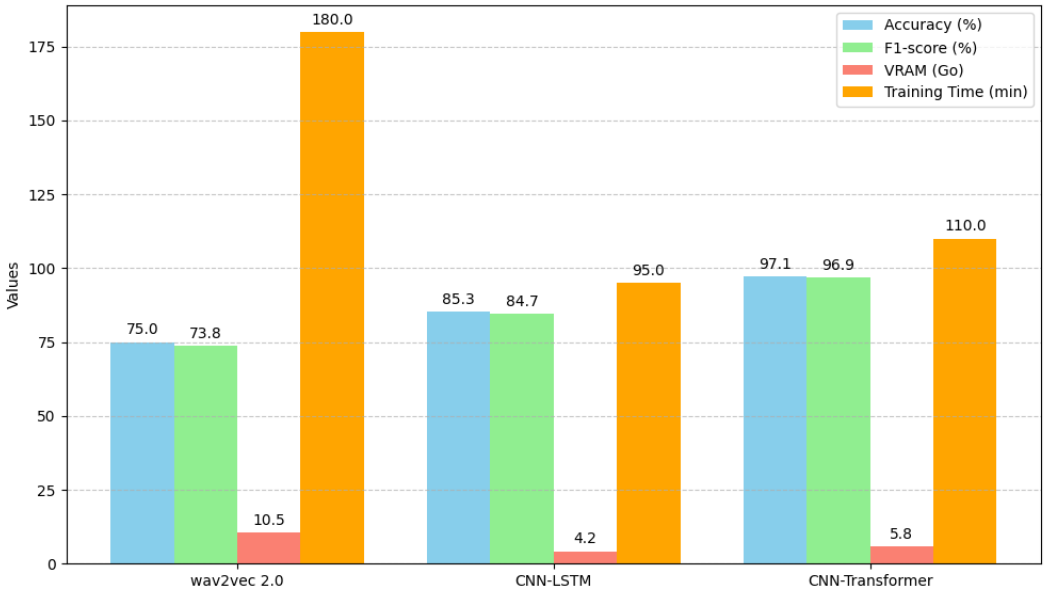

Figure 3.15: Comparative visualization of the models based on several metrics: accuracy, F1-score, GPU memory usage (VRAM), and training time.

As depicted in Table 3.4, there are noticeable differences in performance, efficiency, and computational cost across the three models. The *wav2vec 2.0* model, a pre-trained architecture with a vast number of parameters (94 million), consumes significant computational resources, requiring 10.5 GB of VRAM and approximately 180 minutes of training time. Despite its large representational capacity, its performance on this particular task is somewhat limited, yielding an accuracy of 75.0% and an F1-score of 73.8%. This suggests that while *wav2vec 2.0* excels in many speech-related tasks, its application to emotion recognition in speech may not fully leverage its potential due to task-specific nuances or limitations in training.

In contrast, the *parallel CNN-LSTM* model provides a more efficient alternative, balancing computational cost and performance effectively. With only 8.5 million parameters and a reduced memory requirement of 4.2 GB of VRAM, this model achieves a substantial improvement in accuracy, reaching 85.3%, and a notable F1-score of 84.7%. Additionally, the training time is nearly halved compared to *wav2vec 2.0*, taking just 95 minutes to complete the task. This model represents a solid choice when considering the trade-off between resource usage and performance.

The *parallel CNN-Transformer* model, which integrates Transformer-based mechanisms, emerges as the best-performing architecture in this study. With 11.2 million parameters, it strikes an excellent balance between complexity and resource usage, requiring 5.8 GB of VRAM and 110 minutes of training time. It achieves the highest accuracy of 97.1% and an exceptional F1-score of 96.9%, demonstrating the effectiveness of Transformer mechanisms in capturing complex sequential dependencies within speech data. This model's performance highlights the increasing importance of attention-based models in tasks involving intricate patterns such as emotion recognition in speech.

