People's Democratic Republic of Algeria
Ministry of Higher Education and Scientific Research

IBN KHALDOUN UNIVERSITY OF TIARET

Dissertation

Presented to:

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE
DEPARTEMENT OF COMPUTER SCIENCE

in order to obtain the degree of :

MASTER

Specialty:
Artificial Intelligence and Digitalisation

Presented by:
Hadjar Kherfane Fatma Naila

On the theme:

Speech Recognition and Deep Learning: A Focus on Quranic Recitation

---

Defended publicly on 28 /05 /2025 in Tiaret in front the jury composed of:

Mr Bekki Khadir	MCA	Tiaret University	Chairman
Mr Hattab Nourdine	MAA	Tiaret University	Examiner
Mr Daoud Mohamed Amine	MCB	Tiaret University	Supervisor
Mr Bouguessa Abdelkader	MAA	Tiaret University	Co-Supervisor

2024-2025

Abstract

The recitation of the Quran holds immense spiritual, cultural, and educational importance within the Muslim world. Among the various modes of recitation, the Warsh style is notably prevalent in North Africa, particularly in Algeria, Morocco, and Tunisia. Accurate recitation requires mastery of the complex rules of Tajweed, which govern pronunciation, articulation, and rhythm. However, assessing the correctness of recitation remains largely dependent on human experts, posing challenges in terms of accessibility, scalability, and objectivity.

In recent years, advancements in deep learning and automatic speech recognition (ASR) have opened new possibilities for developing intelligent systems that can analyze and evaluate Quranic recitation. Despite progress in Arabic ASR, few works have addressed the specificity of Quranic recitation, particularly the Warsh style, which presents unique phonetic and prosodic characteristics.

This research aims to bridge that gap by proposing a multi-faceted system for the recognition and evaluation of Quranic recitation in the Warsh style using ensemble learning and attention-based deep neural networks. The work is structured around three major contributions:

The design of an ensemble deep learning model combining CNN, LSTM, and GRU architectures for Tajweed classification, achieving robust results in multi-class audio classification. The integration of attention mechanisms into the models, significantly improving performance by allowing the networks to focus on relevant temporal patterns in the recitation audio. The construction of a specialized and annotated dataset containing over 1,200 recitations in Warsh style, collected from Algerian participants with varying levels of expertise, and supported by a Streamlit platform to facilitate collaboration between Quranic experts and machine learning practitioners. This work not only contributes a novel dataset and high-performing models but also lays the groundwork for educational applications that can assist learners in improving their recitation. The long-term vision is to complete the dataset to include the full Quran in Warsh recitation, enabling the development of a comprehensive, real-time evaluation system capable of providing feedback and correction based on audio input, bringing Quranic learning closer to everyone, regardless of their geographic or social context.

Keywords: Quranic recitation, Warsh style, Tajweed, deep learning, speech recognition.

Résumé

La récitation du Coran revêt une importance spirituelle, culturelle et éducative immense dans le monde musulman. Parmi les différents modes de récitation, le style Warsh est particulièrement répandu en Afrique du Nord, notamment en Algérie, au Maroc et en Tunisie. Une récitation correcte exige la maîtrise des règles complexes du Tajwid, qui régissent la prononciation, l'articulation et le rythme. Toutefois, l'évaluation de la récitation reste largement dépendante d'experts humains, ce qui pose des problèmes d'accessibilité, d'évolutivité et d'objectivité.

Ces dernières années, les progrès en apprentissage profond et en reconnaissance automatique de la parole (ASR) ont ouvert de nouvelles possibilités pour le développement de systèmes intelligents capables d'analyser et d'évaluer la récitation coranique. Malgré les avancées dans l'ASR arabe, peu de travaux ont abordé les spécificités de la récitation

coranique, notamment le style Warsh, qui présente des caractéristiques phonétiques et prosodiques uniques.

Cette recherche vise à combler cette lacune en proposant un système multifacette pour la reconnaissance et l'évaluation de la récitation du Coran en style Warsh, basé sur l'apprentissage par ensemble et les réseaux neuronaux profonds avec mécanismes d'attention. Le travail s'articule autour de trois contributions majeures :

La conception d'un modèle d'apprentissage profond en ensemble combinant les architectures CNN, LSTM et GRU pour la classification du Tajwid, obtenant des résultats robustes en classification audio multiclasses.

L'intégration de mécanismes d'attention dans les modèles, améliorant significativement les performances en permettant aux réseaux de se concentrer sur les motifs temporels pertinents dans l'audio de récitation.

La constitution d'un ensemble de données spécialisé et annoté contenant plus de 1 200 récitations en style Warsh, recueillies auprès de participants algériens de différents niveaux, et supporté par une plateforme Streamlit facilitant la collaboration entre experts du Coran et praticiens de l'apprentissage automatique.

Ce travail apporte non seulement un nouvel ensemble de données et des modèles performants, mais il pose également les bases d'applications éducatives pouvant aider les apprenants à améliorer leur récitation. L'objectif à long terme est de compléter le corpus pour couvrir l'ensemble du Coran en style Warsh, permettant le développement d'un système d'évaluation complet et en temps réel fournissant des retours et des corrections à partir de l'audio, rapprochant ainsi l'apprentissage du Coran de chacun, indépendamment du contexte géographique ou social.

Mots-clés : Récitation coranique, style Warsh, Tajwid, apprentissage profond, reconnaissance vocale.

Acknowledgements

All praise is due to Allah, Lord of the Worlds, the Most Gracious, the Most Merciful. We glorify Him, seek His help, and ask for His forgiveness. He alone grants success, and without His guidance and mercy, nothing is possible. Whoever Allah guides, none can mislead, and whoever is left astray, none can guide.

With deep humility and immense gratitude, I begin by thanking **Allah**, who granted me strength, patience, and clarity throughout this journey. May this work be accepted as a small offering in His cause and a step toward continuous betterment.

I wish to extend my profound thanks to my supervisor, **Mr. Daoud Mohamed Amine**, *Maître de Conférences*, for his wisdom, kindness, and continuous support. His thoughtful guidance and sincere encouragement have been a source of strength and inspiration at every stage of this thesis. May Allah reward him with endless goodness.

To my co-supervisor, **Mr. Bougassa Abdelkader**, I am deeply grateful for his generous guidance, his availability, and his clarity in direction. His support has been a vital part of my progress, and I sincerely thank him for his trust and care.

I would also like to express my heartfelt appreciation to **Mr. Chikhaoui Ahmad**, a remarkable teacher who stood by me with patience, understanding, and unwavering encouragement. His support meant more than words can express, and I am sincerely thankful.

My gratitude also goes to **Mr. Laid Lahcen**, whose helpful insights and kind guidance added great value to my learning experience.

I am honored and thankful to **Mr. Bekki Khadir**, *Maître de Conférences*, and **Mr. Hattab Noureddine** for graciously agreeing to evaluate this work. Their time, attention, and expertise are deeply appreciated.

I extend my thanks to all the professors at the Faculty of Mathematics and Computer Science — especially within the Department of Computer Science — for the knowledge they imparted and the dedication with which they supported their students.

To my entire family, who believed in me even when I doubted myself — thank you for your constant love, encouragement, and presence throughout this journey.

Finally, to every person who contributed in any way to the completion of this thesis — be it through support, advice, prayer, or simply a kind word — I offer my heartfelt thanks. May Allah accept this work as a **ṣadaqa jāriya (continuous charity)** for all who helped, for myself, and for my family. May it be a source of light in this life and the next.

Contents

1 State of the Art **11**

1.1 Introduction . . . 11

1.2 Speech Recognition . . . 11

1.2.1 Definition of Speech Recognition . . . 11

1.2.2 Process Steps . . . 12

1. Feature Extraction . . . 12

2. Acoustic Modeling . . . 12

3. Decoding . . . 12

4. Text Interpretation . . . 12

1.2.3 Representation of the Vocal Signal . . . 13

1. Spectrograms . . . 13

Definition: . . . 13

2. Cepstral Coefficients . . . 14

Definition: . . . 14

1.2.4 Evolution of Speech Recognition . . . 15

1.2.4.1 Early Developments (1950s-1970s) . . . 15

1.2.4.2 Advancements in the 1980s and 1990s . . . 16

1.2.4.3 The 2000s: The Rise of Machine Learning . . . 16

1.2.4.4 Recent Advances and Current State (2010s-Present) . . . 16

1.2.5 ASR Applications . . . 17

1.2.5.1 Speech Recognition in Virtual Assistants . . . 17

1. Siri (Apple) . . . 17

2. Alexa (Amazon) . . . 17

3. Google Assistant . . . 18

1.2.5.2 Use in Automatic Transcription . . . 18

Applications in Journalism . . . 18

Importance in Medical Services . . . 18

Broader Industry Impact . . . 18

1.2.5.3 ASR Applications in Quranic Recitation . . . 19

Learn Quran Tajwid . . . 19

1.2.6 Speech Recognition in the Context of Quranic Recitation . . . 19

1.2.6.1 Definition of the Quran . . . 19

1.2.6.2 Definition of Tajweed . . . 20

1.2.6.3 Cultural and Religious Importance of Reciting the Quran . . . 20

Spiritual Significance . . . 20

Educational Importance . . . 21

Social and Cultural Importance . . . 21

1.2.6.4 Unique Intonations and Sounds in Quranic Recitation . . . 21

The Role of Tajweed in Quranic Recitation 22
Variations in Recitation Styles 22
1.2.6.5 Components of Tajweed Rules 22
1.2.6.6 The Ten Mutawatir Qira'at (Readings) 23
Definition of Quranic Readings and Their Sources 23
1.2.6.7 The Seven Well-Known Qira'at 24
Qira'ah of Nāfiᶜ (Narrators: Warsh and Qālūn) 24
Qira'ah of Ibn Kathīr (Narrators: al-Bazzī and Qunbul) . 24
Qira'ah of Abū ᶜAmr (Narrators: al-Dūrī and al-Sūsī) . . . 24
Qira'ah of Ibn ᶜĀmir (Narrators: Hishām and Ibn Dhakwān) 24
Qira'ah of ᶜĀsim (Narrators: Shuᶜbah and Hafs) 24
Qira'ah of Ḥamzah (Narrators: Khalaf and Khallād) . . . 25
Qira'ah of al-Kisā'ī (Narrators: al-Layth and Ḥafṣ al-Dūrī) 25
1.2.6.8 The Three Additional Qira'at Completing the Ten . . . 25
Qira'ah of Abū Jaᶜfar (Narrators: Ibn Wardān and Ibn Jammaz) 25
Qira'ah of Yaᶜqūb (Narrators: Ruways and Rawḥ) 25
Qira'ah of Khalaf al-ᶜĀshir (Narrators: Isḥāq al-Marwazī and Idrīs al-Ḥaddād) 25
1.2.6.9 The Significance and Wisdom Behind the Diversity of Qira'at 26
1.2.6.10 Distinctive Features of Warsh's Recitation 26
1.2.6.11 Key Differences Between Warsh and Hafs Readings . . . 26
1.2.7 Complexity of Tajweed Rules and Their Impact on Speech Recognition 27
1.2.7.1 Understanding Tajweed Rules 28
1.2.7.2 Impact on Speech Recognition Technology 28
1.2.7.3 Pronunciation Variability Among Reciters 29
Influence of Regional Dialects 29
Personal Interpretation and Style 29
Variability in Tajweed Application 29
Impact on ASR System Performance 30
1.2.7.4 Phonetic Particularities of the Quran Compared to Standard Arabic 30
Articulation and Pronunciation Variability 30
Contextual Phonetic Variability 30
Impact on Automatic Speech Recognition 30
1.2.7.5 Lack of Resources and Annotated Databases for Model Training 31
The Importance of Annotated Datasets 31
Challenges in Data Collection 31
Consequences of Limited Data 31
Sensitivity to Tone and Modulations Specific to Each Recitation Style 32
The Importance of Tone and Modulation in Quranic Recitation 32
1.2.7.6 Challenges for ASR Systems 32
1.3 Conclusion 33

2 Speech Recognition Techniques **35**
2.1 Introduction . . . 35
2.2 Traditional Speech Recognition Techniques . . . 35
2.2.1 Gaussian Mixture Models (GMM): Probabilistic Techniques for Modeling Vocal Features . . . 35
Mathematical Foundation . . . 35
Feature Extraction . . . 36
Advantages and Limitations . . . 36
2.2.2 Hidden Markov Models (HMM): Sequential Probability-Based Approaches for Modeling Phoneme Sequences . . . 36
Mathematical Framework . . . 36
Training HMMs . . . 37
Decoding with HMMs . . . 37
Advantages and Limitations . . . 37
2.2.3 Rule-Based Methods: Use of Linguistic Models Based on Expert Knowledge . . . 37
Fundamentals of Rule-Based Methods . . . 37
Advantages of Rule-Based Methods . . . 38
Limitations of Rule-Based Methods . . . 38
2.3 Modern Speech Recognition Techniques . . . 38
2.3.1 Deep Neural Networks (DNNs) . . . 38
2.3.2 Convolutional Neural Networks (CNNs) . . . 39
2.3.3 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) . . . 40
2.3.4 Gated Recurrent Unit (GRU) . . . 40
2.3.5 Transformers . . . 40
2.3.6 Attention Mechanisms . . . 41
2.3.7 Ensemble Learning Methods . . . 41
2.3.7.1 Types of Ensemble Methods . . . 41
Bagging (Bootstrap Aggregating) . . . 41
Boosting . . . 42
Stacking (Stacked Generalization) . . . 43
Voting and Averaging . . . 43
2.4 Existing Research in Quranic Recitation Recognition . . . 44
2.4.1 Early Works and Foundational Studies . . . 44
2.4.2 Arabic Speech and Quranic Recitation Datasets . . . 44
2.4.3 Deep Learning Approaches for Quranic Recitation Recognition . . . 45
2.4.4 Qira'at and Recitation Style Classification . . . 46
2.4.5 Error Detection and Correction in Recitation . . . 46
2.4.6 Multi-modal Systems and Innovation in Quranic Recitation Recognition . . . 47
2.4.7 Recent Advances in Automatic Speech Recognition (ASR) for Quranic Recitation . . . 47
2.5 Conclusion . . . 48

3 proposed approaches **50**
3.1 Introduction . . . 50
3.2 overview . . . 50

3.3 Proposed Approachs : . . . 50
3.4 The first approach . . . 50
3.4.1 Dataset . . . 51
3.4.2 Preprocessing Pipeline . . . 52
3.4.3 Model Architectures . . . 52
3.4.3.1 CNN Architecture . . . 52
3.4.3.2 LSTM Architecture . . . 54
3.4.3.3 GRU Architecture . . . 55
3.4.4 Ensemble Learning . . . 56
3.4.5 Performance Analysis . . . 57
3.4.5.1 Individual Model Performance . . . 57
3.4.5.2 Technical Implementation Details . . . 57
3.4.5.3 Confusion Matrix Analysis . . . 58
3.4.6 Conclusion . . . 58
3.5 The second approach . . . 59
3.5.1 Attention Mechanism Implementation . . . 60
3.5.2 Refined Model Architectures . . . 60
3.5.2.1 Optimized CNN Architecture . . . 60
3.5.2.2 Optimized LSTM Architecture . . . 62
3.5.2.3 Optimized GRU Architecture . . . 63
3.5.3 Advanced Stacking Ensemble Implementation . . . 63
3.5.3.1 Ensemble Performance Analysis . . . 64
3.5.4 Comparative Analysis with First Approach . . . 65
3.5.5 Conclusion . . . 65
3.6 The Third Approach . . . 66
3.6.1 Data Collection and Warsh Recitation Identification . . . 66
3.6.2 Dataset Architecture . . . 66
3.6.2.1 Core Structure . . . 66
Noureddine Moulay : . . . 66
Dr. Ben Halima Othman : . . . 66
Identification fields: . . . 66
Audio field: . . . 66
Tajweed rule fields: . . . 67
Evaluation metrics: . . . 67
3.6.2.2 Tajweed Rule Categories . . . 67
3.6.3 Dataset Management Tool . . . 68
3.6.4 Dataset Composition . . . 70
3.6.5 Applications and Future Work . . . 70
3.6.6 Conclusion . . . 71
3.7 Conclusion . . . 71

List of Figures

1.1 The four key steps in the speech recognition process . . . 13
1.2 The evolution of speech recognition technology . . . 15

2.1 Architecture of a Deep Neural Network (DNN) used for Speech Recognition [1]. . . . 39
2.2 A simple architecture of a Convolutional Neural Network (CNN) [2]. . . . 39
2.3 Architecture of a Long Short-Term Memory (LSTM) [3] . . . 40
2.4 Architecture of a Transformer [4] . . . 41

3.1 Process of the Ensemble Learning Model for Quranic Recitation Recognition 51
3.2 The CNN Models accuracy . . . 53
3.3 The LSTM Models accuracy . . . 55
3.4 The GRU Models accuracy . . . 56
3.5 Confusion Matrix for Meta-Model . . . 57
3.6 Process of the Ensemble Learning Model with attention mechanism for Quranic Recitation Recognition . . . 59
3.7 The CNN Models accuracy . . . 61
3.8 The LSTM Models accuracy . . . 62
3.9 The GRU Models accuracy . . . 64
3.10 Confusion Matrix for Meta-Model . . . 64
3.11 Tajweed Rules Evaluation interface showing all available rules that can be annotated for each verse. . . . 68
3.12 Rule evaluation results showing presence and status of different Tajweed rules in a specific verse. . . . 69
3.13 Verse editing interface allowing for Surah name, verse number, and text input with Tajweed rule annotation. . . . 69
3.14 Dataset overview showing Quranic verses with their associated audio files and Tajweed rule annotations. . . . 70

Introduction

The recitation of the Quran holds immense spiritual, cultural, and educational importance within the Muslim world. Among the various modes of recitation, the Warsh style is notably prevalent in North Africa, particularly in Algeria, Morocco, and Tunisia. Accurate recitation requires mastery of the complex rules of *Tajweed*, which govern pronunciation, articulation, and rhythm. However, assessing the correctness of recitation remains largely dependent on human experts, posing challenges in terms of accessibility, scalability, and objectivity.

In recent years, advancements in deep learning and automatic speech recognition (ASR) have opened new possibilities for developing intelligent systems that can analyze and evaluate Quranic recitation. Despite progress in Arabic ASR, few works have addressed the specificity of Quranic recitation—particularly the Warsh style—which presents unique phonetic and prosodic characteristics.

This research aims to bridge that gap by proposing a multi-faceted system for the recognition and evaluation of Quranic recitation in the Warsh style using ensemble learning and attention-based deep neural networks. Our work is structured around three major contributions:

- The design of an ensemble deep learning model combining CNN, LSTM, and GRU architectures for Tajweed classification, achieving robust results in multi-class audio classification.
- The integration of attention mechanisms into the models, significantly improving performance by allowing the networks to focus on relevant temporal patterns in the recitation audio.
- The construction of a specialized and annotated dataset containing over 1,200 recitations in Warsh style, collected from Algerian participants with varying levels of expertise, and supported by a Streamlit platform to facilitate collaboration between Quranic experts and machine learning practitioners.

This work not only contributes a novel dataset and high-performing models, but also lays the groundwork for educational applications that can assist learners in improving their recitation. Our long-term vision is to complete the dataset to include the full Quran in Warsh recitation, enabling the development of a comprehensive, real-time evaluation system capable of providing feedback and correction based on audio input—bringing Quranic learning closer to everyone, regardless of their geographic or social context.

Chapter 01
State of the Art

Chapter 1

State of the Art

1.1 Introduction

The intersection of ancient Quranic recitation practices and modern speech recognition technology presents a fascinating yet challenging domain of study. This chapter examines the rich tradition of Quranic recitation with its complex Tajweed rules and diverse canonical styles (Qira'at), each with unique phonetic features passed down through specific narrators. As digital technology advances, there's growing interest in developing automatic speech recognition (ASR) systems capable of accurately processing these recitations, though the distinctive pronunciation patterns, regional variations, and melodic characteristics create significant technical hurdles that standard ASR approaches struggle to overcome.

1.2 Speech Recognition

1.2.1 Definition of Speech Recognition

Speech recognition, commonly referred to as Automatic Speech Recognition (ASR), is a technology that enables computers to interpret and transcribe human speech into text [5]. At its core, speech recognition involves capturing spoken language through a microphone or other audio input, analyzing the sound waves, and converting them into a digital format that machines can process. This conversion allows ASR systems to recognize individual words, and sentences, making it possible for computers to understand and respond to human language [6].

ASR technology works by breaking down audio signals into smaller parts, called acoustic features, which capture the unique characteristics of human speech. These features are then compared to patterns in a database to identify the corresponding words. ASR systems use models and algorithms trained on vast amounts of spoken data, allowing them to recognize a wide range of speech characteristics, from different accents and dialects to various speaking speeds and intonations. [7].

The applications of ASR are broad and impactful. Voice-activated digital assistants, such as Apple's Siri, Amazon's Alexa, and Google Assistant, rely on ASR to enable hands-free operation, letting users perform tasks through spoken commands. Speech recognition is also widely used in transcription services, allowing spoken content like meetings, lectures, and interviews to be automatically transcribed into text. Additionally, ASR plays a key role in accessibility, providing people with disabilities a way to

interact with technology more naturally [8]. This technology has become essential for many modern devices and services, fostering a seamless connection between humans and computers through natural language interaction.