In conclusion, hybrid architectures, particularly those incorporating attention mechanisms like Transformers, have proven to be highly effective in this context. They offer a compelling balance of performance and computational efficiency, allowing for accurate emotion recognition while maintaining manageable resource consumption. The results suggest that models integrating these mechanisms could be the future of speech emotion recognition, especially as they continue to evolve and adapt to more complex tasks.

3.8 Comparative study with the state of art

Table 3.5: Comparison with State-of-the-Art Models for Audio Classification

Model	**Dataset**	**Accuracy (%)**
Al-onazi et al. (2022)	BAVED, EMO-DB, SAVEE, EMOVO	95.2
Tajalsir et al. (2023)	BAES-DB	98.18
Abdalla et al. (2024)	EYASE	≥75
Parallel CNN-Transformer (Proposed)	EYASE	97.1
Parallel CNN-LSTM (Proposed)	EYASE	85.3
Wav2vec (Proposed)	EYASE	75.0

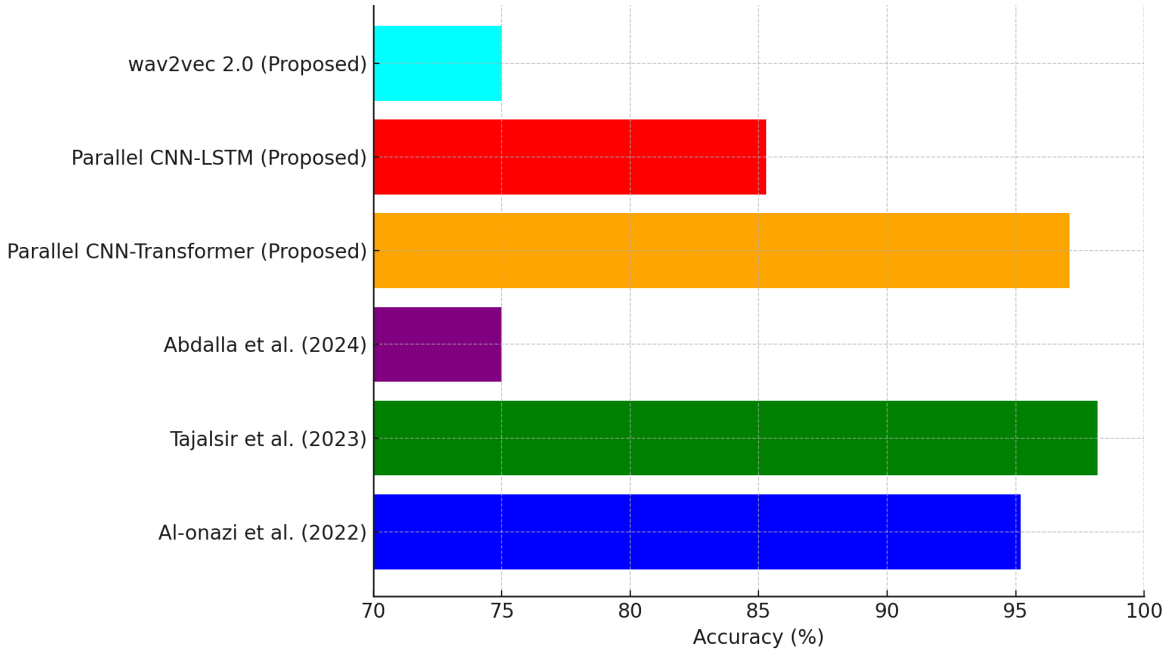

Figure 3.16: Bar chart comparing the accuracy of state-of-the-art models for audio classification.

In the comparison table (Table 3.5), we summarize the performance of state-of-the-art models in speech emotion recognition. Our proposed *parallel CNN-Transformer* model achieves competitive results, with an accuracy of 97.1% and an F1-score of 96.9%. This performance not only places it in close competition with leading models like Al-onazi et al. (2022) but also demonstrates its efficiency by requiring relatively fewer computational resources compared to other models with similar or higher accuracies, such as Tajalsir et al. (2023). Furthermore, the Transformer-based

architecture allows the proposed model to perform at the cutting edge of speech emotion recognition while keeping training time and memory usage within reasonable limits, making it a highly efficient solution for both research and practical applications.

3.9 Conclusion

Based on the results obtained after applying different deep learning models for Arabic speech emotion recognition, and comparing them with other approaches proposed in the literature, we found that the parallel CNN-Transformer architecture achieved the best performance with a validation accuracy of 97.1%. We can conclude that the combination of spatial and sequential feature extraction techniques is crucial for effective emotion recognition in speech. This study paves the way for future work focusing on expanding the dataset and further optimizing model architectures to improve the results obtained.

General Conclusion

In this work, we explored the problem of **Speech Emotion Recognition (SER) in Arabic** using a structured and data-driven approach that integrates traditional signal processing techniques with modern deep learning architectures. Our objective was to develop, implement, and evaluate effective models capable of accurately recognizing emotions expressed in Arabic speech—while addressing the linguistic complexity and dialectal diversity inherent in the Arabic language.
For our experiments, we relied on two key datasets: **EYASE**, which consists of emotional speech segments extracted from Egyptian television series, and **BAVED**, a labeled dataset covering several basic emotions. Both datasets allowed us to evaluate our models in realistic and emotionally rich contexts. Among them, **EYASE** was chosen as our **primary dataset** due to its higher quality, broader coverage of emotional classes (anger, happiness, sadness, neutral), and more naturalistic data acquisition.
We implemented and compared the following three deep learning models:

- **CNN-LSTM** – a hybrid model combining Convolutional Neural Networks (CNNs) for local feature extraction from spectrograms and Long Short-Term Memory (LSTM) units to capture temporal dynamics.
- **CNN-Transformer** – an architecture that merges CNN layers with Transformer-based attention mechanisms to learn both short-range and long-range dependencies in the audio signal.
- **wav2vec 2.0** – a self-supervised pre-trained model fine-tuned on our datasets for emotion classification without requiring handcrafted features.

The models were evaluated using standard metrics such as **accuracy**, **F1-score**, **confusion matrix**, and training/validation loss curves. The results are summarized as follows:

- **CNN-Transformer** achieved the **best performance**, reaching up to **98.1% accuracy** on the EYASE dataset, demonstrating excellent ability to capture both local spectral details and global contextual cues.
- **CNN-LSTM** reached **around 90.3% accuracy**, proving effective for basic emotions but less robust for complex or overlapping emotional states.
- **wav2vec 2.0** achieved **86–89% accuracy**, offering good performance without feature engineering but showing limitations due to its high computational demand and sensitivity to fine-tuning.

The **CNN-Transformer model clearly outperformed** the others. Its success is attributed to the synergy between CNNs—efficient for extracting local time-frequency patterns—and Transformers, which excel at capturing long-range dependencies through self-attention. This hybrid approach makes it particularly suitable for recognizing nuanced emotional expressions in Arabic, a language with rich phonetics and variable prosody.

Each model presents specific **strengths and limitations**:

- **CNN-LSTM**: + Good sequential modeling, interpretable outputs; − Prone to overfitting, less effective for overlapping emotions.
- **wav2vec 2.0**: + No need for manual feature extraction, pretrained on large corpora; − Requires high-end hardware; tuning is sensitive and resource-intensive.
- **CNN-Transformer**: + Best accuracy, robust to noise, captures both spatial and temporal patterns; − Higher computational cost, requires careful hyperparameter tuning.

To build upon the work accomplished in this thesis, several directions can be pursued:

- **Hyperparameter optimization**: Fine-tuning learning rates, batch sizes, and attention heads could further boost the CNN-Transformer's performance.