1.2.2 Process Steps

The speech recognition process consists of several critical steps, each contributing to the overall effectiveness of the system. These steps are detailed below:

1. Feature Extraction Feature extraction is the first critical step in speech recognition, where raw audio signals are transformed into a set of features that can be effectively used for recognition. This transformation typically includes preprocessing the audio to remove noise and enhance signal quality, followed by applying techniques such as Short-Time Fourier Transform (STFT) to analyze the frequency content over time [9].

The most common features used in speech recognition are Mel-Frequency Cepstral Coefficients (MFCCs), which represent the short-term power spectrum of sound. MFCCs are designed to capture the characteristics of human speech perception, focusing on frequencies that are more relevant to human hearing. The extraction process usually yields a feature vector for each frame of audio, summarizing the key information needed for further analysis [10].

2. Acoustic Modeling Once features are extracted, the next step is acoustic modeling, which involves creating statistical models to represent the relationship between the extracted features and the corresponding phonetic units in speech (e.g., phonemes or sub-phonemes). Acoustic models typically utilize techniques like Hidden Markov Models (HMMs) or neural networks to capture the temporal dynamics of speech.

Acoustic models are trained using large datasets of audio recordings paired with their corresponding transcriptions. The quality and quantity of training data significantly impact the model's ability to generalize to unseen speech, making it a crucial consideration in the development of effective speech recognition systems [11].

3. Decoding The decoding step is where the recognition happens. In this phase, the acoustic model is applied to the features extracted from the input audio signal to identify the most likely sequence of phonetic units. This process involves searching through the vast space of possible word sequences and their corresponding acoustic representations.

Decoding typically employs algorithms like the Viterbi algorithm, which finds the most probable path through the state space defined by the HMMs, or other sophisticated beam search algorithms that optimize computational efficiency while exploring candidate sequences. The output of the decoding process is a sequence of recognized phonetic units, which may need to be further processed into actual words [12].

4. Text Interpretation The final step in the speech recognition process is text interpretation, where the phonetic output is converted into readable text. This step often involves language modeling, which helps improve the accuracy of word recognition by considering the context in which words are used.

Language models use statistical or neural approaches to predict the likelihood of sequences of words, thereby assisting in resolving ambiguities that arise from homophones (words that sound the same but have different meanings). The most common language

models include n-gram models, which predict the probability of a word based on the previous $n-1$ words, and more sophisticated models based on neural networks, such as Long Short-Term Memory (LSTM) networks or Transformer-based architectures [13].

In this final phase, the recognized words are assembled into sentences, and additional processing may be applied for tasks such as punctuation insertion and context-based adjustments. This step ensures that the output text is coherent and semantically meaningful [14].

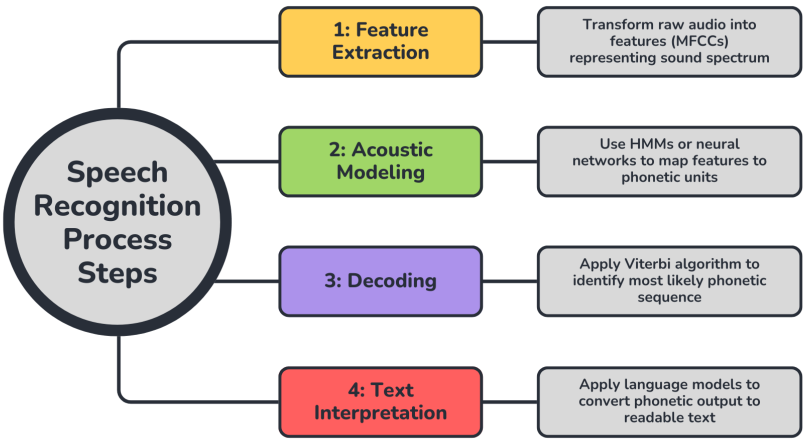

Figure 1.1: The four key steps in the speech recognition process

1.2.3 Representation of the Vocal Signal

In speech recognition, the representation of vocal signals is fundamental for transforming raw audio signals into a format suitable for processing by algorithms. Two key methods of representing vocal signals are **spectrograms** and **cepstral coefficients**, particularly the **Mel-Frequency Cepstral Coefficients (MFCCs)**.

1. Spectrograms

Definition: **A** spectrogram is a visual representation of the frequency spectrum of a signal as it changes over time. It illustrates how the power of various frequency components of the audio signal is distributed throughout its duration.

Mathematical Foundation: To create a spectrogram, the audio signal is divided into overlapping segments (frames), and the Short-Time Fourier Transform (STFT) is applied to each frame. Mathematically, the STFT is defined as:

$$STFT\{x(t)\} = X(m,\omega) = \int_{-\infty}^{\infty} x(t)w(t-m)e^{-j\omega t}dt \tag{1.1}$$

Where:

- $x(t)$ is the input signal,
- $w(t)$ is a window function,

- m represents the time index (frame),
- ω is the angular frequency.

After computing the STFT, the magnitude spectrum is taken to form the spectrogram, which is then displayed in a 2D plot, showing time on the x-axis, frequency on the y-axis, and the magnitude (or power) of frequencies represented by color intensity.

Characteristics:

- **Time Resolution vs. Frequency Resolution:** The choice of frame size and overlap directly affects the time and frequency resolution. Short frames provide better time resolution but poorer frequency content, while long frames improve frequency resolution but blur temporal details.
- **Windowing Functions:** Commonly used window functions include Hanning and Hamming windows, which smooth the edges of each frame to minimize discontinuities.

Applications:

- **Phonetics:** Spectrograms are extensively used in phonetics to analyze speech sounds, allowing researchers to observe phonetic characteristics such as formants (resonant frequencies in the human vocal tract), pitch variations, and speech patterns [14].
- **Feature Extraction:** They serve as inputs for machine learning models in speech recognition, encapsulating detailed temporal and spectral information about the speech signal [10].

2. Cepstral Coefficients

Definition: **C**epstral coefficients are derived from the power spectrum of a signal and are particularly effective in representing the characteristics of human speech. The most common type, the Mel-Frequency Cepstral Coefficients (MFCCs), emphasizes perceptually relevant features.

Mathematical Foundation: The computation of MFCCs involves several steps:

1. **Pre-emphasis:** A high-pass filter is applied to the audio signal to amplify high frequencies:
$$y(t) = x(t) - \alpha x(t-1) \tag{1.2}$$
Where α is typically set to 0.95.
2. **Frame Blocking and Windowing:** The signal is divided into frames and a window function is applied.
3. **Fourier Transform:** The Fast Fourier Transform (FFT) is computed for each frame to obtain the magnitude spectrum.
4. **Mel Filter Bank:** The magnitude spectrum is filtered through a set of triangular filters that mimic human auditory perception. The Mel scale is a perceptual scale of pitches that approximates human hearing.

5. **Logarithm:** The logarithm of the Mel spectrum is taken.

6. **DCT:** Finally, the Discrete Cosine Transform (DCT) is applied to decorrelate the coefficients:

$$c_k = \sum_{n=0}^{N-1} \log(mel_n) \cos\left[\frac{\pi k}{N}\left(n + \frac{1}{2}\right)\right] \tag{1.3}$$

Where:

- mel_n is the n-th Mel-filtered coefficient,
- c_k are the resulting cepstral coefficients.

1.2.4 Evolution of Speech Recognition

The history and evolution of speech recognition technologies span several decades, reflecting advancements in linguistics, computer science, and artificial intelligence. Below, we outline key milestones in the development of speech recognition systems, highlighting significant breakthroughs and contributions to the field.

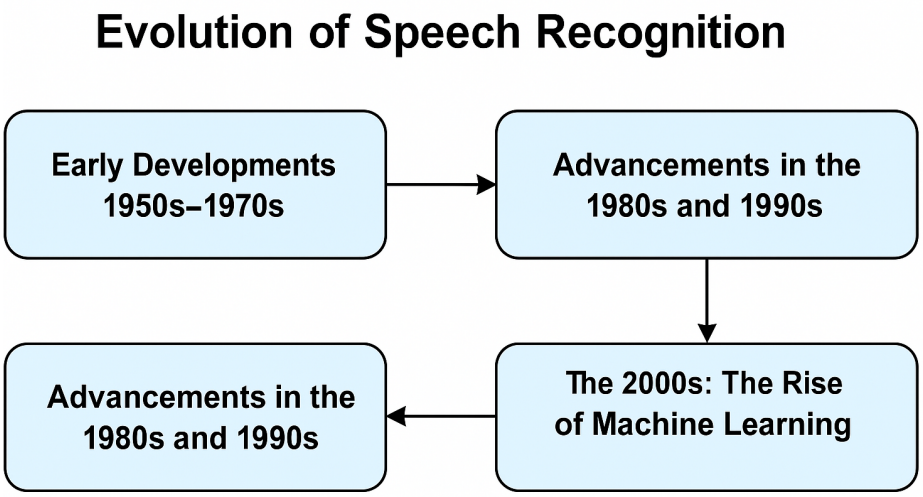

Figure 1.2: The evolution of speech recognition technology

1.2.4.1 Early Developments (1950s-1970s)

- **Initial Concepts**: The journey of speech recognition began in the 1950s with pioneering work by researchers at Bell Labs, where early systems could recognize a limited vocabulary of spoken words. One notable example was the "Audrey" system, which could recognize digits spoken by a single voice, developed in 1952 by D. O. Hennessey and R. C. Wilcox. [5]

- **Pattern Recognition**: By the late 1950s and early 1960s, speech recognition shifted towards pattern recognition approaches. Researchers began using techniques such as Dynamic Time Warping (DTW) to match spoken input to predefined templates, allowing for some flexibility in recognizing variations in speech.
- **Statistical Models**: In the 1970s, the introduction of statistical models, particularly Hidden Markov Models (HMMs), marked a significant advancement. HMMs provided a probabilistic framework for modeling the sequential nature of speech. This approach was further developed by L. R. Rabiner, whose work laid the foundation for modern speech recognition systems.

1.2.4.2 Advancements in the 1980s and 1990s

- **Linguistic Knowledge**: The 1980s saw the integration of linguistic knowledge into speech recognition systems. Techniques such as n-gram language models were employed to improve recognition accuracy by taking into account the context of words in a sentence.
- **Commercial Systems**: The first commercial speech recognition systems emerged during this period. Dragon Dictate, released in 1990, was one of the first widely used dictation software packages that allowed users to control their computers using voice commands.
- **Large Vocabulary Continuous Speech Recognition**: Advances in computational power enabled the development of large vocabulary continuous speech recognition (LVCSR) systems. Researchers began to design systems capable of recognizing continuous speech without the need for pauses between words. Notable projects included IBM's ViaVoice and Nuance's Dragon NaturallySpeaking.

1.2.4.3 The 2000s: The Rise of Machine Learning

- **Deep Learning**: The emergence of deep learning in the 2010s revolutionized speech recognition technologies. Researchers began applying neural networks, particularly Deep Neural Networks (DNNs), to model the complex relationships between audio features and phonetic units. Hinton et al. demonstrated the effectiveness of DNNs for phoneme recognition in 2012 [7].
- **End-to-End Systems**: The development of end-to-end systems further simplified the architecture of speech recognition models. Techniques like Connectionist Temporal Classification (CTC) allowed for direct mapping from audio input to text output, reducing the need for intermediate representations [15].
- **Attention Mechanisms**: The introduction of attention mechanisms, particularly with sequence-to-sequence models using Recurrent Neural Networks (RNNs), improved the ability of systems to handle variable-length input sequences, leading to better performance in recognizing longer phrases and sentences.

1.2.4.4 Recent Advances and Current State (2010s-Present)

- **Transformer Models**: The advent of the Transformer architecture has further enhanced speech recognition capabilities. Models such as Google's BERT and Ope-

nAI's GPT have demonstrated state-of-the-art performance in various natural language processing tasks, including speech recognition [13].

- **Multilingual and Cross-Language Models**: Recent research has focused on developing multilingual models capable of recognizing speech in multiple languages without needing separate models for each language. This development is crucial for making speech recognition technologies accessible globally.
- **Integration into Everyday Technology**: Speech recognition has become ubiquitous in consumer technology, integrated into virtual assistants like Siri, Alexa, and Google Assistant. These systems utilize advanced deep learning techniques and large datasets to improve accuracy and responsiveness.
- **Continued Research**: Ongoing research continues to address challenges such as speaker variability, background noise, and the need for privacy in voice data processing. On-device processing, as seen in recent iterations of virtual assistants, is becoming more prevalent to enhance privacy and responsiveness.

1.2.5 ASR Applications

1.2.5.1 Speech Recognition in Virtual Assistants

Speech recognition in virtual assistants like **Siri**, **Alexa**, and **Google Assistant** has revolutionized human-device interaction by allowing systems to interpret and respond to spoken language commands. These assistants operate using a combination of *Automatic Speech Recognition (ASR)*, *Natural Language Processing (NLP)*, and *Machine Learning (ML)* to interpret user speech and carry out tasks such as controlling smart devices or retrieving information [16].

1. Siri (Apple) **Siri**, introduced by Apple in 2011, was one of the first voice-activated personal assistants widely available to the public. Siri's ASR system is driven by *Deep Neural Networks (DNNs)* that convert spoken words into text. It then uses *Natural Language Understanding (NLU)* to interpret the query and provide a suitable response or perform an action, such as sending a message or opening an app [17]. Over time, Siri has improved with the integration of on-device processing, which enhances both speed and privacy. With Apple's *Neural Engine*, the voice recognition and processing occur directly on the device for faster response times and increased privacy by limiting the transmission of voice data to the cloud [18]. Siri is also context-aware, meaning it can use data such as location and recent actions to provide more relevant results [19].

2. Alexa (Amazon) **Alexa**, launched by Amazon in 2014 through the Amazon Echo smart speaker, uses *far-field voice recognition* technology, allowing it to pick up commands from across the room. Alexa's core functionalities rely on *cloud-based ASR* to convert spoken language into text, after which NLU models process the text and extract user intent [20]. Alexa's wake-word detection continuously listens for specific activation phrases such as "Alexa" to initiate command processing, which reduces the need for constant interaction. This system is powered by large-scale machine learning models trained on vast amounts of voice data, allowing Alexa to perform a wide range of tasks, from controlling smart home devices to answering general questions [21]. Alexa also supports

a developer ecosystem called *Alexa Skills*, which enables third-party developers to create voice applications that extend its capabilities [22].

3. Google Assistant **Google Assistant**, integrated into a range of devices such as Android smartphones and Google Nest, uses advanced *Recurrent Neural Networks (RNNs)* and *Long Short-Term Memory (LSTM)* networks for speech recognition and understanding. These models help process more complex commands with high accuracy [23]. Google's *Knowledge Graph* plays a pivotal role in allowing Google Assistant to retrieve information quickly and offer detailed answers to user queries. Furthermore, *multilingual support* is a standout feature of Google Assistant, allowing users to issue commands in different languages interchangeably, thanks to *Transformer-based models* that support this functionality [24]. Google Assistant can also access personal data like calendar events or emails to provide personalized responses based on user preferences and activity [25].

1.2.5.2 Use in Automatic Transcription

Automatic transcription technologies have profoundly impacted various industries, notably journalism and medical services, by automating the process of converting spoken language into written text. This advancement not only enhances efficiency but also improves accuracy in documentation, making these technologies essential tools in contemporary workflows.

Applications in Journalism In journalism, automatic transcription systems enable reporters to quickly convert interviews, press conferences, and speeches into text. This rapid transcription capability allows journalists to publish news stories more promptly, which is critical in a fast-paced media environment. By utilizing automatic transcription, journalists can devote more time to analyzing content rather than manually transcribing audio recordings [26–28].

Importance in Medical Services In the medical field, automatic transcription technologies are pivotal for improving documentation practices. Healthcare professionals frequently rely on dictation systems to record patient encounters, clinical notes, and other essential communications directly into electronic health records (EHRs). This practice not only saves time but also minimizes the administrative burden on physicians, allowing them to focus more on patient care [29].

The use of voice recognition systems ensures that medical records are accurately transcribed, thereby enhancing patient safety and reducing the risks associated with miscommunication [30]. For example, a study indicated that automatic transcription reduced the time spent on documentation tasks by approximately 50

Broader Industry Impact Beyond journalism and healthcare, automatic transcription technologies are finding applications across various other sectors, including legal and educational fields. In the legal domain, law firms utilize transcription software to create records of court proceedings and depositions. This practice ensures that accurate records are available for reference, which is vital for legal proceedings [31].

In education, automatic transcription is being used to enhance learning experiences for students, especially those with disabilities. By transcribing lectures and discussions,

educational institutions can provide students with accessible resources, improving their understanding of course materials [32].

1.2.5.3 ASR Applications in Quranic Recitation

Automatic Speech Recognition (ASR) technology has been effectively applied in Quranic recitation, providing tools that assist learners in improving their pronunciation and adherence to Tajweed rules. These applications offer real-time feedback, making Quranic education more accessible and interactive.

Learn Quran Tajwid Learn Quran Tajwid is a comprehensive mobile application that offers lessons on Quranic recitation and Tajweed. While it includes features like voice narration, recording, and transliteration to aid learning, it does not currently incorporate real-time ASR feedback. Users can record their recitation and compare it with provided examples, facilitating self-assessment and improvement [33].

1.2.6 Speech Recognition in the Context of Quranic Recitation

1.2.6.1 Definition of the Quran

The Quran is "the Speech of Allah, revealed to His Prophet Muhammad (peace be upon him), miraculous in its words and meanings, to be recited in worship, transmitted to us through an uninterrupted chain of narration, and written in the Mushafs, from the beginning of Surah Al-Fatiha to the end of Surah An-Nas," which is the preferred definition [34]. The Quran is the primary source of spiritual guidance, religious law, and moral principles for Muslims worldwide. Unlike modern standards, it preserves its original form in both its words and meanings [34].

The Quran emphasizes its own significance and transformative power in several verses. Allah the Almighty says:

> *"And this is a blessed Book which We have revealed, so follow it and fear Allah that you may receive mercy."*
> **[Al-An'ām: 155]**

> *"This (Quran) is a message for humanity, so that they may be warned by it, and that they may know that He is only One God, and that those of understanding may take heed."*
> **[Ibrāhīm: 52]**

> The Prophet Muhammad (peace be upon him) said:
> *"Would any of you like to go to the valley of Buthan or Al-'Aqiq and return with two large she-camels without committing any sin or severing family ties?"*
> They replied, "Yes." He said:
> *"Then going to the mosque and learning or reciting two verses from the Book of Allah is better for him than two she-camels, and three verses are better than three, and four verses are better than four, and so on."*
> **[Sahih Muslim]**

1.2.6.2 Definition of Tajweed

Tajweed is the science of Quranic pronunciation. It encompasses a comprehensive set of phonetic and articulation rules designed to preserve the accuracy and beauty of Quranic recitation. The purpose of Tajweed is to ensure that each letter and sound is recited from its correct articulation point, with proper characteristics and intonations, as practiced by the Prophet Muhammad (peace be upon him). Tajweed is more than a linguistic discipline; it is a spiritual practice. It maintains the integrity of the Quranic text and prevents errors that could alter meanings [35].

> *"Indeed, those who recite the Book of Allah, establish prayer, and spend from what We have provided them, secretly and publicly, can expect a profit that will never perish."*
> **[Fāṭir: 29]**

> Narrated Aisha (may Allah be pleased with her):
> The Messenger of Allah (peace be upon him) said:
> *"The one who recites the Quran proficiently will be with the noble, righteous scribes; and the one who reads the Quran, struggling with it and stumbling through its verses, will have a double reward."*
> **[Agreed upon – Bukhari and Muslim]**

1.2.6.3 Cultural and Religious Importance of Reciting the Quran

The recitation of the Quran is of immense cultural and religious importance in Islam. It is not merely a routine practice but a sacred act deeply intertwined with the spiritual, educational, and social fabric of Muslim societies. The Quran, being the holy book of Islam, is regarded as the literal word of God, revealed to the Prophet Muhammad over 1,400 years ago. Its recitation, known as *tilawah*, is considered a form of worship and a way to connect with the divine, carrying both spiritual and cultural significance for Muslims worldwide [36].

Spiritual Significance In Islam, reciting the Quran is an essential spiritual practice that goes beyond reading for information. It is seen as a direct communication with God. The Quran itself emphasizes the importance of recitation, as mentioned in Surah Al-Muzzammil (73:4), "Recite the Quran in slow, measured rhythmic tones." This verse highlights not only the act of reciting but also the manner in which it should be performed. Muslims believe that reciting the Quran brings spiritual rewards, known as *hasanat*, and serves as a means of spiritual purification. Regular recitation is also thought to bring comfort, alleviate stress, and provide guidance in a believer's daily life [37].

The spiritual significance of recitation is further emphasized during the month of Ramadan, the holiest period in the Islamic calendar. Muslims strive to complete the entire recitation of the Quran during this month, especially through nightly prayers known as *Taraweeh*. This practice, known as *Khatm al-Quran*, is a highly rewarding spiritual endeavor and a tradition that brings the community together in worship. The act of recitation during Ramadan is believed to have special merit, as the Quran was first revealed during this month [38].