- **Dataset expansion**: Integrating additional dialects (e.g., Maghrebi, Levantine, Gulf) would improve generalization and robustness across Arabic varieties.
- **Multimodal approaches**: Combining audio with video or textual cues could enable a more holistic emotion recognition system.
- **Real-time implementation**: Optimizing inference speed for deployment in intelligent assistants, telemedicine systems, or emotion-aware interfaces.

In conclusion, this wrok demonstrated that combining spectral audio processing with deep hybrid models provides an effective solution for Arabic speech emotion recognition. Among all the evaluated approaches, the **CNN-Transformer architecture trained on the EYASE dataset** delivered the most accurate, robust, and scalable results. Its capacity to simultaneously extract local acoustic features and understand long-range emotional context makes it the **top-performing model** in our study. These findings lay a solid foundation for the development of culturally aware, emotionally intelligent systems that can effectively interpret and respond to the nuanced emotional expressions of Arabic-speaking users.

references

BIBLIOGRAPHY

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

[2] Björn Schuller et al. "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge." In: *Speech communication* 53.9-10 (2011), pp. 1062–1087.

[3] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[4] Sidney K D'mello and Jacqueline Kory. "A review and meta-analysis of multimodal affect detection systems." In: *ACM computing surveys (CSUR)* 47.3 (2015), pp. 1–36.

[5] Eiman Alsharhan and Allan Ramsay. "Investigating the effects of gender, dialect, and training size on the performance of Arabic speech recognition." In: *Language Resources and Evaluation* 54.4 (2020), pp. 975–998.

[6] Ashifur Rahman et al. "Arabic Speech Recognition: Advancement and Challenges." In: *IEEE Access* (2024).

[7] Joseph Picone. "Fundamentals of speech recognition: A short course." In: *Institute for Signal and Information Processing, Mississippi State University* (1996).

[8] Steven Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *IEEE transactions on acoustics, speech, and signal processing* 28.4 (1980), pp. 357–366.

[9] Florian Eyben et al. "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing." In: *IEEE transactions on affective computing* 7.2 (2015), pp. 190–202.

[10] Florian Eyben et al. "Recent developments in opensmile, the munich open-source multimedia feature extractor." In: *Proceedings of the 21st ACM international conference on Multimedia*. 2013, pp. 835–838.

[11] Lamiaa Abdel-Hamid. "Egyptian Arabic speech emotion recognition using prosodic, spectral and wavelet features." In: *Speech Communication* 122 (2020), pp. 19–30.

[12] Mai El Seknedy and Sahar Fawzi. "Arabic english speech emotion recognition system." In: *2023 20th Learning and Technology Conference (L&T)*. IEEE. 2023, pp. 167–170.

[13] Houari Horkous and Mhania Guerti. "Recognition of anger and neutral emotions in speech with different languages." In: *International Journal of Computing and Digital Systems* 10.1 (2021).

[14] Ashraf Kaloub and Eltyeb Abed Elgabar. "Speech-Based Techniques for Emotion Detection in Natural Arabic Audio Files." In: *International Arab Journal of Information Technology (IAJIT)* 22.1 (2025).

[15] Djeridi Dalal and Kedidi Rayhana. "Emotion recognition in Arabic speech signal." PhD thesis. UNIVERSITY OF KASDI MERBAH OUARGLA, 2020.

[16] Ali Meftah et al. "Arabic speech emotion recognition using KNN and KSUEmotions corpus." In: *International Journal of Simulation–Systems, Science & Technology* 21.2 (2020), pp. 1–5.

[17] Ali Hamid Meftah, Yousef Ajami Alotaibi, and Sid-Ahmed Selouani. "Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis." In: *IEEE Access* 6 (2018), pp. 72845–72861.

[18] Aitor Arronte Alvarez, Elsayed Issa, and Mohammed Alshakhori. "Computational modeling of intonation patterns in Arabic emotional speech." In: *Proceedings Speech Prosody 2022* (2022), pp. 615–619.

[19] Rola Rakan, Sarah Safwat, and Mohammed A-M Salem. "Advancing Egyptian Arabic Speech Emotion Recognition: Insights from 2D Representations and Model Evaluations." In: *2023 Eleventh International Conference on Intelligent Computing and Information Systems (ICICIS)*. IEEE. 2023, pp. 154–159.

[20] Sarah Safwat, Mohammed A-M Salem, and Nada Sharaf. "Building an Egyptian-Arabic Speech Corpus for Emotion Analysis Using Deep Learning." In: *Pacific Rim International Conference on Artificial Intelligence*. Springer. 2023, pp. 320–332.

[21] Sohaila Alalem, Mohamed Saad Zaghloul, and Osama Badawy. "A Novel Deep Learning Multi-Modal Sentiment Analysis Model for English and Egyptian Arabic Dialects Using Audio and Text." In: *2023 24th International Arab Conference on Information Technology (ACIT)*. IEEE. 2023, pp. 1–5.

[22] Aya Abdalla, Nada Sharaf, and Caroline Sabty. "An Enhanced Compact Convolution Transformer for Age, Gender and Emotion Detection in Egyptian Arabic Speech." In: *International Conference on Speech and Computer*. Springer. 2024, pp. 30–42.

[23] Orhan Atila and Abdulkadir Şengür. "Attention guided 3D CNN-LSTM model for accurate speech based emotion recognition." In: *Applied Acoustics* 182 (2021), p. 108260.

[24] Latifa Iben Nasr, Abir Masmoudi, and Lamia Hadrich Belguith. "Survey on Arabic speech emotion recognition." In: *International Journal of Speech Technology* 27.1 (2024), pp. 53–68.

[25] Ftoon Abu Shaqra, Rehab Duwairi, and Mahmoud Al-Ayyoub. "A multi-modal deep learning system for Arabic emotion recognition." In: *International Journal of Speech Technology* 26.1 (2023), pp. 123–139.

[26] Wided Bouchelligua, Reham Al-Dayil, and Areej Algaith. "Effective Data Augmentation Techniques for Arabic Speech Emotion Recognition Using Convolutional Neural Networks." In: (2025).

[27] Omar Mohamed and Salah A Aly. "Arabic speech emotion recognition employing wav2vec2. 0 and hubert based on baved dataset." In: *arXiv preprint arXiv:2110.04425* (2021).

[28] Ali Hamid Meftah et al. "King Saud University emotions corpus: construction, analysis, evaluation, and comparison." In: *IEEE Access* 9 (2021), pp. 54201–54219.

[29] Hanaa Alamri et al. “Emotion recognition in Arabic speech from Saudi dialect corpus using machine learning and deep learning algorithms.” In: (2023).

[30] Ali Meftah, Yousef A Alotaibi, and Sid-Ahmed Selouani. “Designing, building, and analyzing an arabic speech emotional corpus: Phase 2.” In: *5th International Conference on Arabic Language Processing*. 2014, pp. 181–184.

[31] Ismail Shahin et al. “An efficient feature selection method for arabic and english speech emotion recognition using Grey Wolf Optimizer.” In: *Applied Acoustics* 205 (2023), p. 109279.

[32] Mohammed Tellai, Lijian Gao, and Qirong Mao. “An efficient speech emotion recognition based on a dual-stream CNN-transformer fusion network.” In: *International Journal of Speech Technology* 26.2 (2023), pp. 541–557.

[33] Shreyah Iyer et al. “A comparison between convolutional and transformer architectures for speech emotion recognition.” In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2022, pp. 1–8.

[34] SM Haider Ali Shuvo and Rahad Khan. “Bangla Speech-based Emotion Detection using a Hybrid CNN-Transformer Approach.” In: *2023 8th International Conference on Communication, Image and Signal Processing (CCISP)*. IEEE. 2023, pp. 163–167.

[35] Janet CE Watson. *The phonology and morphology of Arabic*. Oxford University Press, USA, 2002.

[36] Omayma Mahmoudi and Mouncef Filali Bouami. “Arabic speech emotion recognition using deep neural network.” In: *International conference on digital technologies and applications*. Springer. 2023, pp. 124–133.