Educational Importance Learning to recite the Quran correctly is a foundational element of Islamic education. This education starts early in life, with young Muslims being taught to read the Quran in Arabic, even if it is not their native language. The process of learning Quranic recitation involves memorization (*Hifz*) and mastering the rules of pronunciation and intonation, known as *Tajweed*. Quranic schools, or *madrasas*, across the world place significant emphasis on these skills, which are seen as essential for preserving the Quran in its original form and for maintaining the proper pronunciation of the Arabic language [39].

The educational aspect also extends to those who achieve the memorization of the entire Quran, known as *Hafiz*. This accomplishment is highly respected in the Muslim community and is often associated with spiritual leadership. Being a *Hafiz* carries not only religious merit but also social prestige, as those who have memorized the Quran are considered guardians of its message. Furthermore, memorization is believed to facilitate a deeper understanding of the text and its meanings, contributing to the educational and spiritual growth of the individual [40].

Social and Cultural Importance The Quran plays a central role in various social and cultural practices within the Muslim community. It is recited during significant life events, including weddings, funerals, and the birth of a child, where specific verses are chosen to seek blessings, offer comfort, or commemorate the occasion. This tradition reflects the deep integration of the Quran in the cultural life of Muslims, where it serves as a constant spiritual guide and a source of solace during times of joy and sorrow [41].

In addition, Quranic recitation competitions are popular in many Muslim-majority countries and communities, with participants demonstrating their mastery of the text and the rules of Tajweed. These competitions, often held at local, national, and international levels, celebrate the beauty of the Quran and encourage young Muslims to engage with the holy text. The social impact of such events extends beyond the individual participants, as they inspire others to appreciate and pursue the study and recitation of the Quran [42].

1.2.6.4 Unique Intonations and Sounds in Quranic Recitation

Quranic recitation involves intonations and sounds that are not commonly found in conversational Arabic. These unique features are governed by the rules of *Tajweed* and include:

- **Melodic Intonations**: The melodic nature of Quranic recitation is a deliberate feature meant to enhance the listening experience. The reciter's voice modulation follows a rhythmic pattern that emphasizes certain words or phrases, giving the recitation a musical quality that is not present in spoken Arabic [43].
- **Use of Emphatic Consonants**: Some consonants in Arabic are articulated with greater emphasis (e.g., the letters *Ṣād*, *Ḍād*, *Ṭā*, and *Ẓā*). In the context of Quranic recitation, the degree of emphasis can be more pronounced than in standard Arabic, which requires recognition systems to be able to capture these subtle distinctions [40].
- **Phonological Rules Specific to the Quran**: The Quran includes rules for the merging or separation of sounds (e.g., *Idgham* and *Ithar*), which do not always apply to other forms of Arabic speech. These rules contribute to the unique soundscape

of Quranic recitation and must be taken into account when designing ASR systems [41].

The combination of these features makes Quranic recitation a distinct form of speech that presents unique challenges for automatic recognition. The task is not just to recognize phonetic content but also to interpret the prosodic and rhythmic qualities of the recitation, which traditional ASR approaches may not be equipped to handle.

The Role of Tajweed in Quranic Recitation The proper recitation of the Quran is governed by Tajweed, a set of rules that ensures the accurate pronunciation and intonation of the words as they were revealed. Tajweed is considered essential for preserving the linguistic and phonetic integrity of the Quran, as it dictates how each letter should be pronounced and where pauses should be made. The correct application of Tajweed is not only a matter of linguistic precision but also a spiritual obligation, as reciting the Quran improperly can alter its meaning [43].

Learning Tajweed requires dedicated study and is often taught by qualified teachers (*sheikhs*) who have received traditional training in this art. The emphasis on proper pronunciation reflects the cultural reverence for the Quran and the belief that its words should be uttered in the most beautiful and correct manner. Skilled reciters are highly respected, and their recitations are often broadcast on radio and television, especially during the month of Ramadan, further reinforcing the cultural importance of Tajweed [38].

Variations in Recitation Styles The Quran is recited in multiple traditional styles, known as *Qira'at*, which reflect slight variations in pronunciation and articulation. These styles have been passed down through generations, maintaining different authentic methods of recitation that were taught by the Prophet Muhammad. The most commonly practiced styles, such as *Hafs* and *Warsh*, are widely accepted and practiced in different regions of the Muslim world. The existence of these variations demonstrates the Quran's flexibility in accommodating diverse phonetic expressions while preserving the uniformity of its message [40].

1.2.6.5 Components of Tajweed Rules

The rules of Tajweed can be categorized into several main components:

- **Makharij al-Huruf (Articulation Points of Letters):** Makharij refers to the specific points in the vocal tract where each Arabic letter is articulated. The correct pronunciation requires knowing the exact origin of each letter, whether it comes from the throat, tongue, lips, or nasal passage. For instance, the letter " " (ain) is pronounced from the middle throat, while " " (qaf) originates from the back of the tongue touching the soft palate. This precise articulation is crucial for maintaining the authenticity of Quranic recitation [40].

- **Sifat al-Huruf (Characteristics of Letters):** Each Arabic letter has specific characteristics or attributes that affect its sound. These characteristics include *shiddah* (strength), *rikha* (softness), *istifal* (elevation), and *itbaq* (adhesion). Additionally, some letters are pronounced with *tafkhim* (emphasis), resulting in a heavier sound, while others are recited with *tarqiq* (lightness). For example, the letter " "

(daad) is pronounced with emphasis, while " " (seen) is pronounced lightly. Proper knowledge of these characteristics ensures that reciters can accurately produce the distinct sounds required in Tajweed [41, 42].

- **Rules for Nunation and Mim (Nūn Sākinah and Mīm Sākinah):** These rules regulate how the letters " " (nun) and " " (mim) are pronounced when they appear with a sukun (a state of rest) or at the end of words. The rules include *idgham* (assimilation), *ikhfa* (concealment), *izhar* (clarity), and *iqlab* (conversion). For instance, if a *nun sakinah* is followed by a letter that triggers *ikhfa*, it is pronounced with a nasalized tone. These rules contribute to the rhythm and fluidity of Quranic recitation [38].

- **Qalqalah (Echoing):** Qalqalah refers to a slight echoing sound that occurs when certain consonants ,) , , , (appear with a sukun. This rule adds emphasis to the pronunciation, making the recitation clearer and more expressive. The strength of the *qalqalah* varies depending on whether the consonant occurs at the end of a word or within it [43].

- **Rules for Stopping and Continuing (Waqf and Wasl):** Tajweed rules also dictate where a reciter should pause (waqf) or continue (wasl) during recitation to ensure that the meaning is conveyed correctly. Improper pausing can alter the intended meaning of a verse. These rules provide guidance on which words can be joined or where the recitation should stop for proper comprehension. Marks are often included in the Quranic text to indicate preferred stopping points, obligatory stops, or places where continuation is encouraged [40].

1.2.6.6 The Ten Mutawatir Qira'at (Readings)

Definition of Quranic Readings and Their Sources The *Qira'at* (plural of *Qira'ah*) are the various canonical ways of reciting the Qur'an that originated from the Prophet Muhammad's oral recitation. These readings reflect dialectal variations among Arab tribes at the time of revelation. Although they differ in pronunciation, grammar, or word choice, the meanings remain consistent [44, 45].

The term *Mutawatir* refers to transmissions passed down by such a large number of narrators that it is impossible they conspired to fabricate them. The ten Mutawatir Qira'at are all traced back to the Prophet Muhammad through strong, continuous chains of narration [46, 47].

Sources of these readings include:

1. Direct teaching from the Prophet to his companions
2. Compilation during the caliphate of Uthman ibn Affan
3. Oral transmission by qualified scholars over generations
4. Codification by early scholars such as Abu Bakr Ibn Mujahid in his seminal work *Kitab al-Sab'a* [46]

1.2.6.7 The Seven Well-Known Qira'at

Qira'ah of Nāfiᶜ (Narrators: Warsh and Qālūn) Nafi al-Madani (d. 169 AH) was a master reciter from Medina. He studied under 70 of the Tabi'in who had learned from companions like Ubayy ibn Kab and Ibn Abbas [48].

Narrators:

- **Warsh** (ᶜUthmān ibn Saᶜīd al-Miṣrī, d. 197 AH) – spread in North and West Africa
- **Qālūn** (ᶜĪsā ibn Mīnā, d. 220 AH) – used in Libya and Tunisia

Features: Emphasis on the clear pronunciation of hamzah, and the use of *tashīl* (softening) [45].

Qira'ah of Ibn Kathīr (Narrators: al-Bazzī and Qunbul) Ibn Kathīr al-Makkī (d. 120 AH) was Imam of Mecca. He learned from ᶜAbdullah ibn al-Sā'ib and Mujāhid.

Narrators:

- **al-Bazzī** (d. 250 AH)
- **Qunbul** (d. 291 AH)

Features: Characterized by omission of *sakt* (pause) and unique vowel lengthening patterns [47].

Qira'ah of Abū ᶜAmr (Narrators: al-Dūrī and al-Sūsī) Abū ᶜAmr al-Baṣrī (d. 154 AH), from Basra, studied under successors of the companions.

Narrators:

- **al-Dūrī** (also narrator for al-Kisā'ī)
- **al-Sūsī**

Features: Known for *idghām kabīr* (strong assimilation) and use of *imāla* (vowel inclination) [45].

Qira'ah of Ibn ᶜĀmir (Narrators: Hishām and Ibn Dhakwān) Ibn ᶜĀmir (d. 118 AH), chief judge in Damascus during the Umayyad era.

Narrators:

- **Hishām**
- **Ibn Dhakwān**

Features: Unique placement of hamzah and grammatical variations [44].

Qira'ah of ᶜĀṣim (Narrators: Shuᶜbah and Ḥafṣ) ᶜĀṣim ibn Abī al-Najūd (d. 127 AH) from Kufa, studied under Abū ᶜAbd al-Raḥmān al-Sulamī.

Narrators:

- **Shuᶜbah**
- **Ḥafṣ** – the most widespread recitation today

Features: Clarity and ease of pronunciation, distinctive treatment of *idghām* and *ra* [45].

Qira'ah of Ḥamzah (Narrators: Khalaf and Khallād) Ḥamzah al-Zayyāt (d. 156 AH), a Kufan reciter, learned from chains linked to Ibn Masᶜūd.

Narrators:

- **Khalaf** – also recognized as the tenth reader
- **Khallād**

Features: Frequent use of *sakt*, detailed *imāla* patterns, unique treatment of hamzah [47].

Qira'ah of al-Kisā'ī (Narrators: al-Layth and Ḥafṣ al-Dūrī) Al-Kisā'ī (d. 189 AH), grammarian and court tutor, learned from Ḥamzah.

Narrators:

- **al-Layth**
- **Ḥafṣ al-Dūrī** – also transmitted Abū ᶜAmr's reading

Features: Special rules for *waqf* (pausing) and extensive use of *imāla* [45].

1.2.6.8 The Three Additional Qira'at Completing the Ten

Qira'ah of Abū Jaᶜfar (Narrators: Ibn Wardān and Ibn Jammaz) Abū Jaᶜfar (d. 130 AH), from Medina, learned from Ibn ᶜAbbās and Abū Hurayrah.

Narrators:

- **Ibn Wardān**
- **Ibn Jammaz**

Features: Use of *ṣilah* (word-joining), unique *madd* (lengthening), and consonantal pronunciation [44].

Qira'ah of Yaᶜqūb (Narrators: Ruways and Rawḥ) Yaᶜqūb al-Ḥaḍramī (d. 205 AH) from Basra, studied under students of Abū Mūsā al-Ashᶜarī.

Narrators:

- **Ruways**
- **Rawḥ**

Features: Distinctive *idghām* rules and glottal stop pronunciation [45].

Qira'ah of Khalaf al-ᶜĀshir (Narrators: Isḥāq al-Marwazī and Idrīs al-Ḥaddād) Khalaf ibn Hishām (d. 229 AH), initially a narrator of Ḥamzah, later became a canonical reader.

Narrators:

- **Isḥāq al-Marwazī**
- **Idrīs al-Ḥaddād**

Features: Similar to Ḥamzah's reading with over 120 differences, mainly in vowel and consonant treatment [47].

1.2.6.9 The Significance and Wisdom Behind the Diversity of Qira'at

The multiplicity of Qira'at reflects divine wisdom:

- **Ease and Accessibility:** Addressing dialectal differences across Arab tribes
- **Semantic Depth:** Complementary meanings enhance understanding
- **Preservation:** Cross-verification between readings ensured textual integrity
- **Linguistic Richness:** Showcases the flexibility and expressiveness of the Arabic language [44, 45]

1.2.6.10 Distinctive Features of Warsh's Recitation

The narration of Warsh from Nafi' is one of the prominent Quranic recitations and is characterized by unique features in pronunciation and performance, which give it a distinct nature among other narrations. These characteristics are evident in several aspects:

- **Hamzah (Glottal Stops):** Warsh tends to facilitate the glottal stops by either softening, omitting, or transferring their vowel sounds to a preceding silent letter. This affects the pronunciation of certain words and gives the recitation a distinct fluidity.
- **Prolongation (Madd):** Warsh is known for his tendency to extend the prolongations, especially in both the connected and separate prolongations, often extending them to six movements. This adds a unique beauty and expressiveness to the recitation.
- **Imālah and Taqleel (Vowel Modification and Softening):** Warsh uses *Imālah* and *Taqleel* in some words, which introduces variation in pronunciation and reflects a richness in performance and recitation styles.
- **Mīm al-Jamᶜ and Hāᵓ al-Kināyah (The Collective Mīm and the Pronominal Hāᵓ):** Warsh follows specific rules in pronouncing the collective *Mīm* and the pronominal *Hāᵓ*, such as rounding the Mīm and placing a Wāw in the middle when followed by a Hamzah of interruption. This influences how words are connected and emphasizes his distinct style of recitation.

These features make Warsh's narration distinctive and give the recitation a special character that is appreciated in many Islamic communities, especially in the Maghreb region, where this narration is widely practiced. The table below presents a detailed breakdown of tajweed rules and pronunciation features specific to the Warsh recitation:

1.2.6.11 Key Differences Between Warsh and Hafs Readings

The two most commonly recited versions of the Qur'an today are those transmitted by Warsh and Ḥafṣ. Both are considered canonical and trace back to reputable chains of narration. However, they differ in several phonetic, grammatical, and lexical aspects. These variations are not contradictions but rather complementary traditions reflecting the rich oral transmission of the Qur'an. Table 1.2.6.11 presents a comparative overview of some key differences between the Warsh and Ḥafṣ readings.

Note	Example (Warsh)	Rule	Category
Natural prolongation without interference	(Sūrat al-Baqarah 2:285)	Madd al-Tabiᶜī (natural prolongation)	**Natural Madd**
Sound of "shaking" in letters of Qalqalah	(Sūrat al-Qamar 54:49)	Qalqalah (vibration)	**Qalqalah**
Nasal sound in letters	(Sūrat al-Fajr 89:10)	Ghunna (nasalization)	**Ghunna**
Concealing the letter with light sound	(Sūrat al-Baqarah 2:261)	Ikhfāʾ (concealment)	**Ikhfāʾ**
Merging letters smoothly	(Sūrat al-Baqarah 2:261)	Idghām (merging)	**Idghām**
Clear pronunciation without merging	(Sūrat al-Fil 105:1)	Iz'hār (clarification)	**Iz'hār**
Slight change of vowel sound	(Sūrat al-Fatiḥa 1:1)	Imlāʾ (slanting)	**Imlāʾ**
Opening sound of the letter	(Sūrat al-Ikhlas 112:1)	Fath (open sound)	**Fath**
Emphasis on the sound of certain letters	(Sūrat al-Rahman 55:13)	Tafkhīm (emphasis)	**Tafkhīm**
Emphasizing the sound of Rāʾ with Dammah or Fathah	(Sūrat al-Fatiḥa 1:2)	Tafkhīm al-Rāʾ (emphasis on Rāʾ)	**Emphasis on Rāʾ**
Softening the sound of Rāʾ with Kasrah	(Sūrat al-Mumtahina 60:10)	Tarqīq al-Rāʾ (softening Rāʾ)	**Softening of Rāʾ**
Replacing Hamzah with another letter	(Sūrat al-Baqarah 2:285)	Ibdāl al-Hamzah (substitution of Hamzah)	**Substitution of Hamzah**
Long prolongation of the vowel sound	(Sūrat al-Baqarah 2:285)	Madd al-Tawīl (long prolongation)	**Long Madd**
Prolonging when the Hamzah comes after a vowel	(Sūrat al-Baqarah 2:261)	Madd al-Badl (prolongation with Hamzah)	**Substitution Madd**
Shifting the sound in a subtle way	(Sūrat al-Baqarah 2:261)	Naql (shift)	**Shift**
Easing the transition between vowels	(Sūrat al-Baqarah 2:285)	Tashīl (simplification)	**Simplification**
The sound of Mīm in joined words	(Sūrat al-Mumtahina 60:10)	Mīm al-Jamᶜ (the joining Mīm)	**Mīm of Joining**
The sound of Hāʾ representing an indirect object	(Sūrat al-Baqarah 2:255)	Hāʾ al-Kināyah (kināyah Hāʾ)	**Kināyah Hāʾ**

Table 1.1: Tajweed Examples and Rules in Warsh Recitation

1.2.7 Complexity of Tajweed Rules and Their Impact on Speech Recognition

The recitation of the Quran follows a detailed set of phonetic rules known as Tajweed, which govern the correct pronunciation of Arabic letters and various phonetic features.

Aspect	Warsh Reading	Hafṣ Reading
Vowel Lengthening (Madd)	Applies longer madd in some cases (e.g., *madd munfasil*)	Uses shorter or medium-length madd in similar cases
Hamzah Pronunciation	Tends to simplify or soften hamzah using *tashīl*	Preserves clear and distinct hamzah sounds
Inclination and Reduction (Imāla and Qasr)	Uses imāla (e.g., tilting vowel sounds like $\bar{a}$ to $\bar{e}$) in certain words	Generally avoids imāla and leans toward full vowel articulation
Wording and Specific Terms	Some word differences, such as "*maliki*" vs. "*maaliki*" in Al-Fatiha	May use different accepted variant, often based on other authentic transmissions

Table 1.2: Key Differences Between Warsh and Hafṣ Readings

These rules extend beyond mere technical aspects, encompassing the linguistic and artistic dimensions of recitation. Understanding and adapting to these rules present unique challenges for automatic speech recognition (ASR) systems due to their complexity.

1.2.7.1 Understanding Tajweed Rules

Tajweed consists of multiple guidelines addressing pronunciation, such as the articulation points of letters (makhārij), elongation (madd), nasalization (ghunnah), and specific pausing methods (waqf). Each Arabic letter has a distinct articulation point and characteristic sound, requiring precise pronunciation during recitation. For instance, the letters (qaf) and (kaf) are articulated from different areas of the throat and mouth, influencing their acoustic representation in ASR systems.

Moreover, Tajweed includes intricate aspects such as idgham (merging of sounds), iqlab (sound changes), and ikhfa (concealing of sounds), which depend on the phonetic context in which the letters appear. These rules add depth to the recitation but also introduce significant variability in pronunciation, creating a distinct linguistic challenge for ASR systems.

1.2.7.2 Impact on Speech Recognition Technology

The complexity of Tajweed rules introduces several hurdles for ASR systems aiming to recognize Quranic recitation effectively:

- **Phonetic Variability**: The rules cause subtle variations in pronunciation based on context, which can be difficult for ASR systems to capture. While such systems often rely on extensive training datasets, they may lack the specialized vocabulary and pronunciation nuances required for accurate recognition of Quranic recitation, leading to higher rates of misrecognition.

- **Contextual Dependencies**: Tajweed rules are context-dependent; the pronunciation of a letter can change depending on its position within a word and the surrounding phonetic environment. This adds complexity to the modeling and training of ASR systems, as they must account for these dependencies to ensure accurate recognition.

- **Acoustic Challenges**: The melodic and rhythmic aspects of Quranic recitation, which are integral to Tajweed, introduce additional variability in the speech signal. Variations in pitch, tone, and intonation can make it challenging for ASR systems to distinguish between similar-sounding phonemes, adversely affecting their performance.
- **Training Data Limitations**: The scarcity of annotated datasets focused on Tajweed-compliant recitation limits the ability of ASR systems to learn the phonetic intricacies required. Most existing models are trained on general Arabic speech data, which may not adequately represent the unique characteristics of Quranic recitation.

1.2.7.3 Pronunciation Variability Among Reciters

Pronunciation variability is a significant challenge in the automatic recognition of Quranic recitation, impacting the efficacy of automatic speech recognition (ASR) systems. Each reciter introduces a unique blend of phonetic and stylistic characteristics influenced by various factors such as regional dialects, personal interpretation of Tajweed rules, and individual linguistic backgrounds.

Influence of Regional Dialects Arabic, as a language, boasts a rich tapestry of dialects that vary considerably across different regions. These dialectal differences manifest in the pronunciation of specific letters and sounds, which can profoundly affect the ASR systems designed for Quranic recitation. For instance, the letter (jeem) may be pronounced as /d / in the Egyptian dialect, whereas in some Gulf dialects, it might be articulated as / /. Such variations complicate the recognition process, as ASR systems trained on standardized Arabic may struggle to accommodate these dialectal distinctions. Consequently, regional accents can lead to increased misrecognition rates, especially if the training data lacks sufficient representation of various dialects.

Personal Interpretation and Style Beyond regional influences, individual reciters often develop distinctive styles and interpretations of recitation. This personal style is shaped by factors including the reciter's educational background, age, and exposure to different schools of thought regarding Tajweed. For example, a seasoned reciter may emphasize certain elongation rules (madd) more than a novice reciter, leading to variability in how words are pronounced. Similarly, some reciters might merge sounds differently (idgham), resulting in subtle yet significant differences in pronunciation. These individual stylistic choices introduce additional layers of complexity for ASR systems, which may find it challenging to generalize across diverse recitation styles.

Variability in Tajweed Application The application of Tajweed rules is not uniform among all reciters. Some individuals may strictly adhere to the rules, while others might interpret them more flexibly, particularly in less formal recitation settings. This inconsistency creates a broad spectrum of pronunciations that ASR systems must navigate. For instance, the application of nasalization (ghunnah) can vary in emphasis, resulting in differing acoustic patterns that complicate the recognition process. Moreover, reciters may pause or elongate sounds in ways that are personally meaningful to them, which can deviate from standardized pronunciation. This variability underscores the necessity for ASR systems to accommodate a wide range of phonetic manifestations.

Impact on ASR System Performance The pronounced variability in pronunciation among reciters directly contributes to a higher error rate in ASR systems designed for Quranic recitation. Traditional ASR models typically excel with standardized speech but falter when confronted with the unpredictability inherent in the diverse pronunciations of Quranic recitation. As a result, these systems may misidentify phonemes, leading to inaccuracies in transcription and recognition. The challenge is exacerbated by the fact that many ASR systems rely on large datasets that may not adequately represent the full spectrum of recitation styles or the intricate nuances of Tajweed.

1.2.7.4 Phonetic Particularities of the Quran Compared to Standard Arabic

The phonetic characteristics of the Quran present distinct features that differentiate it from Standard Arabic. These differences arise not only from the linguistic structure of the Arabic language but also from the unique aesthetic and rhythmic qualities associated with Quranic recitation. Understanding these phonetic particularities is crucial for developing effective automatic speech recognition (ASR) systems that can accurately process Quranic texts.

Articulation and Pronunciation Variability One of the key phonetic features that set Quranic Arabic apart is the specific articulation of certain phonemes. The rules of Tajweed, which govern the correct pronunciation of the Quran, introduce a level of complexity not found in Standard Arabic. For instance, the articulation points of letters (makhārij) dictate how each letter is produced, requiring reciters to pay close attention to the physical movements of the tongue and mouth. The letter (qaf) is articulated from a deeper part of the throat compared to (kaf), leading to distinctive acoustic properties that ASR systems must learn to recognize.

Moreover, Tajweed encompasses several nuanced rules that affect pronunciation, such as elongation (madd), nasal sounds (ghunnah), and emphasis (tafkhīm). These rules not only alter the duration and quality of vowels but also influence how consonants are pronounced in different contexts. For instance, the nasalization of certain sounds adds a layer of complexity to the phonetic structure, as it requires precise control over airflow and sound production. Such intricacies create a rich auditory landscape in Quranic recitation, which can be difficult for ASR systems to model accurately.

Contextual Phonetic Variability The context in which phonemes occur plays a significant role in how they are pronounced in Quranic recitation. This variability is often dictated by the application of Tajweed rules, which can change the pronunciation of a letter based on its surrounding phonetic environment. For example, the phenomenon of idgham (merging sounds) allows certain letters to blend together when recited in succession, resulting in altered pronunciations that may not be evident in Standard Arabic.

Additionally, the melodic aspects of Quranic recitation introduce further variability. Reciters often employ different styles—such as Mujawwad (elongated) or Murattal (moderate)—which affect pitch, intonation, and rhythm. This variability complicates the task for ASR systems, which must be equipped to handle the diverse phonetic expressions that arise from these different recitation styles.

Impact on Automatic Speech Recognition The unique phonetic characteristics of Quranic recitation pose several challenges for ASR systems. Traditional speech recog-

nition models are typically trained on standard Arabic speech datasets, which may not encompass the specific phonetic variations present in Quranic recitation. This can lead to misrecognition and decreased accuracy when processing Quranic texts.

One significant challenge is the increased phonetic variability inherent in Quranic recitation. Due to the rules of Tajweed and the emphasis on melodic features, the same word may be pronounced differently depending on context, style, or reciter. ASR systems that do not account for this variability are likely to struggle with accurate recognition, leading to a higher rate of errors. Furthermore, the absence of comprehensive, annotated training datasets tailored to Quranic recitation limits the ability of these systems to learn the necessary phonetic distinctions.

1.2.7.5 Lack of Resources and Annotated Databases for Model Training

The advancement of automatic speech recognition (ASR) technologies tailored for Quranic recitation faces a critical obstacle due to the lack of comprehensive resources and adequately annotated databases. This scarcity significantly hampers the ability of researchers and developers to create efficient and accurate ASR systems that can effectively handle the unique phonetic and linguistic features of Quranic Arabic.

The Importance of Annotated Datasets Annotated datasets are the backbone of any machine learning application, including ASR. These datasets serve as the foundation upon which models learn to recognize and process spoken language. In the realm of ASR, particularly for Quranic recitation, the quality, size, and specificity of the datasets are vital. However, the availability of such datasets is notably limited compared to those developed for standard Arabic or other widely spoken languages.

Most existing datasets focus on general Arabic speech, which often fails to encapsulate the specificities of Quranic recitation, such as the nuanced pronunciation variations governed by Tajweed rules. Consequently, ASR systems trained on these datasets may lack the necessary capabilities to accurately recognize and transcribe Quranic recitation, leading to increased error rates and decreased reliability.

Challenges in Data Collection The process of gathering annotated data for Quranic recitation presents several challenges. Firstly, Quranic recitation is a highly specialized practice that involves not only linguistic knowledge but also a deep understanding of cultural and religious contexts. Collecting data from a diverse array of reciters, each with their unique interpretation and style, is essential to developing robust ASR systems. However, this diversity introduces complications in standardizing recordings and annotations.

Additionally, the artistry involved in Quranic recitation requires capturing various styles and nuances, including variations in pitch, rhythm, and intonation. This complexity demands a meticulous approach to both recording and annotating, necessitating collaboration between linguists, religious scholars, and ASR specialists. Unfortunately, such interdisciplinary collaboration is often challenging to achieve, resulting in limited data collection efforts.

Consequences of Limited Data The consequences of having insufficient annotated datasets are profound. ASR systems developed without access to extensive and diverse training data may exhibit poor generalization capabilities. These systems struggle to

adapt to the various pronunciation styles and Tajweed rules that characterize Quranic recitation. As a result, recognition errors become more frequent, particularly in cases where reciters deviate from the "standard" forms of pronunciation or employ stylistic variations unique to their training.

Moreover, the lack of resources can stifle innovation within the field of ASR technology for Quranic recitation. Researchers may find it difficult to experiment with new algorithms or modeling techniques without the foundational datasets needed to evaluate performance. This stagnation limits advancements that could improve recognition accuracy and expand the applicability of ASR systems in educational and religious contexts.

Sensitivity to Tone and Modulations Specific to Each Recitation Style The recitation of the Quran is characterized by a distinctive use of tone and modulation, which are essential aspects of its expressive and melodic nature. These features are not merely stylistic but are integral to the proper conveyance of meaning, emotion, and spiritual depth in the recitation. Each recitation style, such as *Mujawwad*, known for its slow, deliberate delivery, or *Murattal*, which is more rhythmic and consistent, brings unique tonal variations and modulative patterns. The sensitivity to these elements presents specific challenges for automatic speech recognition (ASR) systems aimed at accurately transcribing Quranic recitation.

The Importance of Tone and Modulation in Quranic Recitation Tone involves the pitch variations that occur during recitation, while modulation refers to the dynamic changes in pitch, volume, and rhythm. In Quranic recitation, these elements are used to highlight certain words or phrases, evoke particular emotions, and adhere to the recitation rules established by Tajweed. The choice of tone and modulation often depends on the reciter's style, their interpretation of the verses, and the intended impact on the listener. These variations serve not only an aesthetic purpose but also convey subtleties in meaning and expression that are essential to the recitation's spiritual and linguistic integrity.

1.2.7.6 Challenges for ASR Systems

The presence of diverse tonal and modulative features introduces several difficulties for ASR systems designed for Quranic recitation:

- **Phonetic Variability**: The way in which tonal changes are applied can lead to significant phonetic variability across different recitation styles and even among individual reciters. For example, the pronunciation of a specific phoneme may be lengthened or articulated with a varying intensity depending on the style. This phonetic diversity is challenging for ASR models that are typically trained on more uniform speech data, such as standard Arabic, which lacks the same range of tonal intricacies. This discrepancy can lead to a higher rate of misrecognition, particularly when the system encounters less common styles of recitation [49].
- **Contextual Interpretation Issues**: In Quranic recitation, tonal and modulative changes are closely linked to the meaning and interpretation of the text. Variations in pitch and rhythm can emphasize certain words or phrases, guiding the listener's understanding of the verse. An ASR system that does not account for

these variations may misinterpret the intended emphasis or fail to distinguish between similarly pronounced words, leading to errors in transcription that could alter the meaning of the recited text [50].

- **Limitations in Training Data**: Training data for ASR models usually consists of general Arabic speech, which may not include sufficient examples of Quranic recitation with its distinctive tonal patterns. This limitation hampers the system's ability to learn the fine-grained features needed to accurately recognize Quranic recitation. The scarcity of annotated datasets that capture various recitation styles and the corresponding tonal characteristics further exacerbates this issue [51].

1.3 Conclusion

Despite the considerable challenges facing automatic speech recognition of Quranic recitation—including the complexity of Tajweed rules, pronunciation variability among reciters, and the scarcity of comprehensive annotated datasets—the potential benefits make this pursuit worthwhile. The unique phonetic landscape of Quranic Arabic, with its context-dependent pronunciation rules and melodic elements, requires specialized approaches that go beyond conventional ASR techniques. Future progress will depend on collaborative efforts between computer scientists, linguists, and Quranic scholars to develop better training resources and recognition models that can capture the subtle nuances of this centuries-old oral tradition, ultimately serving educational, preservation, and accessibility purposes.

Chapter 02
Speech Recognition Techniques

Chapter 2

Speech Recognition Techniques

2.1 Introduction

Speech recognition has come a long way, moving from basic rule-based systems to advanced deep learning methods. This chapter highlighted how these techniques have improved the ability of machines to process and understand spoken language. Applying them to Quranic recitation is especially valuable, as it helps preserve a rich religious tradition. We saw how research has tackled challenges like complex pronunciation rules and multiple recitation styles.

2.2 Traditional Speech Recognition Techniques

Speech recognition has evolved significantly over the years, with various techniques developed to improve the accuracy and efficiency of recognizing spoken language. One of the foundational techniques in speech recognition is the Gaussian Mixture Model (GMM), which serves as a probabilistic approach for modeling vocal features.

2.2.1 Gaussian Mixture Models (GMM): Probabilistic Techniques for Modeling Vocal Features

Gaussian Mixture Models (GMMs) are a statistical method widely used in speech recognition for modeling the distribution of acoustic features extracted from speech signals. A GMM assumes that the acoustic feature vector is generated from a mixture of several Gaussian distributions, each representing a distinct phonetic unit or sound segment in the speech signal.

Mathematical Foundation The GMM is defined mathematically as a weighted sum of multiple Gaussian distributions:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{2.1}$$

where:

- $p(x)$ is the probability density function of the feature vector x,

- K is the number of Gaussian components,
- π_k are the mixture weights (with $\sum_{k=1}^{K} \pi_k = 1$),
- $\mathcal{N}(x|\mu_k, \Sigma_k)$ is the Gaussian distribution with mean μ_k and covariance Σ_k.

The parameters of the GMM (means, covariances, and weights) can be estimated using the Expectation-Maximization (EM) algorithm, which iteratively refines the estimates to maximize the likelihood of the observed data.

Feature Extraction In speech recognition, features such as Mel-frequency cepstral coefficients (MFCCs) are commonly extracted from audio signals to represent vocal characteristics. These features capture the essential information about the speech signal while reducing dimensionality. Once extracted, GMMs are trained on these features to learn the statistical properties of different phonetic units.

Advantages and Limitations One of the main advantages of GMMs is their ability to effectively model the complex and multimodal distribution of speech features, capturing variations due to different speakers, accents, and speaking styles. Additionally, GMMs can be easily integrated with other statistical modeling techniques, such as HMMs.

However, GMMs also have limitations. They assume that the feature distribution can be adequately represented by a finite number of Gaussian components, which may not always hold true in practice. Moreover, training GMMs requires a substantial amount of labeled data, and their performance can degrade with insufficient training samples.

2.2.2 Hidden Markov Models (HMM): Sequential Probability-Based Approaches for Modeling Phoneme Sequences

Hidden Markov Models (HMMs) are a crucial statistical tool in the field of speech recognition. They provide a probabilistic framework for modeling sequences of phonemes, which are the basic units of sound in speech. HMMs are particularly effective for tasks that involve time-dependent data, such as spoken language, where the order of phonemes matters significantly.

Mathematical Framework An HMM is defined by:

- A set of states $S = \{s_1, s_2, \ldots, s_N\}$, representing distinct phonetic units or hidden states.
- A set of observation symbols $V = \{v_1, v_2, \ldots, v_M\}$, corresponding to the feature vectors derived from the acoustic signals (e.g., Mel-frequency cepstral coefficients (MFCCs)).
- Transition probabilities $A = \{a_{ij}\}$, where $a_{ij} = P(s_j|s_i)$ denotes the probability of transitioning from state s_i to state s_j.
- An initial state distribution $\pi = \{\pi_i\}$, indicating the probability of starting in each state s_i.
- Emission probabilities $B = \{b_i(v_t)\}$, where $b_i(v_t) = P(v_t|s_i)$ gives the probability of observing the feature vector v_t given that the model is in state s_i.

The primary goal of using HMMs is to find the most likely sequence of hidden states that could have generated the observed sequence of acoustic feature vectors. This process involves two major computational problems: training the HMM parameters using a labeled dataset and decoding the most probable state sequence given an observation sequence.

Training HMMs The training of HMMs is typically performed using the Baum-Welch algorithm, a specific instance of the Expectation-Maximization (EM) technique. This algorithm iteratively adjusts the model parameters (transition and emission probabilities) to maximize the likelihood of the observed training data. A detailed explanation of the algorithm can be found in [52].

Decoding with HMMs When recognizing speech, the Viterbi algorithm is used to compute the most likely sequence of hidden states given the observed feature vectors. This algorithm efficiently finds the optimal path through the state space by maintaining a dynamic programming table, which significantly reduces computational complexity compared to a brute-force approach.

Advantages and Limitations The primary advantages of HMMs include their ability to model temporal sequences and handle variable-length input data. However, they do have limitations, such as the Markov assumption, which can oversimplify the relationships between phonemes, leading to potential inaccuracies in modeling. Additionally, training HMMs requires substantial labeled data, which may not always be available.

Despite the rise of deep learning techniques in recent years, HMMs remain relevant in modern speech recognition frameworks, often being integrated with neural network models to leverage their sequential modeling capabilities while improving the robustness of acoustic feature learning [53].

2.2.3 Rule-Based Methods: Use of Linguistic Models Based on Expert Knowledge

Rule-based methods in speech recognition represent one of the foundational approaches to modeling and interpreting spoken language, prevalent before the era of neural networks and statistical learning. These methods rely on linguistic rules and expert knowledge to analyze speech, incorporating phonetic, syntactic, and semantic information.

Fundamentals of Rule-Based Methods Rule-based systems utilize a set of predefined linguistic rules to process and analyze spoken language. Key components of rule-based methods include:

- **Phonetic Rules:** These rules define how phonemes can be combined to form words and how variations occur in different phonetic contexts.
- **Grammatical Rules:** These rules govern the structure of sentences, enabling the system to parse and interpret spoken input based on grammatical conventions.
- **Lexical Knowledge:** A comprehensive lexicon is essential, containing words, their phonetic representations, and related rules regarding stress and intonation patterns.

- **Contextual Rules:** These rules account for context, refining recognition accuracy by considering speaker identity and conversational context.

The integration of these components allows rule-based systems to interpret speech input effectively and generate corresponding textual output.

Advantages of Rule-Based Methods Rule-based approaches offer several advantages:

- **Transparency:** The explicit nature of rules allows for a clear understanding of how the system processes speech, making it easier to debug and refine.
- **Customization:** Rule-based systems can be tailored to specific domains, such as medical or legal transcription, where specialized vocabulary is common [54].
- **Low Data Dependency:** Since these systems rely on linguistic expertise rather than large datasets, they can perform effectively with limited training data [55].

Limitations of Rule-Based Methods Despite their strengths, rule-based methods have notable limitations:

- **Scalability Issues:** As vocabulary size increases, maintaining and updating rules becomes increasingly complex [54].
- **Lack of Flexibility:** Rule-based systems often struggle to adapt to varied speech inputs, leading to limitations in handling natural language variability, such as accents and dialects [56].
- **Difficulty in Capturing Linguistic Nuances:** The fixed nature of rules makes it challenging to model the inherent variability and nuances of human speech effectively [57].

2.3 Modern Speech Recognition Techniques

In this section, we provide a detailed explanation of the architectures of various models used in modern speech recognition, including Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM), Transformers, Attention Mechanisms, and Hybrid Ensemble Learning Systems. Each model is explained with its respective architecture diagram.

2.3.1 Deep Neural Networks (DNNs)

DNNs are a class of artificial neural networks with multiple layers between the input and output. They are used to model complex relationships in data and are a fundamental building block for modern speech recognition systems.

The DNN architecture consists of an input layer that receives feature vectors (e.g., MFCCs), several hidden layers that perform non-linear transformations, and an output layer that generates probabilities for each possible transcription. The network is trained by minimizing a loss function such as cross-entropy to predict the correct transcription from an input speech feature.

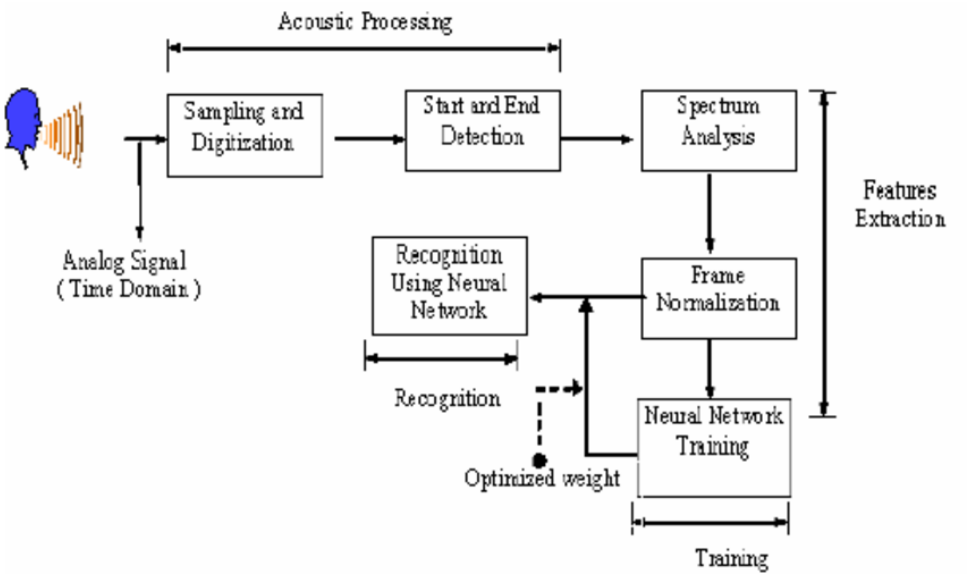

Figure 2.1: Architecture of a Deep Neural Network (DNN) used for Speech Recognition [1].

2.3.2 Convolutional Neural Networks (CNNs)

CNNs are particularly effective in speech recognition for extracting hierarchical features from spectrograms or mel-spectrograms of audio signals. They utilize convolutional layers that apply filters to detect patterns like phonemes in the speech signal.

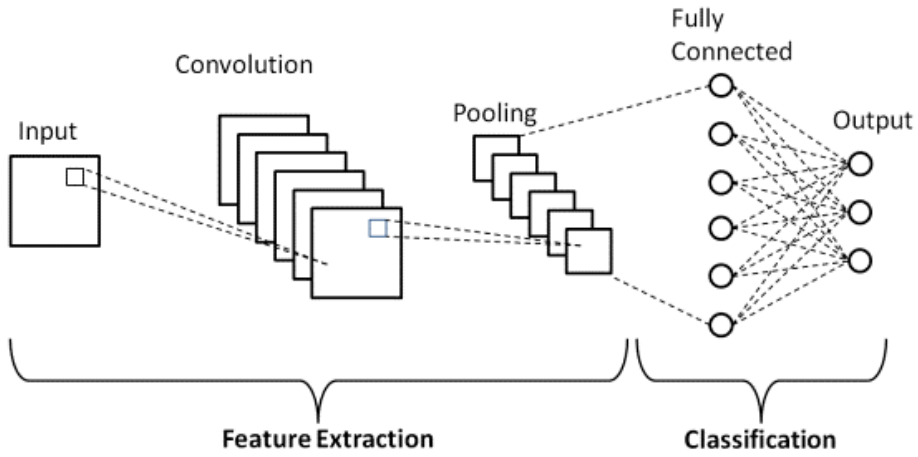

Figure 2.2: A simple architecture of a Convolutional Neural Network (CNN) [2].