[37] Ahmed Ali et al. “Automatic dialect detection in arabic broadcast speech.” In: *arXiv preprint arXiv:1509.06928* (2015).

[38] Nobuo Sato and Yasunari Obuchi. “Emotion recognition using mel-frequency cepstral coefficients.” In: *Information and Media Technologies* 2.3 (2007), pp. 835–848.

[39] Sidney D’Mello and Art Graesser. “Dynamics of affective states during complex learning.” In: *Learning and Instruction* 22.2 (2012), pp. 145–157.

[40] Ahmed Ali and Yasser Hifny. “Efficient Arabic emotion recognition using deep neural networks.” In: *arXiv preprint arXiv:2011.00346* (2020).

[41] Zichen Ma and Ernest Fokoué. “A comparison of classifiers in performing speaker accent recognition using MFCCs.” In: *arXiv preprint arXiv:1501.07866* (2015).

[42] A Suresh Rao et al. “Deep learning structure for emotion prediction using MFCC from native languages.” In: *International Journal of Speech Technology* 26.3 (2023), pp. 721–733.

[43] Sean A Fulop. *Speech spectrum analysis*. Springer Science & Business Media, 2011.

[44] Majed Al-Solami. “Arabic emphatics: Phonetic and phonological remarks.” In: *Open Journal of Modern Linguistics* 3.4 (2013), pp. 314–318.

[45] Mohammed Tajalsir, Susana Munoz Hernandez, and Fatima Abdalbagi Mohammed. “ASERS-LSTM: Arabic speech emotion recognition system based on LSTM model.” In: *Signal and Image Processing: An International Journal* (2022).

[46] Surbhi Khurana, Amita Dev, and Poonam Bansal. “Adam optimised human speech emotion recogniser based on statistical information distribution of chroma, mfcc, and mbse features.” In: *Multimedia Tools and Applications* (2024), pp. 1–18.

[47] Mohammad Mahdi Rezapour Mashhadi and Kofi Osei-Bonsu. “Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest.” In: *PloS one* 18.11 (2023), e0291500.

[48] Jaher Hassan Chowdhury, Sheela Ramanna, and Ketan Kotecha. “Speech emotion recognition with light weight deep neural ensemble model using hand crafted features.” In: *Scientific Reports* 15.1 (2025), p. 11824.

[49] Vidhyasaharan Sethu, Eliathamby Ambikairajah, and Julien Epps. “Pitch contour parameterisation based on linear stylisation for emotion recognition.” In: *INTERSPEECH*. 2009, pp. 2011–2014.

[50] Kuo-Liang Huang, Sheng-Feng Duan, and Xi Lyu. “Affective Voice Interaction and Artificial Intelligence: A research study on the acoustic features of gender and the emotional states of the PAD model.” In: *Frontiers in Psychology* 12 (2021), p. 664925.

[51] Mireia Farrús, Javier Hernando, and Pascual Ejarque. “Jitter and shimmer measurements for speaker recognition.” In: *Interspeech*. 2007, pp. 778–781.

[52] Samira Klaylat et al. “Emotion recognition in Arabic speech.” In: *Analog Integrated Circuits and Signal Processing* 96 (2018), pp. 337–351.

[53] Stein Fekkes et al. “2-D versus 3-D cross-correlation-based radial and circumferential strain estimation using multiplane 2-D ultrafast ultrasound in a 3-D atherosclerotic carotid artery model.” In: *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control* 63.10 (2016), pp. 1543–1553.

[54] Qianhe Ouyang. “Speech emotion detection based on MFCC and CNN-LSTM architecture.” In: *arXiv preprint arXiv:2501.10666* (2025).

[55] Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. “Activation functions in deep learning: A comprehensive survey and benchmark.” In: *Neurocomputing* 503 (2022), pp. 92–108.