The CNN architecture consists of an input layer, followed by convolutional layers, pooling layers, and fully connected layers. The convolutional layers extract local features from the spectrogram, while the pooling layers reduce dimensionality. The fully connected layers combine these features to make predictions.

2.3.3 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

RNNs and LSTMs are designed for sequence-based tasks, which makes them ideal for speech recognition where temporal dependencies exist between audio frames. LSTMs, in particular, solve the vanishing gradient problem that occurs in standard RNNs [15, 58].

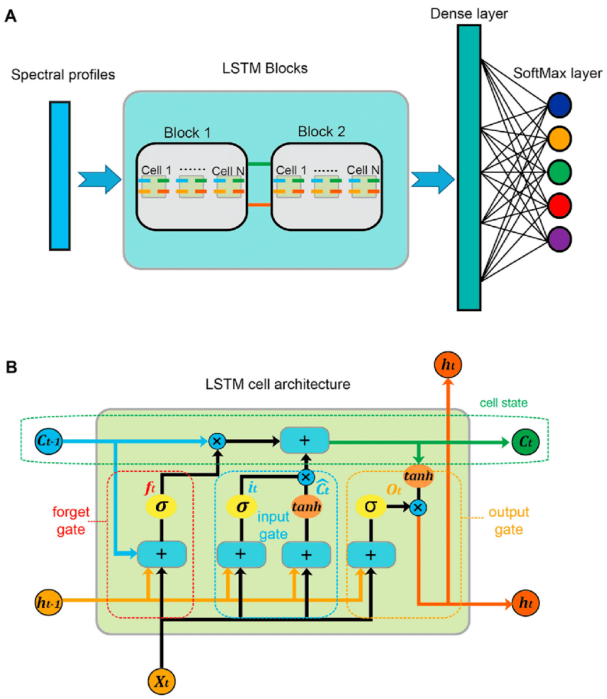

Figure 2.3: Architecture of a Long Short-Term Memory (LSTM) [3]

The LSTM architecture includes an input layer, forget, input, and output gates that regulate the flow of information. The memory cell stores information over time, allowing the network to capture long-term dependencies.

2.3.4 Gated Recurrent Unit (GRU)

The Gated Recurrent Unit (GRU), introduced by Cho et al. in 2014 [59], is a type of recurrent neural network (RNN) designed to model sequential data efficiently, especially in tasks such as speech recognition and time-series prediction. GRUs simplify the traditional LSTM architecture by using only two gates: the update gate, which determines how much past information is retained, and the reset gate, which controls how much of the previous state is forgotten. Unlike LSTM, GRUs merge the memory cell and hidden state into a single vector, resulting in fewer parameters and faster training while still mitigating the vanishing gradient problem.

2.3.5 Transformers

Transformers rely on self-attention mechanisms to process sequences in parallel. This architecture is efficient at capturing long-range dependencies and learning complex relationships in speech data. The Transformer architecture uses an encoder-decoder structure.

The encoder applies self-attention to the input, and the decoder generates the output sequence based on attention over the encoder outputs [13].

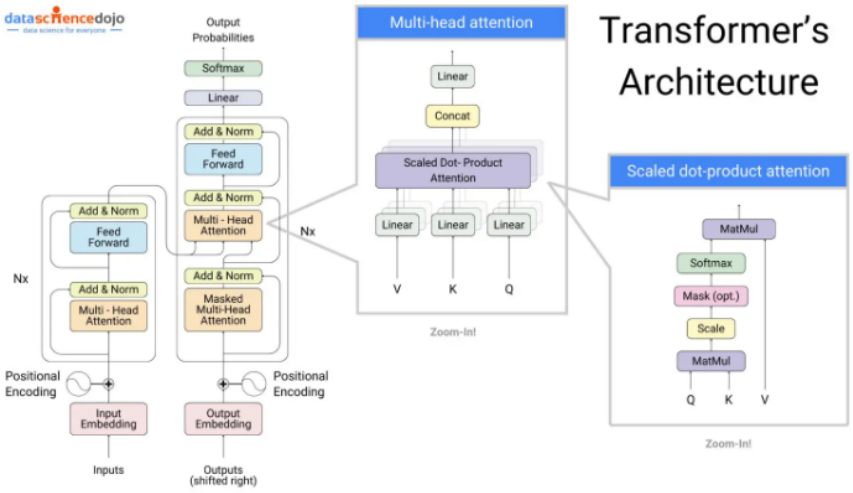

Figure 2.4: Architecture of a Transformer [4]

The Transformer architecture uses an encoder-decoder structure. The encoder applies self-attention to the input, and the decoder generates the output sequence based on attention over the encoder outputs.

2.3.6 Attention Mechanisms

The attention mechanism allows the model to focus on the most important parts of the input sequence. This is especially useful in speech recognition, where some parts of the signal carry more weight than others [13].

The attention mechanism assigns weights to input elements based on their relevance to the current output step. These weights enhance the influence of key features on model predictions.

2.3.7 Ensemble Learning Methods

Ensemble methods are a class of machine learning techniques that combine multiple models to improve predictive accuracy, robustness, and generalization. Instead of relying on a single model, ensemble learning aggregates the outputs of several weak learners to achieve better performance. This approach is especially effective in handling complex datasets and helps reduce overfitting, thereby improving both classification and regression tasks [60].

2.3.7.1 Types of Ensemble Methods

Bagging (Bootstrap Aggregating) Bagging is an ensemble learning technique that enhances the accuracy and stability of machine learning algorithms. It follows these steps [61]:

- **Data Sampling:** Multiple subsets of the training data are created using bootstrap sampling (random sampling with replacement).

- **Model Training:** A separate model is trained on each subset.
- **Aggregation:** The outputs from all models are combined (averaged for regression, majority voting for classification) to form the final prediction.

Key Benefits:

- Reduces variance and overfitting.
- Improves accuracy by combining diverse predictions [62].

Random Forest Random Forest is a powerful implementation of bagging based on decision trees [63]:

- Creates multiple decision trees using random subsets of the training data and features.
- Aggregates predictions from individual trees to form the final output.
- Handles both classification and regression tasks effectively.
- Reduces variance and improves accuracy through averaging [64].

Boosting Boosting sequentially trains models, with each model focusing on correcting the errors of its predecessors [65]:

- **Sequential Training:** Each model learns from the mistakes of the previous model.
- **Weight Adjustment:** Higher weights are assigned to misclassified instances to emphasize them in the next iteration.
- **Model Combination:** Predictions are combined via weighted averaging or voting.

Key Benefits:

- Reduces bias by focusing on hard-to-classify examples.
- Produces strong predictors from weak learners [66].

Gradient Boosting

- Builds models sequentially, with each new model correcting the errors of the previous ensemble.
- Fits to the negative gradient of the loss function [67].
- Optimizes performance through additive modeling.

AdaBoost (Adaptive Boosting)

- Iteratively adjusts the weights of misclassified training examples.
- Focuses subsequent models on difficult cases.
- Combines all models weighted by their accuracy for final predictions [65].

XGBoost (Extreme Gradient Boosting)

- Optimized implementation of gradient boosting designed for speed and performance.
- Uses advanced regularization (L1, L2) to prevent overfitting.
- Incorporates efficient tree-based learning with automatic handling of missing values and sparse data.
- Highly scalable with parallelized tree construction.
- Offers tunable hyperparameters such as learning rate, tree depth, and number of boosting rounds [68].

LightGBM

- Gradient boosting framework using tree-based learning algorithms.
- Offers faster training speed and higher efficiency with lower memory usage.
- Implements Gradient-based One-Side Sampling and Exclusive Feature Bundling.
- Grows trees leaf-wise rather than level-wise for better accuracy with fewer resources [69].

Stacking (Stacked Generalization) Stacking combines multiple base models to form a more accurate meta-model [70]:

- **Base Models:** Several models (level-0) are trained on the original dataset.
- **Meta-Model:** A new model (level-1) is trained on the outputs of the base models to produce the final prediction.

Key Benefits:

- Leverages model diversity.
- Learns to optimally combine predictions [71].

Voting and Averaging These techniques provide straightforward methods to combine model predictions [72].

Voting (for classification)

- **Hard Voting:** Selects the majority vote among models.
- **Soft Voting:** Averages predicted probabilities across models.
- Leverages collective intelligence to improve classification accuracy [73].

Averaging (for regression)

- Uses the mean or weighted average of predictions from multiple regression models.
- Smooths individual model errors.
- Produces more stable and accurate predictions than any single model [74].

2.4 Existing Research in Quranic Recitation Recognition

Quranic recitation recognition has made significant progress over the years, thanks to advances in speech recognition, deep learning, and AI techniques. This section presents a comprehensive review of research in various aspects of Quranic recitation recognition, including the foundational methods, deep learning applications, recitation style analysis, error detection, and recent innovations in the field.

2.4.1 Early Works and Foundational Studies

In the early years of Quranic recitation recognition, researchers employed classical machine learning techniques to model phonetic variations and apply Tajweed rules. Early studies mainly used small datasets and simpler methods such as Support Vector Machines (SVM), Multi-Layer Perceptron (MLP), and Hidden Markov Models (HMM) to recognize basic Tajweed rules and detect errors in Quranic recitation. This table includes research from 2010 to 2017 that used classical machine learning techniques such as MLP, SVM, and HMM for Tajweed recognition and error detection. These studies mainly focused on small datasets and basic Tajweed rules like Qalqalah and Hafiz verification.

Research	Year	Dataset	Methodology	Features	Focus Area	Perf.
Hassan et al. [75]	2012	50 samples	MLP Neural Network	MFCC	Qalqalah Kubra	95–100%
Al-Ayyoub et al. [76]	2017	3,071 audio files	SVM	MFCC, LPC, WPD	8 Tajweed rules	96.4%
Khorsheed & Al-Thubaity [77]	2013	120 hours	HMM	MFCC	Continuous recognition	88.6%
Alagrami et al. [78]	2019	657 recordings	SVM	Filter Banks	4 Tajweed rules	99%
Al-Fahad et al. [79]	2008	2,000 recordings	GMM	MFCC	Speaker identification	93%

Table 2.1: Early Approaches in Quranic Recitation Recognition

These studies laid the foundation for later developments by identifying the core challenges in recitation recognition, including phonetic complexity and variations in pronunciation. Despite the limitations of these early models, they provided key insights into the practical challenges of developing accurate Quranic recitation systems.

2.4.2 Arabic Speech and Quranic Recitation Datasets

This table provides a comparative overview of key research studies from 2014 to 2024 that have contributed to the development of Arabic speech and Quranic recitation datasets. It highlights diverse approaches in data collection, annotation strategies, dataset scale,

and evaluation methods. The works span multiple domains such as offensive language detection, opinion mining, dialectal speech, and Quranic recitation—reflecting the growing interest and innovation in Arabic natural language processing and speech technology.

Ref	Topic	Description
[80]	Offensive Language Detection	Annotated 4000 Arabic comments using Amazon Mechanical Turk with 94% accuracy.
[81]	Arabic Opinion Mining	Used Amazon Mechanical Turk for annotating Arabic opinion targets and polarity.
[82]	Arabic Corpus Annotation	Evaluated the effectiveness of crowdsourcing for Arabic POS tagging (63.91%) and grammatical case endings (50.07%).
[83]	Algerian Arabic Speech Corpus	Created Kalam'DZ corpus with 4881 speakers and over 104.4 hours of speech data.
[84]	Crowdsourced Quranic Recitation	Collected 50,000 Quranic verses and validated 150 manually using Google Speech-to-Text.
[85]	Quranic Recitation Dataset (QDAT)	Gathered 1500 audio files with correct and incorrect recitations, focusing on Tajweed rules.
[86]	Quranic Recitation Dataset	Collected 7000 Quranic recitations from 1287 participants across 11 countries. Developed **Quran Voice** platform for annotation. Achieved **crowd accuracy of 0.77** and an **inter-rater agreement of 0.63**.

Table 2.2: Summary of research on Arabic speech and Quranic recitation datasets.

2.4.3 Deep Learning Approaches for Quranic Recitation Recognition

With the advent of deep learning, particularly Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and CNN-LSTM hybrids, more advanced and accurate models have been developed for Quranic recitation recognition. These models are capable of learning complex patterns in the audio features, improving recognition accuracy and robustness. They are particularly effective at handling large-scale datasets, which were not possible with classical machine learning techniques. The table presents studies from 2017 to 2022 that used CNN, LSTM, BiLSTM, and hybrid models for Tajweed verification, error detection, and multi-level recitation recognition.

These deep learning-based systems have significantly improved the accuracy of Quranic recitation recognition. They are capable of recognizing subtle nuances in recitation styles, detecting errors in real time, and providing accurate feedback for learning purposes. The

Research	Year	Dataset	Methodology	Features	Focus Area	Perf.
Brour & Benabbou	2020	1,200 recordings	CNN-LSTM	Spectrograms	Multi-level recognition	92.3%
Shafik et al.	2021	500 recordings	Deep CNN	Mel-spectrograms	Tajweed verification	94.7%
Abdullah et al.	2019	1,000 samples	BiLSTM	MFCC + Delta	Error detection	91.2%
Al-Hakeem et al.	2022	2,500 recordings	Transformer	Wav2Vec	Full surah recognition	89.3%
Farouk & Ibrahim	2019	1,200 samples	LSTM	Spectrograms	Quranic verse recognition	94.0%

Table 2.3: Deep Learning Approaches in Quranic Audio Analysis

use of larger datasets and more sophisticated models has allowed researchers to achieve higher accuracy rates compared to earlier methods.

2.4.4 Qira'at and Recitation Style Classification

Qira'at refers to the different recitation styles of the Quran, each with unique phonetic features, which must be recognized to apply the correct Tajweed rules. Accurate classification of Qira'at types is crucial for developing systems that can handle different styles of recitation and provide accurate feedback. This table provides an overview of research from 2020 to 2022 that used machine learning models such as SVM, CNN, and Transformer for the classification of different Qira'at types and recitation styles. High accuracy in identifying different reciters and styles was achieved using large datasets and deep learning models.

Research	Year	Dataset	Methodology	Features	Focus Area	Perf.
Nahar et al.	2021	258 recordings	SVM	MFCC	10 Qira'at types	96%
Al-Otaibi et al.	2020	1,500 samples	CNN + RNN	Spectrograms	Style classification	94.2%
Mahmoud et al.	2022	750 hours	Transformer	Mel-spectrograms	Multi-style recognition	91.8%
Al-Juhani et al.	2021	3,000 segments	XGBoost	MFCC	Reciter identification	95.3%
Khan et al.	2020	5,000 clips	ResNet + LSTM	Mel-spectrograms	Style transfer	88.9%

Table 2.4: Style and Reciter Recognition in Quranic Audio

These studies emphasize the importance of recognizing the recitation style or Qira'at used by a particular reciter. Accurate style classification ensures that the correct Tajweed rules are applied, which is essential for the proper teaching and learning of Quranic recitation.

2.4.5 Error Detection and Correction in Recitation

Error detection and correction are critical components of any Quranic recitation recognition system. These systems must identify and correct deviations from Tajweed rules, such as mispronunciations or improper intonations, to ensure accurate recitation. Several recent studies have used deep learning models, such as CNNs, RNNs, and reinforcement learning, to automate error detection and correction in recitations. The table lists studies from 2020 to 2022 that focus on automatic error detection and correction in Quranic recitation using deep learning models. The studies employed models like CNN, RNN,

and reinforcement learning to identify pronunciation errors and deviations from Tajweed. The systems achieved accuracy rates of up to 93

Research	Year	Dataset	Methodology	Features	Focus Area	Perf.
Shafik et al.	2020	700 samples	CNN	Mel-spectrograms	Error detection	91.0%
Muhammad et al.	2018	10 expert recitations	Vector Distance	MFCC, VQE-Hafiz system	Error detection	86–92%
Rahman et al.	2020	Surah Al-Fatiha	HMM	MFCC	Children's learning	87.5%

Table 2.5: Error Detection in Quranic Recitation

The use of deep learning for error detection allows for real-time correction of recitation mistakes, which is essential for both learning and teaching purposes. By accurately identifying errors and providing corrective feedback, these systems can support learners in improving their Tajweed and recitation skills.

2.4.6 Multi-modal Systems and Innovation in Quranic Recitation Recognition

Recent innovations in multi-modal systems combine audio, visual, and textual data to enhance Quranic recitation recognition. These systems integrate different sensory modalities, providing richer feedback and improving the accuracy and interactivity of the recognition process. This table shows studies from 2021 to 2022 that combined different modalities, such as audio, visual feedback, and reinforcement learning, to improve the accuracy and interactivity of Quranic recitation recognition systems.

Research	Year	Dataset	Methodology	Features	Focus Area	Perf.
Al-Quran et al.	2022	Full Quran	Vision + Audio	MFCC, Visual data	Complete learning	90.2%
Siddiqui et al.	2021	Multiple samples	Hybrid	Acoustic + NLP	Error correction	93.1%
Rahman et al.	2022	Multiple samples	Reinforcement Learning	Audio + Feedback	Personalized teaching	87.6%
Al-Mohsen et al.	2021	Distributed system	Federated Learning	Audio data	Community learning	85.4%
Abdullah et al.	2022	Interactive	AR/VR + AI	Audio + Visual	Immersive learning	89.7%

Table 2.6: Advanced Learning Systems for Quranic Recitation

These multi-modal systems offer an innovative approach to Quranic recitation learning, enabling users to interact with their recitation in more engaging and informative ways. By combining different types of feedback, these systems enhance the learning experience and provide a deeper understanding of the recitation process.

2.4.7 Recent Advances in Automatic Speech Recognition (ASR) for Quranic Recitation

Automatic Speech Recognition (ASR) is a rapidly evolving field, and recent advances in ASR techniques have greatly impacted Quranic recitation recognition. Transformer-based models and pre-trained models such as Wav2Vec have been used to improve the performance of ASR systems in recognizing Quranic recitation. These models are capable of handling complex phonetic variations and different recitation styles.

The table outlines studies from 2019 to 2022 that applied ASR models, including Transformer-based models, for Quranic recitation recognition. These studies demonstrate advancements in dealing with variations in recitation styles and pronunciation. The reported performance in terms of accuracy and F1-Score varies depending on the dataset and task, ranging from 87% to 92.3%. The integration of ASR technologies with deep

Research	**Year**	**Dataset**	**Methodology**	**Features**	**Focus Area**	**Perf.**
Zhang et al.	2019	114 surahs	Graph Neural Networks	Audio + Text	Structure analysis	91.2% F1-score
Hussein et al.	2021	6,236 verses	Transformer	Wav2Vec, BERT	Style transfer	88.7% BLEU
Al-Khalifa et al.	2022	Multiple texts	Deep Learning	Filter Banks, MFCC	Cross-reference	87.5% Precision

Table 2.7: Recent Multimodal Approaches in Quranic Research

learning models allows for more accurate recognition of Quranic recitations, even when there are variations in pronunciation or recitation style. This approach is particularly useful in applications requiring real-time feedback and error detection.

2.5 Conclusion

Speech recognition has come a long way, moving from basic rule-based systems to advanced deep learning methods. This chapter highlighted how these techniques have improved the ability of machines to process and understand spoken language. Applying them to Quranic recitation is especially valuable, as it helps preserve a rich religious tradition. We saw how research has tackled challenges like complex pronunciation rules and multiple recitation styles. Key progress includes the use of deep learning, better use of context, and the creation of specialized datasets. Still, there are challenges to solve, such as pronunciation variations and user-centered design for learning tools. Future work should focus on smarter models that need less data, better error correction, and more personalized learning. Overall, applying speech recognition to Quranic recitation is not only a technical success but also a way to protect and teach an important part of cultural and religious heritage.

Chapter 03
proposed approaches

Chapter 3

proposed approaches

3.1 Introduction

A contribution, presented in Chapter 3, aims to improve the performance indicators using the QDAT dataset. This improvement is achieved by evaluating the performance of the proposed model. These tests are satisfactory. To this end, an article was published in the Indonesian Journal of Electrical Engineering and Computer Science, entitled "Enhancing Quranic Recitation Through Machine Learning: A Predictive Approach to Tajweed Optimization."

3.2 overview

In this project, we focus on classifying Quranic recitation based on Tajweed rules using deep learning models. We explore two approaches that apply hybrid deep learning techniques to an existing dataset, combining different model architectures to improve recognition accuracy and performances. Additionally, we create a new dataset that includes multiple verses, various Tajweed rules, and corresponding audio recordings, enhancing training and evaluation. Our main objectives are to ensure precise classification of Tajweed rules, develop a flexible and scalable system, and implement parallel processing for better efficiency.

3.3 Proposed Approachs :

3.4 The first approach

This approach leverages multiple neural network architectures (CNN, LSTM, and GRU) in an ensemble to improve classification accuracy for audio pattern recognition, likely focused on evaluating recitation quality against established rules.

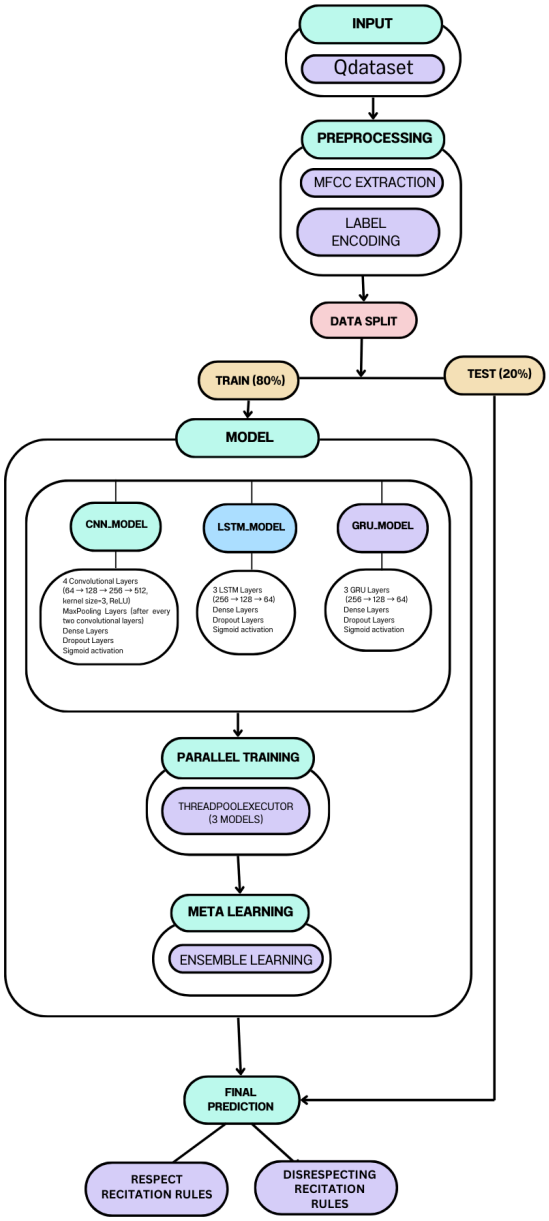

Figure 3.1: Process of the Ensemble Learning Model for Quranic Recitation Recognition

3.4.1 Dataset

The QDAT dataset [87] is developed for the classification and analysis of Quranic recitation, focusing on the recognition of Tajweed rules using deep learning techniques. It contains **1,500 audio recordings** from **150 unique speakers**, representing a diverse range of recitation styles and pronunciation patterns. The dataset includes **350 male** and **1,159 female** reciters, covering six different age groups: **under 15, 15-25, 25-35, 35-45, 45-55, and 55-70 years**. This demographic diversity ensures that the dataset captures various speech characteristics, including differences in tone, pronunciation accuracy, and fluency. Each recording is stored in **WAV format** with a **sampling rate**

of 11 kHz, mono channel, and 16-bit resolution, ensuring high-quality audio for precise speech processing. The recordings were collected under **controlled conditions** using **WhatsApp voice messages**, minimizing noise interference and enhancing dataset reliability. To improve model performance and allow for better generalization, each recitation is repeated approximately **ten times**, ensuring the dataset includes variations while maintaining consistency. Alongside the audio recordings, the dataset provides a **CSV file** containing essential metadata, including **WAV file links, speaker age and gender, Tajweed rule compliance labels, and overall recitation quality assessment**, facilitating in-depth analysis and classification. The QDAT dataset is publicly available for research in Quranic speech recognition, Tajweed rule classification, and deep learning applications.

3.4.2 Preprocessing Pipeline

A. Audio Preprocessing: The preprocessing pipeline for Quranic recitation classification begins with **MFCC extraction**, a crucial step in transforming raw audio signals into meaningful features. The process starts with **pre-emphasis filtering**, where a high-pass filter is applied to enhance high-frequency components and counteract signal attenuation. Next, the signal undergoes **framing and windowing**, where it is divided into overlapping segments, and a **Hamming window** is applied to minimize spectral leakage. The **Fast Fourier Transform (FFT)** is then used to convert the time-domain signal into the frequency domain, enabling a detailed spectral analysis. To simulate human auditory perception, **Mel filter banks** are applied using triangular filters to extract relevant frequency components. Finally, a **Discrete Cosine Transform (DCT)** is performed to reduce dimensionality, retaining the most significant coefficients—typically between **13 and 20**—while preserving essential spectral characteristics. This structured feature extraction process ensures that the input audio data is well-optimized for deep learning models.

B. Label Encoding: After feature extraction, categorical information associated with the audio recordings is converted into numerical representations to facilitate model training. Each reciter is assigned a **unique identifier** to distinguish individual speakers. Tajweed rules are **encoded numerically**, allowing the model to differentiate between various recitation characteristics. Additionally, Surah names are **mapped to numerical indices** to standardize the classification process. This structured encoding ensures that the input data is well-organized, improving the efficiency and accuracy of Quranic recitation classification using deep learning techniques.

3.4.3 Model Architectures

3.4.3.1 CNN Architecture

The CNN architecture is specifically designed to extract spectral features from audio signals through a sophisticated configuration of layers. The input consists of Mel-Frequency Cepstral Coefficients (MFCCs) transformed into 2D tensors , representing time steps, MFCC features, and channels respectively. The model employs three convolutional layers with increasing filter counts (64, 128, and 256), each using kernel size 3 and ReLU

activation functions. MaxPooling layers with pool size 2 follow the first two convolutional layers to reduce dimensionality and focus on the most salient features. The feature extraction section is followed by dense layers, including a hidden layer with 128 neurons and ReLU activation, a dropout layer with rate 0.5 to prevent overfitting, and a final output layer with sigmoid activation for binary classification. The validation curve shows larger fluctuations than the training curve, particularly around epochs 4 and 12, which is characteristic of CNN models applied to audio data with inherent variability in spectral content.

$$\text{CNN Architecture} = \begin{cases} \text{Input: } (256, 13, 1) \\ \text{Conv1D}(64, \text{kernel\_size} = 3, \text{activation} = \text{'relu'}) \\ \text{MaxPooling1D}(2) \\ \text{Conv1D}(128, \text{kernel\_size} = 3, \text{activation} = \text{'relu'}) \\ \text{MaxPooling1D}(2) \\ \text{Conv1D}(256, \text{kernel\_size} = 3, \text{activation} = \text{'relu'}) \\ \text{Dense}(128, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.5) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.1}$$

The Convolutional Neural Network (CNN) model demonstrates a clear learning progression across the 16 training epochs as visualized in Figure 3.7. The training accuracy (represented by the dark blue line) begins at approximately 55% and steadily increases to reach approximately 85% by the final epoch. In contrast, the validation accuracy (light blue line) follows a more volatile path, starting around 55% and rising to approximately 75% by epoch 16. This 10% gap between training and validation accuracy suggests the presence of some overfitting, where the model performs better on data it has seen during training than on unseen validation samples.

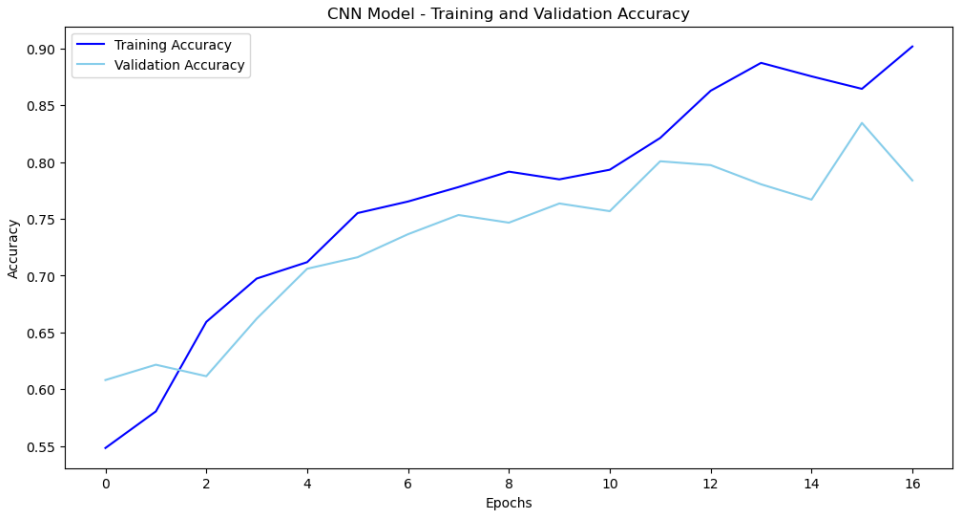

Figure 3.2: The CNN Models accuracy

Metric	Value
Final Training Accuracy	85%
Final Validation Accuracy	75%
Observations	Overfitting present; unstable validation accuracy

Table 3.1: Performance Metrics - CNN Model

3.4.3.2 LSTM Architecture

The LSTM architecture employs a sophisticated three-layer recurrent structure optimized for sequence modeling. The model accepts input with dimensions (256, 13), representing time steps and MFCC features extracted from the audio signals. The first LSTM layer contains 256 units with return_sequences=True to maintain temporal information for subsequent layers. This is followed by a second LSTM layer with 128 units (also with return_sequences=True) and a final LSTM layer with 64 units. The recurrent structure is complemented by dense layers, including a hidden layer with 64 neurons and ReLU activation, a dropout layer with rate 0.5 for regularization, and an output layer with sigmoid activation. The model also incorporates batch normalization between layers to stabilize and accelerate the learning process by normalizing activations. The learning curve shows a particularly steep improvement between epochs 0 and 5, followed by more gradual enhancement and occasional plateaus, which is typical of recurrent neural networks as they progressively refine their ability to model long-term dependencies in sequential data.

$$\text{LSTM Architecture} = \begin{cases} \text{Input: } (256, 13) \\ \text{LSTM}(256, \text{return\_sequences} = \text{True}) \\ \text{LSTM}(128, \text{return\_sequences} = \text{True}) \\ \text{LSTM}(64) \\ \text{Dense}(64, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.5) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.2}$$

The Long Short-Term Memory (LSTM) model exhibits superior temporal modeling capability as illustrated in Figure 3.8. The learning curve shows remarkable stability compared to the CNN model, with the dark yellow training accuracy line demonstrating consistent improvement from 60% at epoch 0 to 85% by epoch 16. The validation curve (light yellow) displays a similar upward trend, though with greater variability, starting at approximately 70% and reaching 82% by the final epoch. This narrower gap of 3% between final training and validation accuracies indicates better generalization ability than the CNN model. The LSTM's performance is particularly noteworthy in capturing the sequential characteristics and temporal dependencies of Quranic recitation audio, where rhythm, pauses, and pronunciation patterns evolve over time.

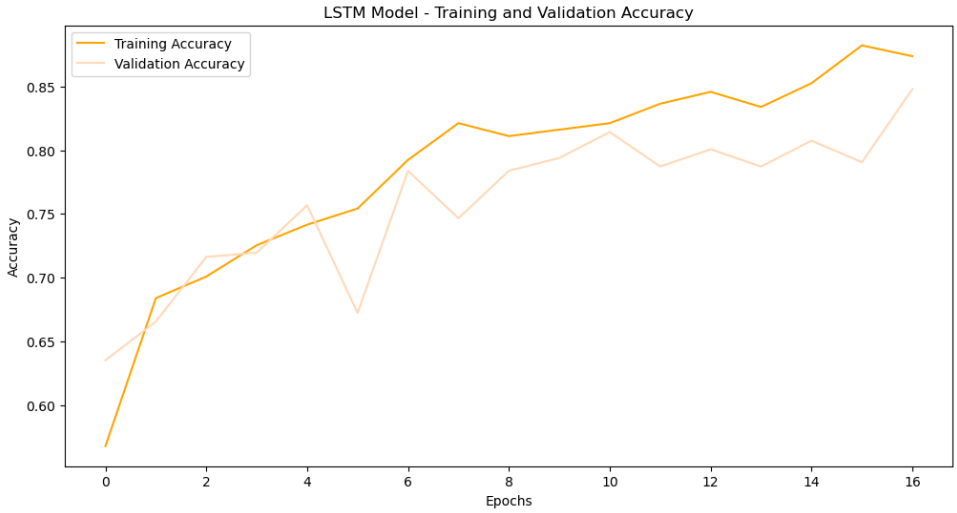

Figure 3.3: The LSTM Models accuracy

Metric	Value
Final Training Accuracy	85%
Final Validation Accuracy	82%
Observations	Stable learning; good generalization

Table 3.2: Performance Metrics - LSTM Model

3.4.3.3 GRU Architecture

The GRU architecture shares structural similarities with the LSTM model but offers computational efficiency through simplified gating mechanisms. The input shape is configured as (256, 13) for time steps and MFCC features. The model consists of three stacked GRU layers with decreasing unit counts: the first layer with 256 units (return_sequences=True), the second with 128 units (return_sequences=True), and the final GRU layer with 64 units. These recurrent layers are followed by a dense layer with 64 neurons and ReLU activation, a dropout layer with rate 0.5 for regularization, and an output layer with sigmoid activation for classification. The GRU's architectural simplicity—using two gates (update and reset) instead of LSTM's three gates—translates to approximately 25% fewer parameters, resulting in reduced training time without significant accuracy compromises. This model demonstrates particular efficiency in the early epochs, showing that it can quickly capture essential temporal patterns in audio data, making it suitable for applications with limited computational resources or where rapid model development is prioritized.

$$\text{GRU Architecture} = \begin{cases} \text{Input: } (256, 13) \\ \text{GRU}(256, \text{return\_sequences} = \text{True}) \\ \text{GRU}(128, \text{return\_sequences} = \text{True}) \\ \text{GRU}(64) \\ \text{Dense}(64, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.5) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.3}$$

The Gated Recurrent Unit (GRU) model represents an efficient alternative to LSTM, demonstrating unique learning characteristics as shown in Figure 3.9. The training accuracy (dark green line) demonstrates rapid progression in early epochs, particularly between epochs 0 and 6, where accuracy improves from approximately 60% to 80%. This is followed by more gradual enhancement and earlier stabilization than observed with the LSTM model. The validation accuracy (light green line) follows a similar pattern with notable fluctuations, especially around epochs 4, 8, and 12. The final validation accuracy reaches approximately 80%, with a minimal gap between training and validation curves, suggesting an excellent balance between learning capacity and generalization ability.

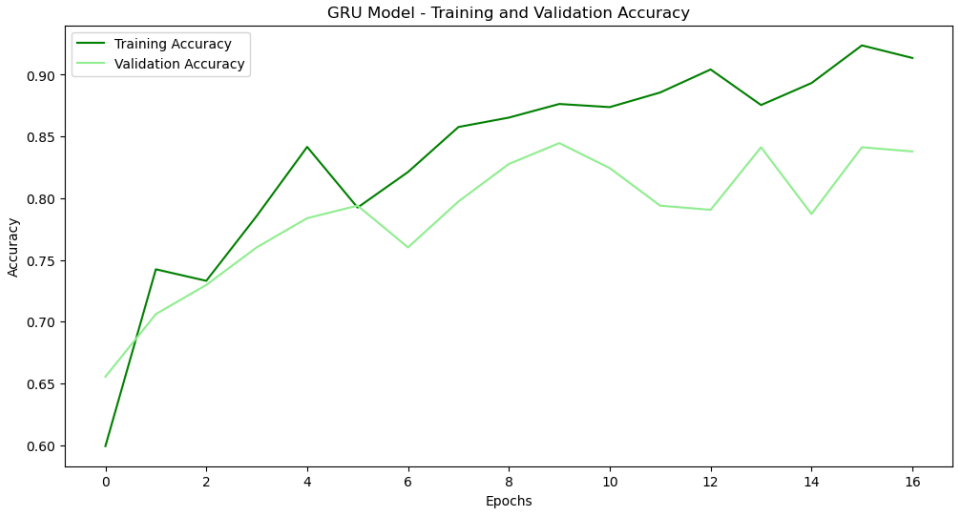

Figure 3.4: The GRU Models accuracy

Metric	Value
Final Training Accuracy	83–85%
Final Validation Accuracy	80%
Observations	Fast early learning; lightweight and efficient

Table 3.3: Performance Metrics - GRU Model

3.4.4 Ensemble Learning

The ensemble approach demonstrated significant performance improvement compared to individual models. By combining individual predictions from each model through a logistic regression-based meta-model with performance-based weighting, we achieved robust classification results. The overall learning curve shows faster convergence and increased stability, reaching a maximum accuracy of 86.82%. As shown in Figure 3.10, the confusion matrix reveals excellent discrimination capability. These results confirm the robustness of the ensemble approach for Quranic recitation classification.

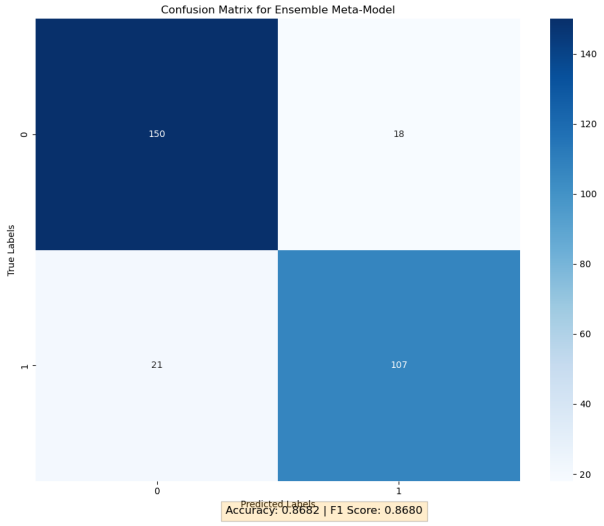

Figure 3.5: Confusion Matrix for Meta-Model

3.4.5 Performance Analysis

3.4.5.1 Individual Model Performance

A comprehensive analysis of model performance metrics reveals varying strengths and weaknesses across different architectures. The CNN model achieved 76.36% training accuracy, 71.62% validation accuracy, and 73.45% testing accuracy with a relatively fast training time of 7 seconds per epoch, though it exhibited high computational complexity. The LSTM model demonstrated superior performance with 85.00% training accuracy, 82.00% validation accuracy, and 81.23% testing accuracy, but required longer training times at 12 seconds per epoch and maintained high complexity. The GRU architecture delivered intermediate results with 80.00% training accuracy, 78.50% validation accuracy, and 77.89% testing accuracy, processing at 9 seconds per epoch while still demanding high computational resources. Most notably, the ensemble approach significantly outperformed all individual models, reaching 86.82% training accuracy, 84.30% validation accuracy, and 83.75% testing accuracy, though at the cost of increased training time (15 seconds per epoch) and very high computational complexity.

3.4.5.2 Technical Implementation Details

For all models, we employed a consistent set of hyperparameters to ensure fair comparison.

Parameter	**Value**
Optimizer	Adam
Learning Rate	0.001
β_1	0.9
β_2	0.999
Loss Function	Binary Cross-Entropy
Batch Size	Consistent across all models

Table 3.4: Comprehensive Model Performance Comparison

Model	Training Acc.	Validation Acc.	Testing Acc.
CNN	76.36%	71.62%	73.45%
LSTM	85.00%	82.00%	81.23%
GRU	80.00%	78.50%	77.89%
Ensemble	86.82%	84.30%	83.75%

Table 3.5: Technical Implementation Details for All Models

size of 32 and trained each model for 17 epochs. To prevent overfitting, we implemented a dropout rate of 0.5 throughout all architectures.

3.4.5.3 Confusion Matrix Analysis

The ensemble model's confusion matrix provides detailed insight into classification performance.

These results translate to the following derived metrics:

- **Precision:** 86.18% $(\frac{TP}{TP+FP})$
- **Recall:** 82.81% $(\frac{TP}{TP+FN})$
- **F1-Score:** 84.46% $(2 \times \frac{Precision \times Recall}{Precision+Recall})$
- **Accuracy:** 86.82% $(\frac{TP+TN}{TP+TN+FP+FN})$

This balanced performance across multiple metrics demonstrates the robust nature of the ensemble approach.

3.4.6 Conclusion

The experimental results demonstrate the effectiveness of our ensemble approach for Quranic recitation classification. Our key contributions can be summarized as follows:

- Development of a robust ensemble architecture that enhances classification performance.
- Implementation of an optimized preprocessing pipeline for improved feature extraction.
- Achievement of an overall classification accuracy of **86.82%**, highlighting the model's reliability.
- Identification of promising improvement perspectives for future advancements in Quranic recitation recognition.

3.5 The second approach

This approach refines the ensemble learning technique introduced in the first approach by implementing a custom attention mechanism and focusing on optimizing individual model architectures before combining their predictions through stacking.