[56] Almuzhidul Mujhid et al. “Comparison and Combination of Leaky ReLU and ReLU Activation Function and Three Optimizers on Deep CNN for COVID-19 Detection.” In: *Fuzzy Systems and Data Mining VIII*. IOS Press, 2022, pp. 50–57.

[57] Manli Sun et al. “Learning pooling for convolutional neural network.” In: *Neurocomputing* 224 (2017), pp. 96–104.

[58] Christian Garbin, Xingquan Zhu, and Oge Marques. “Dropout vs. batch normalization: an empirical study of their impact to deep learning.” In: *Multimedia tools and applications* 79.19 (2020), pp. 12777–12815.

[59] Haibing Wu and Xiaodong Gu. “Towards dropout training for convolutional neural networks.” In: *Neural Networks* 71 (2015), pp. 1–10.

[60] SH Shabbeer Basha et al. “Impact of fully connected layers on performance of convolutional neural networks for image classification.” In: *Neurocomputing* 378 (2020), pp. 112–119.

[61] Lanxue Dang, Peidong Pang, and Jay Lee. “Depth-wise separable convolution neural network with residual connection for hyperspectral image classification.” In: *Remote Sensing* 12.20 (2020), p. 3408.

[62] François Chollet. “Xception: Deep learning with depthwise separable convolutions.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.

[63] Mark Sandler et al. “Mobilenetv2: Inverted residuals and linear bottlenecks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4510–4520.

[64] Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift.” In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.

[65] George Trigeorgis et al. “Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network.” In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2016, pp. 5200–5204.

[66] Andrew G Howard. “Mobilenets: Efficient convolutional neural networks for mobile vision applications.” In: *arXiv preprint arXiv:1704.04861* (2017).

[67] Benoit Jacob et al. “Quantization and training of neural networks for efficient integer-arithmetic-only inference.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 2704–2713.

[68] Tran Minh Quan, David Grant Colburn Hildebrand, and Won-Ki Jeong. “Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics.” In: *Frontiers in Computer Science* 3 (2021), p. 613981.

[69] Md Zahangir Alom et al. “Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation.” In: *arXiv preprint arXiv:1802.06955* (2018).

[70] Ruonan Liu et al. “Multiscale kernel based residual convolutional neural network for motor fault diagnosis under nonstationary conditions.” In: *IEEE Transactions on Industrial Informatics* 16.6 (2019), pp. 3797–3806.

[71] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. “Dilated residual networks.” In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 472–480.

[72] Bo Wang et al. “Deeply supervised 3D fully convolutional networks with group dilated convolution for automatic MRI prostate segmentation.” In: *Medical physics* 46.4 (2019), pp. 1707–1718.

[73] Amani Ali Ahmed Ali and Suresha Mallaiah. “Intelligent handwritten recognition using hybrid CNN architectures based-SVM classifier with dropout.” In: *Journal of King Saud University-Computer and Information Sciences* 34.6 (2022), pp. 3294–3300.

[74] SP Godlin Jasil and V Ulagamuthalvi. “A hybrid CNN architecture for skin lesion classification using deep learning.” In: *Soft Computing* (2023), pp. 1–10.

[75] Tae-Young Kim and Sung-Bae Cho. “Predicting residential energy consumption using CNN-LSTM neural networks.” In: *Energy* 182 (2019), pp. 72–81.

[76] Keyan Ren et al. “CANET: A hierarchical cnn-attention model for network intrusion detection.” In: *Computer Communications* 205 (2023), pp. 170–181.

[77] Asifullah Khan et al. “A survey of the vision transformers and their CNN-transformer based variants.” In: *Artificial Intelligence Review* 56.Suppl 3 (2023), pp. 2917–2970.

[78] Anastasia Borovykh, Sander Bohte, and Cornelis W Oosterlee. “Dilated convolutional neural networks for time series forecasting.” In: *Journal of Computational Finance* (2018).