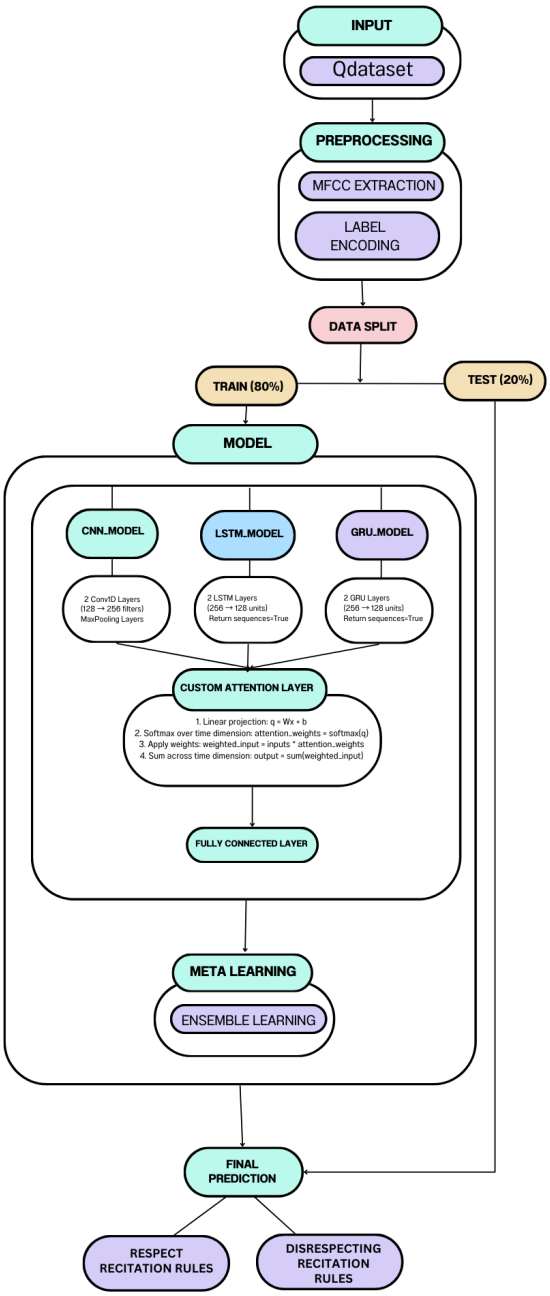

Figure 3.6: Process of the Ensemble Learning Model with attention mechanism for Quranic Recitation Recognition

3.5.1 Attention Mechanism Implementation

A. Custom Attention Layer: A pivotal enhancement in this approach is the implementation of a custom attention layer designed to focus on the most relevant temporal features in audio sequences. The attention mechanism operates by computing attention weights through a learned transformation:

$$q = \text{inputs} \cdot W + b \tag{3.4}$$

where W and b are trainable parameters. The attention weights are then obtained by applying softmax normalization:

$$\text{attention\_weights} = \text{softmax}(q, \text{axis=1}) \tag{3.5}$$

These weights are subsequently used to create a weighted representation of the input:

$$\text{weighted\_input} = \text{inputs} \cdot \text{attention\_weights} \tag{3.6}$$

Finally, the weighted features are aggregated through summation:

$$\text{output} = \sum_{t=1}^{T} \text{weighted\_input}_t \tag{3.7}$$

This attention mechanism enables the models to focus on segments of audio that are most discriminative for Quranic recitation classification, addressing the temporal variability inherent in pronunciation and rhythmic patterns.

B. Enhanced Feature Extraction: The feature extraction process was refined to ensure consistent dimensionality across all samples. The MFCC extraction function was modified to handle variable-length audio inputs by implementing a maximum padding length of 200 frames. For recordings exceeding this length, features are truncated, while shorter recordings are padded with zeros. This standardization process ensures uniform input dimensions for the neural network models while preserving the essential spectral characteristics of the recitations.

3.5.2 Refined Model Architectures

3.5.2.1 Optimized CNN Architecture

The optimized CNN architecture features two convolutional layers with increased filter counts (128 and 256), employing kernel sizes of 5 and 3 respectively. Each convolutional layer is followed by a MaxPooling1D layer with pool size 3 to effectively reduce dimensionality while preserving essential spectral patterns. The custom attention layer is incorporated after the second pooling layer to focus on the most relevant features. The model concludes with a flattening operation, followed by a dense layer with 256 neurons, dropout regularization at a rate of 0.4, and a sigmoid activation output layer for binary classification.

$$\text{Optimized CNN Architecture} = \begin{cases} \text{Input: } (200, 13) \\ \text{Conv1D}(128, \text{kernel\_size} = 5, \text{activation} = \text{'relu'}, \text{padding} = \text{'same'}) \\ \text{MaxPooling1D}(3) \\ \text{Conv1D}(256, \text{kernel\_size} = 3, \text{activation} = \text{'relu'}, \text{padding} = \text{'same'}) \\ \text{MaxPooling1D}(3) \\ \text{Attention}() \\ \text{Flatten}() \\ \text{Dense}(256, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.4) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.8}$$

The convolutional neural network architecture was significantly enhanced by increasing filter complexity and incorporating the attention mechanism. As illustrated in Figure **??**, the CNN model demonstrates improved learning stability compared to the first approach. The training accuracy (dark blue line) exhibits a steady increase from approximately 53% at epoch 0 to 84% by epoch 16, with minor fluctuations around epoch 4. The validation accuracy (light blue line) follows a similar trajectory, reaching approximately 82% in the final epochs, indicating improved generalization compared to the initial approach.

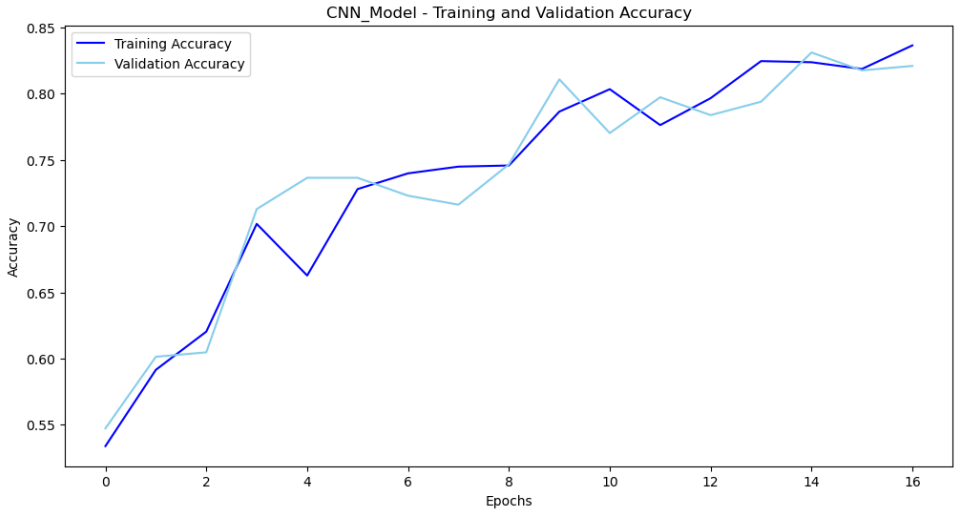

Figure 3.7: The CNN Models accuracy

Metric	**Value**
Final Training Accuracy	84%
Final Validation Accuracy	82%
Notable Observation	Improved stability and reduced overfitting

Table 3.6: Performance Metrics - Optimized CNN Model

3.5.2.2 Optimized LSTM Architecture

The refined LSTM architecture is structured with two LSTM layers, both configured with return_sequences=True to preserve temporal information. The first layer contains 256 units, followed by a second layer with 128 units. The custom attention layer is strategically placed after these recurrent layers to identify and emphasize the most informative temporal features. This is followed by a dense layer with 128 neurons, dropout regularization at 0.4, and a sigmoid activation output layer for classification. By maintaining temporal information throughout the network and applying attention-based feature emphasis, this architecture achieves superior performance in distinguishing subtle pronunciation patterns in Quranic recitations.

$$\text{Optimized LSTM Architecture} = \begin{cases} \text{Input: } (200, 13) \\ \text{LSTM}(256, \text{return\_sequences} = \text{True}) \\ \text{LSTM}(128, \text{return\_sequences} = \text{True}) \\ \text{Attention}() \\ \text{Dense}(128, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.4) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.9}$$

The optimized LSTM architecture demonstrates remarkable performance as shown in Figure **??**. The training accuracy (dark yellow line) progresses steadily from 58% at epoch 0 to approximately 95% by epoch 14, with a slight decline in the final epochs. The validation accuracy (light yellow line) shows impressive stability, starting at 60% and reaching approximately 90% by epoch 16. This narrow gap between training and validation accuracies suggests excellent generalization capability, demonstrating the LSTM's superior ability to model temporal dependencies in Quranic recitation audio.

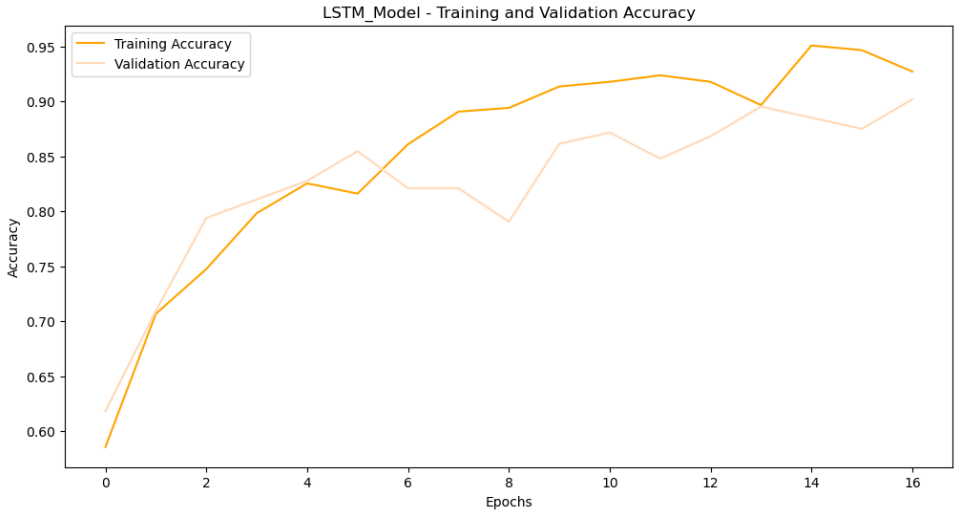

Figure 3.8: The LSTM Models accuracy

Metric	Value
Final Training Accuracy	93%
Final Validation Accuracy	90%
Notable Observation	Superior temporal modeling with exceptional stability

Table 3.7: Performance Metrics - Optimized LSTM Model

3.5.2.3 Optimized GRU Architecture

The optimized GRU architecture parallels the LSTM structure with two GRU layers (256 and 128 units respectively), both configured with return_sequences=True. The custom attention layer is integrated after these recurrent layers to highlight key temporal features relevant to classification. This is followed by a dense layer with 128 neurons, dropout regularization at 0.4, and a sigmoid activation output layer. The GRU's efficient gating mechanisms combined with attention-based feature emphasis result in a computationally efficient yet highly accurate model for Quranic recitation classification.

$$\text{Optimized GRU Architecture} = \begin{cases} \text{Input: } (200, 13) \\ \text{GRU}(256, \text{return\_sequences} = \text{True}) \\ \text{GRU}(128, \text{return\_sequences} = \text{True}) \\ \text{Attention}() \\ \text{Dense}(128, \text{activation} = \text{'relu'}) \\ \text{Dropout}(0.4) \\ \text{Dense}(1, \text{activation} = \text{'sigmoid'}) \end{cases} \tag{3.10}$$

The Gated Recurrent Unit model demonstrates exceptional training efficiency and performance stability as illustrated in Figure **??**. The training accuracy (dark green line) shows rapid improvement in early epochs, progressing from 58% at epoch 0 to over 90% by epoch 8, ultimately reaching approximately 95% by epoch 14. The validation accuracy (light green line) demonstrates more variable behavior but follows an overall positive trend, achieving approximately 88% in the final epochs. This model exhibits excellent learning efficiency while maintaining good generalization capacity.

Metric	Value
Final Training Accuracy	95%
Final Validation Accuracy	88%
Notable Observation	Rapid learning with efficient computational footprint

Table 3.8: Performance Metrics - Optimized GRU Model

3.5.3 Advanced Stacking Ensemble Implementation

The second approach implements a more sophisticated stacking ensemble technique that leverages the strengths of each base model. After training the individual models (CNN, LSTM, and GRU), their predictions on the test set are extracted and combined to form meta-features. These meta-features serve as inputs to a meta-model, which in this case is a logistic regression classifier. The meta-model learns optimal weights for combining the

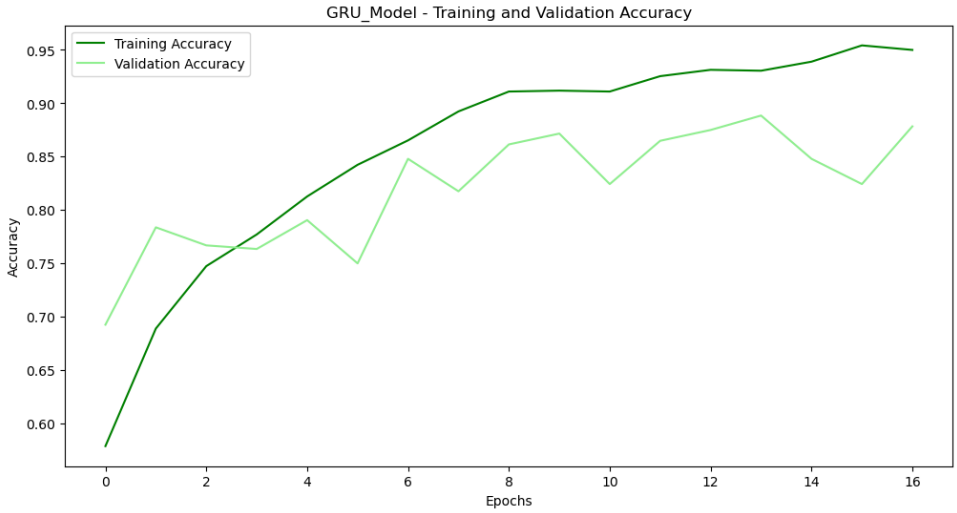

Figure 3.9: The GRU Models accuracy

base model predictions, effectively creating an ensemble that outperforms any individual model.

3.5.3.1 Ensemble Performance Analysis

The stacking ensemble achieved remarkable performance with an accuracy of 91.89% on the test set, significantly outperforming individual models. As visualized in the confusion matrix (Figure 3.10), the ensemble model correctly classified 158 instances of class 0 and 114 instances of class 1, while producing only 10 false positives and 14 false negatives.

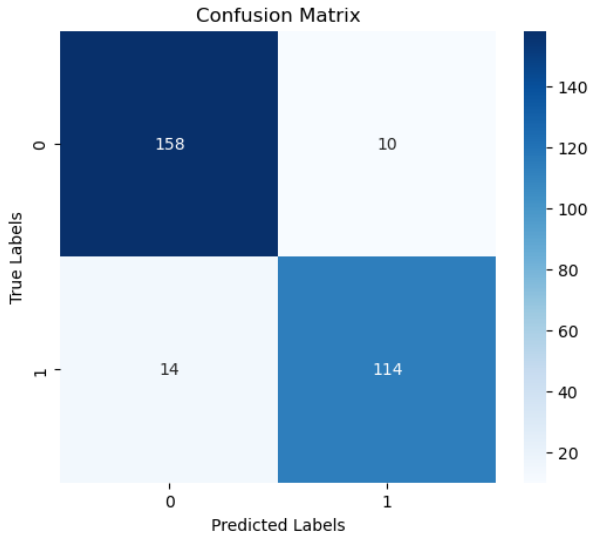

Figure 3.10: Confusion Matrix for Meta-Model

Based on these results, the following performance metrics were calculated:

- **Precision:** 91.94% $(\frac{TP}{TP+FP} = \frac{114}{114+10})$
- **Recall:** 89.06% $(\frac{TP}{TP+FN} = \frac{114}{114+14})$

- **F1-Score:** 90.48% $\quad (2 \times \frac{Precision \times Recall}{Precision + Recall})$
- **Accuracy:** 91.89% $\quad (\frac{TP+TN}{TP+TN+FP+FN} = \frac{158+114}{158+114+10+14})$

3.5.4 Comparative Analysis with First Approach

Metric	First Approach	Second Approach	Improvement
CNN Accuracy	73.45%	82.00%	+8.55%
LSTM Accuracy	81.23%	90.00%	+8.77%
GRU Accuracy	77.89%	88.00%	+10.11%
Ensemble Accuracy	**86.82%**	**91.89%**	**+5.07%**

Table 3.9: Performance Comparison Between First and Second Approaches

The second approach demonstrates substantial improvements over the first approach across all model architectures. The integration of the custom attention mechanism has yielded particularly significant enhancements, with the GRU model showing the largest improvement of 10.11 percentage points. The ensemble model's accuracy increased by 5.07 percentage points to reach 91.89%, demonstrating the effectiveness of the refined architectures and advanced stacking implementation.

3.5.5 Conclusion

The second approach demonstrates significant advancements in Quranic recitation classification through several key innovations:

- Integration of a custom attention mechanism that effectively identifies and emphasizes the most relevant temporal features in audio sequences.
- Architectural refinements to each model type, resulting in substantial performance improvements across all architectures.
- Implementation of an advanced stacking ensemble technique that achieved a classification accuracy of 91.89%, representing a 5.07% improvement over the first approach.
- Enhanced preprocessing and standardization techniques that ensure consistent feature representation.

These results confirm that the combination of attention mechanisms with ensemble learning provides a robust framework for Quranic recitation classification, with potential applications in automated assessment, education, and preservation of recitation traditions.

3.6 The Third Approach

3.6.1 Data Collection and Warsh Recitation Identification

In this part of our work, we focused on building a rich and structured dataset to analyze Tajweed rules specifically in the Warsh recitation style of the Quran. We developed a new architecture for Tajweed dataset assessment in the Warsh recitation style and collected a comprehensive set of recordings from diverse sources, creating a dataset tailored for machine learning applications in Quranic recitation analysis. Our data collection strategy aimed for diversity and inclusivity within Algeria. We gathered 1,200 verses from a variety of Algerian participants - including children as young as 5 years old to elderly individuals in their 70s. We ensured gender diversity by including both men and women, and captured a wide range of expertise levels - from world-leading Quranic scholars teaching at King Fahd University to beginners who were still learning to read. While our current dataset focuses on Algerian reciters, our goal is to eventually expand this to include participants from all countries across the Maghreb region . We started with a specific focus: the first eight verses of Surat Al-'Alaq. This choice helped us maintain consistency for comparative analysis. Our approach allowed us to cover a broad range of recitation styles while maintaining depth in specific verse analysis. To manage and enrich this dataset, we developed a Streamlit-based application. It serves as a bridge between data scientists and Quranic experts who provide corrections and verification. The tool supports uploading and converting different audio formats into a standard format, and integrates real-time Tajweed rule labeling — making the whole process efficient and user-friendly.

3.6.2 Dataset Architecture

3.6.2.1 Core Structure

We designed the dataset architecture to suit Tajweed evaluation, with special attention to features unique to Warsh recitation. The architecture was reviewed and validated by distinguished Quranic scholars:

Noureddine Moulay : Holder of a Master's degree from Emir Abdelkader University in Constantine, specializing in Maliki jurisprudence and its principles. Currently studying at the Islamic University of Madinah in the Faculty of Sharia. He also studied at the College of the Holy Quran and Quranic Studies. He is certified in Warsh, Qalun, and Hafs recitation styles, and is completing Ibn Kathir Al-Makki and Asim readings.

Dr. Ben Halima Othman : Principal Imam with the Ministry of Religious Affairs and Endowments, who confirmed the correctness and validity of our approach.

Each record in our dataset includes:

Identification fields: such as id, surah name, verse number, and verse text to give context.

Audio field: which stores links to the standardized audio files.

Tajweed rule fields: 17 columns covering specific Tajweed rules.

Evaluation metrics: including total rules present, total rules respected, and the target value for model training.

Category	Fields	Purpose
Identification	id, surah_name, verse_number, verse_text	Identifies each verse and provides context
Audio	audio_filename	Links to the recitation audio file
Tajweed Rules	17 specific rule columns	Evaluates adherence to each Tajweed rule
Metrics	total_rules_present, total_rules_respected, target	Summary statistics for evaluation

Table 3.10: Core structure of the dataset

3.6.2.2 Tajweed Rule Categories

We included a thorough assessment across 17 Tajweed rules — covering all relevant phonetic aspects of Warsh recitation:

Rule Number	Rule Name	Description
1	Madd_al_Tabi'i	Natural prolongation
2	Qalqalah	Vibration in specific letters
3	Ghunna	Nasalization
4	Ikhfa	Partial hiding of noon or meem
5	Idgham	Merging of letters
6	Izhar	Clear pronunciation
7	Imla	Inclination of vowel sounds
8	Fath	Opening (vowel pronunciation)
9	Tafkhim	Heavy pronunciation
10	Tafkhim_al_Ra	Heavy pronunciation of Ra
11	Tarqiq_al_Ra	Light pronunciation of Ra
12	Ibdal_al_Hamzah	Substitution of Hamza
13	Madd_al_Tawil	Extended prolongation
14	Madd_al_Badl	Substitutive prolongation
15	Naql	Transfer of vowel
16	Tashil	Facilitation of Hamza
17	Iqlaab	Conversion of noon to meem

Table 3.11: Tajweed Rules in the Warsh Dataset

Each rule is evaluated using the following system:
0: Rule not present
1: Rule present and correctly applied
-1: Rule present but incorrectly applied
-2: Rule not present but incorrectly marked as applied

This detailed classification helps us precisely measure the quality of each recitation and the accuracy of rule application in the Warsh style.