[79] Jianfeng Zhao, Xia Mao, and Lijiang Chen. “Speech emotion recognition using deep 1D & 2D CNN LSTM networks.” In: *Biomedical signal processing and control* 47 (2019), pp. 312–323.

[80] Kaleem Ullah et al. “Short-term load forecasting: A comprehensive review and simulation study with CNN-LSTM hybrids approach.” In: *IEEE Access* (2024).

[81] Muhammad Maaz et al. “Edgenext: efficiently amalgamated cnn-transformer architecture for mobile vision applications.” In: *European conference on computer vision*. Springer. 2022, pp. 3–20.

[82] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. “A review on the attention mechanism of deep learning.” In: *Neurocomputing* 452 (2021), pp. 48–62.

[83] N Romero Aquino et al. “The effect of data augmentation on the performance of convolutional neural networks.” In: *Braz. Soc. Comput. Intell* 10 (2017).

[84] Han Li. “Domain Adaptive Multi-task Learning for Complex Weather Images.” In: *Proceedings of the International Conference on Computer Vision and Deep Learning*. 2024, pp. 1–5.

[85] Alex Sherstinsky. “Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network.” In: *Physica D: Nonlinear Phenomena* 404 (2020), p. 132306.

[86] Yong Yu et al. “A review of recurrent neural networks: LSTM cells and network architectures.” In: *Neural computation* 31.7 (2019), pp. 1235–1270.

[87] Alex Graves and Alex Graves. “Long short-term memory.” In: *Supervised sequence labelling with recurrent neural networks* (2012), pp. 37–45.

[88] Gang Liu and Jiabao Guo. “Bidirectional LSTM with attention mechanism and convolutional layer for text classification.” In: *Neurocomputing* 337 (2019), pp. 325–338.

[89] Zhiyong Cui et al. “Deep bidirectional and unidirectional LSTM recurrent neural network for network-wide traffic speed prediction.” In: *arXiv preprint arXiv:1801.02143* (2018).

[90] Rusul L Abduljabbar, Hussein Dia, and Pei-Wei Tsai. “Unidirectional and bidirectional LSTM models for short-term traffic prediction.” In: *Journal of Advanced Transportation* 2021.1 (2021), p. 5589075.

[91] Zhiyong Cui et al. “Stacked bidirectional and unidirectional LSTM recurrent neural network for forecasting network-wide traffic state with missing values.” In: *Transportation Research Part C: Emerging Technologies* 118 (2020), p. 102674.

[92] Abir Rahali and Moulay A Akhloufi. “End-to-end transformer-based models in textual-based NLP.” In: *Ai* 4.1 (2023), pp. 54–110.

[93] Anthony Gillioz et al. “Overview of the Transformer-based Models for NLP Tasks.” In: *2020 15th Conference on computer science and information systems (FedCSIS)*. IEEE. 2020, pp. 179–183.

[94] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. “Ammus: A survey of transformer-based pretrained models in natural language processing.” In: *arXiv preprint arXiv:2108.05542* (2021).

[95] Xiufeng Qian et al. “Deep learning-based identification of maize leaf diseases is improved by an attention mechanism: Self-attention.” In: *Frontiers in plant science* 13 (2022), p. 864486.

[96] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. "Multi-head attention: Collaborate instead of concatenate." In: *arXiv preprint arXiv:2006.16362* (2020).

[97] Sanyuan Chen et al. "Beats: Audio pre-training with acoustic tokenizers." In: *arXiv preprint arXiv:2212.09058* (2022).

[98] Khaled Koutini et al. "Efficient training of audio transformers with patchout." In: *arXiv preprint arXiv:2110.05069* (2021).

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

[100] Heng-Jui Chang, Shu-wen Yang, and Hung-yi Lee. "Distilhubert: Speech representation learning by layer-wise distillation of hidden-unit bert." In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 7087–7091.

[101] Nan Cao et al. "Whisper: Tracing the spatiotemporal process of information diffusion in real time." In: *IEEE transactions on visualization and computer graphics* 18.12 (2012), pp. 2649–2658.