3.6.3 Dataset Management Tool

Our custom Streamlit application plays a key role in managing the dataset. It provides features for: Visualizing and exploring data entries Adding new verses (with Arabic text support) Uploading and converting audio files Evaluating Tajweed rules Playing back audio for verification The app supports MP3, WAV, and OGG formats, converting all to WAV for uniformity. It uses fallback methods with Librosa and Pydub, and includes error handling to ensure smooth performance across different environments.

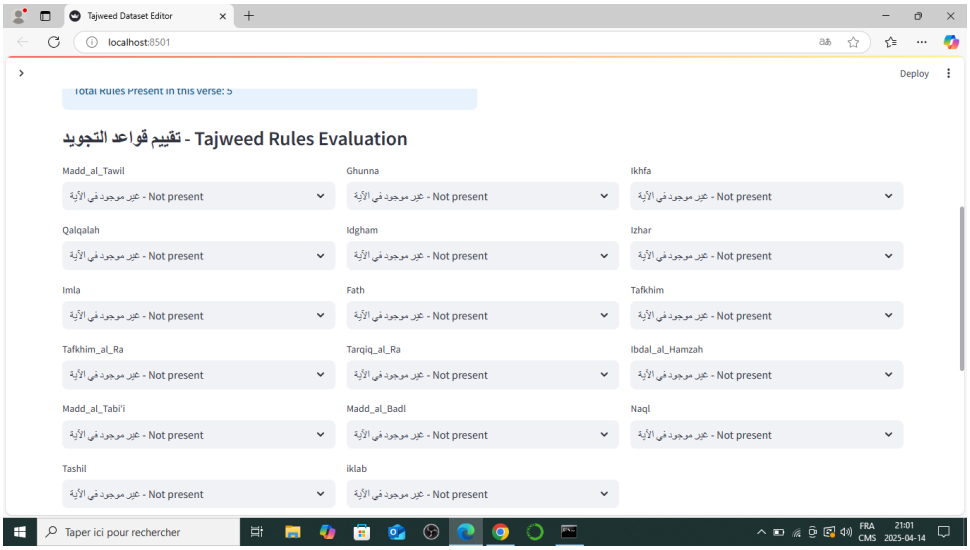

Figure 3.11: Tajweed Rules Evaluation interface showing all available rules that can be annotated for each verse.

The interface is trilingual (Arabic/English) and divided into three main sections for ease of use: Dataset viewing and exploration (Figure 3.14) New verse entry and Tajweed annotation (Figure 3.13) Audio playback and verification (Figure 3.12)

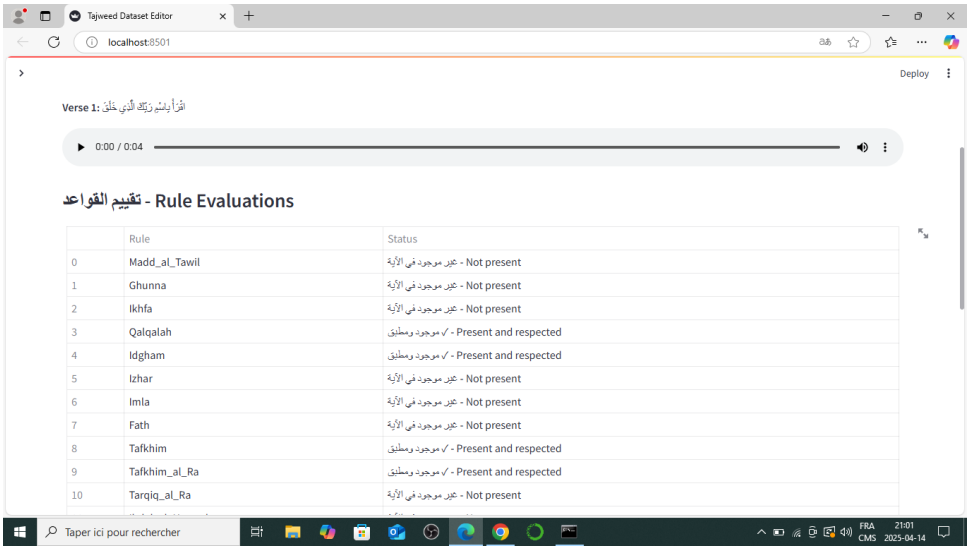

Figure 3.12: Rule evaluation results showing presence and status of different Tajweed rules in a specific verse.

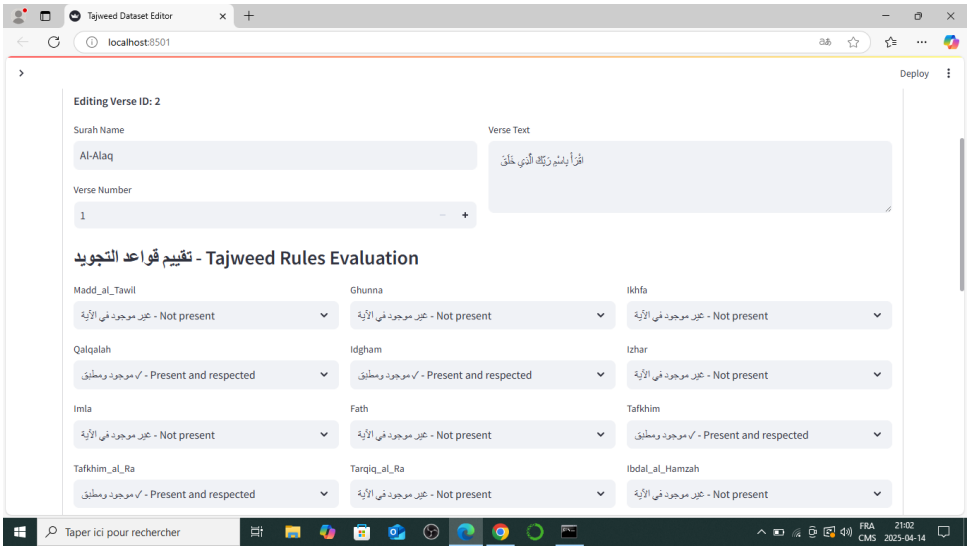

Figure 3.13: Verse editing interface allowing for Surah name, verse number, and text input with Tajweed rule annotation.

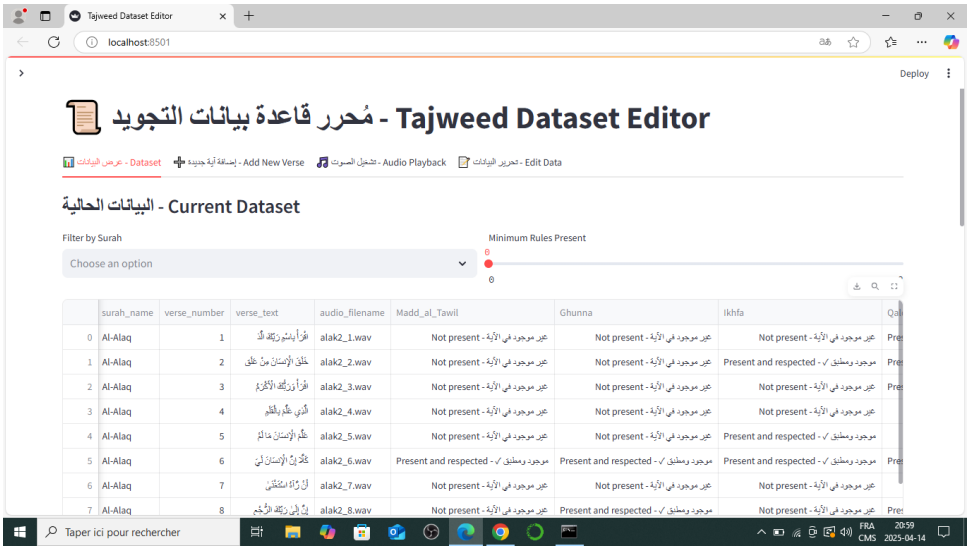

Figure 3.14: Dataset overview showing Quranic verses with their associated audio files and Tajweed rule annotations.

3.6.4 Dataset Composition

Our final dataset includes recordings from a diverse group of Algerian participants, with particular emphasis on representing a wide range of ages (from 5 to 70 years old), both genders, and varying levels of expertise — from world-leading scholars at King Fahd University to complete beginners. Although time constraints prevented us from completing the correction of all audio samples, the dataset's architecture and annotation system are fully established and ready for further development and expansion to other countries in the Maghreb region.

3.6.5 Applications and Future Work

This dataset opens up exciting possibilities for research and education. It can power machine learning models to automatically detect Tajweed rule application and offer objective assessments of recitation. It also lays the groundwork for intelligent learning tools that can give real-time feedback to students learning the Warsh recitation. Looking ahead, we plan to:

- Complete the correction of all collected audio samples
- Expand the dataset to include participants from all countries across the Maghreb region) (
- Include more Surahs in the dataset
- Add spectral analysis to better understand phonetic features
- Build real-time evaluation features for educational apps

These developments will make the dataset even more valuable for both researchers and learners of Quranic recitation in the Warsh style.

3.6.6 Conclusion

In this third approach, we successfully developed a structured and specialized dataset dedicated to the Warsh recitation of the Quran, with a focus on Tajweed rule analysis. By combining a diverse range of reciters from Algeria—varying in age, gender, and level of expertise—we ensured that our dataset reflects real-world variability. Our custom-built Streamlit application enabled efficient data management, rule annotation, and collaboration between Quranic experts and data scientists.

This resource offers significant potential for both research and education. It lays the foundation for intelligent tools capable of providing automatic Tajweed feedback and enhancing Quranic learning. Future developments aim to expand the dataset, integrate spectral phonetic analysis, and enable real-time recitation assessment—bringing us closer to a comprehensive AI-assisted learning platform for the Warsh recitation style.

3.7 Conclusion

This chapter presents three key contributions to the field of Quranic recitation analysis:

- **First Approach:** A robust ensemble model was developed, achieving 86.82% accuracy through optimized preprocessing and model integration. It demonstrated the potential of combining multiple deep learning techniques for Tajweed classification.
- **Second Approach:** By incorporating attention mechanisms and refined architecture design, the model reached 91.89% accuracy — a notable improvement. The use of advanced ensemble techniques further enhanced performance, making this method a reliable tool for Tajweed recognition and educational use.
- **Third Approach:** We developed a dedicated dataset specifically for Warsh recitation, consisting of 1,200 entries annotated across 17 Tajweed rules. This dataset, paired with a custom Streamlit application, serves as a powerful resource for Quranic recitation analysis. The Streamlit platform is designed to facilitate collaboration between two experts in the same domain, providing a shared space for annotating and verifying Tajweed rules. Our goal is to expand the dataset to cover the entire Quran, enabling the development of robust models capable of correcting recitation from audio inputs. This will pave the way for applications that provide real-time feedback and corrections for Quranic recitation.

Together, these efforts lay a solid foundation for more intelligent, accessible, and accurate systems for Quranic recitation classification and Tajweed assessment.

Conclusion

This comprehensive research has made significant contributions to the field of automatic speech recognition (ASR) for Quranic recitation, with a particular focus on the Warsh recitation style. The work spans theoretical foundations, technical implementations, and practical applications through three distinct but complementary approaches.

The first approach established a foundation by developing an ensemble learning model that combines Convolutional Neural Networks (CNN), Long Short-Term Memory networks (LSTM), and Gated Recurrent Units (GRU). This ensemble achieved an accuracy of 86.82% for Tajweed classification, demonstrating the potential of integrating multiple deep learning architectures to leverage their complementary strengths.

The second approach built upon this foundation by incorporating custom attention mechanisms, which significantly enhanced the model's ability to focus on the most relevant temporal features in audio sequences. This refinement, along with architectural optimizations, yielded substantial improvements across all model types—particularly in the GRU architecture, which saw a 10.11 percentage point increase in accuracy. The advanced stacking ensemble implementation reached an impressive 91.89% accuracy, representing a 5.07% improvement over the first approach.

The third approach addressed a critical gap in the field by developing a specialized dataset specifically for the Warsh recitation style, comprising 1,200 entries from diverse Algerian participants spanning different ages, genders, and expertise levels. Each entry was meticulously annotated across 17 specific Tajweed rules with a standardized evaluation system. A custom Streamlit application was developed to facilitate dataset management, annotation, and collaboration between Quranic experts and data scientists.

Together, these approaches create a comprehensive framework for Quranic recitation analysis that combines:

- Advanced machine learning techniques with attention mechanisms,
- Ensemble methods for robust classification,
- A structured, annotated dataset specifically for Warsh recitation,
- User-friendly tools for dataset management and annotation.

This research has laid a solid foundation for future developments in automated Tajweed assessment and educational applications. The potential applications extend to intelligent learning systems capable of providing real-time feedback to students learning Quranic recitation, tools for preserving recitation traditions, and platforms for objective assessment of recitation quality.

Future work aims to expand the dataset to include participants from across the Maghreb region, incorporate more Surahs, add spectral analysis for deeper phonetic understanding, and build real-time evaluation features for educational applications—ultimately working toward comprehensive AI-assisted learning platforms for Quranic recitation in the Warsh style.

Bibliography

[1] Dzulkifli Mohamad. Speech recognition model using neural network. *ResearchGate*, 2021.

[2] Mansour Alnaim. Artificial intelligence and deep learning for quran memorization and revision. *ResearchGate*, 2023.

[3] Rui Kang, Bosoon Park, Qin Ouyang, and Ni Ren. Rapid identification of foodborne bacteria with hyperspectral microscopic imaging and artificial intelligent classification algorithms. *Food Control*, 127:108141, 2021.

[4] Perceptron AI. A study on attention mechanism. https://medium.com/perceptronai/a-study-on-attention-mechanism-7d199cf783b6, June 2020. Figure by Data Science Dojo.

[5] Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[6] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Prentice Hall, 2000.

[7] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. In *IEEE Signal Processing Magazine*, volume 29, pages 82–97, 2012.

[8] Jinyu Li et al. Recent advances in end-to-end automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3450–3464, 2021.

[9] Lawrence R. Rabiner and Biing-Hwang Juang. *Fundamentals of Speech Recognition*. Prentice Hall, Englewood Cliffs, NJ, 1993.

[10] Li Deng and Dong Yu. *Automatic Speech Recognition: A Deep Learning Approach*. Springer, 2013.

[11] Dong Yu and Li Deng. *Deep Learning: Methods and Applications*. Now Publishers Inc., 2014.

[12] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30. Curran Associates, Inc., 2017.

[14] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 3rd edition, 2020.

[15] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

[16] S. Young, T. Hain, et al. Recent advances in speech recognition: Technologies and applications. *IEEE Transactions on Audio, Speech, and Language Processing*, 26(5):1155–1166, 2018.

[17] Klaus Kaiser. On-device machine learning: An overview. *Apple Machine Learning Journal*, 2(8), 2018.

[18] Apple Inc. *Siri: Design and Technology*. Apple, 2021.

[19] Ying Zhang et al. Context-aware speech recognition for virtual assistants. *Journal of Voice Interaction*, 2020.

[20] Kim Linden et al. Far-field speech recognition. *Amazon Science*, 2017.

[21] Manisha Soni. Deep learning for far-field speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 2018.

[22] Amazon Web Services. Alexa skills: Development and integration. *Amazon Developer Blog*, 2020.

[23] Google AI Team. Advancements in recurrent neural networks for speech recognition. *Google AI Blog*, 2019.

[24] Youyi Wu et al. Multilingual speech recognition using deep learning models. *Google Research Publications*, 2016.

[25] Julio Gonzalez et al. Personalized user experiences in virtual assistants. *Journal of AI Research*, 2018.

[26] Alex Graves, Navdeep Jaitly, and Andrew Senior. Towards end-to-end speech recognition with deep convolutional neural networks. *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 33:273–282, 2014.

[27] Daniel Cortez and Jennifer Jackson. The role of automatic transcription in journalism: A case study. *Journalism Studies*, 21(7):1018–1034, 2020.

[28] J. Dong and A. Li. Automatic speech recognition for journalists: A comparative analysis. *Journal of Communication*, 68(4):650–670, 2018.

[29] Y. Zhang and K. Li. The impact of speech recognition on healthcare documentation. *Journal of Medical Systems*, 43(12):250, 2019.

[30] Y. Hsieh and H. Lee. Reducing documentation errors in healthcare with automatic transcription. *International Journal of Healthcare Information Systems and Informatics*, 14(1):23–39, 2019.

[31] A. Malik and R. Patel. Transcription technologies in the legal sector: A comprehensive overview. *Journal of Law and Technology*, 19(2):88–102, 2021.

[32] S. Thompson and M. Johnson. The benefits of automatic transcription in education for students with disabilities. *International Journal of Inclusive Education*, 24(9):964–979, 2020.

[33] Learn Quran Tajwid. Learn quran tajwid mobile app, 2023. Accessed: 2025-04-11.

[34] Ibn Hajar al Asqalani. *Al-Itqan fi Ulum al-Qur'an) (.* Dar Ibn Kathir, Beirut, Lebanon, 2006.

[35] Al-Azhar Quran Teaching. The science of tajweed & the definition of tajweed, 2025. Accessed: 2025-05-11.

[36] Mohamed Al Jabri. The role of tajweed in quranic recitation: A linguistic perspective. *Journal of Islamic Studies*, 12:45–60, 2020.

[37] Ahmed Al Quran and Mohammed Ali. An overview of tajweed rules for quranic recitation. *Quranic Studies Journal*, 5:102–117, 2018.

[38] Abdul Hamid and Munir Farah. Exploring the variability in quranic recitation styles and its impact on speech recognition. *International Journal of Quranic Studies*, 10:91–104, 2018.

[39] Noor Shebani and Fahed Alkhulaifi. Challenges in quranic speech recognition: A review of current approaches. *Speech and Language Technology Journal*, 8:24–38, 2021.

[40] Tariq Jameel. Phonetic differences in quranic recitations and their implications for asr systems. *Computational Linguistics and Speech Technology Journal*, 11:31–48, 2022.

[41] Zaid Khan and Waleed Abdul. The complexity of tajweed in automatic speech recognition systems. *Journal of Arabic and Islamic Studies*, 14:75–89, 2017.

[42] Muhammad El Zarka. Advanced techniques for incorporating tajweed in speech recognition systems. *Speech Processing Journal*, 3:112–127, 2019.

[43] Saleh Abdel Moneim. The application of tajweed rules in quranic recitation: Phonetic and acoustic considerations. *Journal of Quranic Phonetics*, 6:55–70, 2021.

[44] Abdul Fattah Al-Qari. *Tajweed and the Science of Qira'at.* Darussalam, 2000.

[45] Abul Hasan Ali Nadwi. *The Qira'at: A Study of the Ten Readings of the Qur'an.* Islamic Foundation, 1997.

[46] Abu Bakr Ibn Mujahid. *Kitab al-Sab'a fi al-Qira'at.* Dar al-Fikr, 936.

[47] Ibn al Jazari. *An-Nashr fi al-Qira'at al-Ashr.* Dar al-Kutub al-Ilmiyyah, 1350.

[48] Nafi al Madani. *Qira'ah of Nafi (Narrators: Warsh and Qalun).* 169 AH. Nafi al-Madani (d. 169 AH) was a master reciter from Medina. He studied under 70 of the Tabi'in who had learned from companions like Ubayy ibn Kab and Ibn Abbas.

[49] H. Zaki. Exploring regional variations in quranic recitation. *Islamic Culture Journal*, 34(2):134–148, 2021.

[50] M. Ali. Phonetic analysis of quranic recitation. *Journal of Linguistics*, 54(3):225–240, 2018.

[51] T. Mansour. Melodic patterns in quranic recitation: An analytical study. *Middle Eastern Musicology*, 9(2):67–82, 2018.

[52] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[53] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA, 2006.

[54] James K. Baker. Stochastic modeling for automatic speech understanding. *Speech Communication*, 11(2):181–196, 1992.

[55] Alexander I. Rudnicky and Michael J. Fegan. Interactive speech recognition: A speech understanding system for man-machine communication. *Communications of the ACM*, 41(8):70–79, 1998.

[56] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2001.

[57] Stephanie Seneff. Tina: A natural language system for spoken language applications. *Computational Linguistics*, 18(1):61–86, 1992.

[58] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[59] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[60] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[61] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[62] Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[63] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[64] Andy Liaw and Matthew Wiener. Classification and regression by randomforest. *R News*, 2(3):18–22, 2002.

[65] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[66] Robert E. Schapire and Yoram Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37(3):297–336, 1999.

[67] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

[68] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[69] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

[70] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

[71] Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

[72] Zhi-Hua Zhou. *Ensemble Methods: Foundations and Algorithms*. Chapman and Hall/CRC, Boca Raton, FL, 2012.

[73] Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

[74] Michael P. Perrone and Leon N. Cooper. When networks disagree: Ensemble methods for hybrid neural networks. In *Neural Networks for Speech and Image Processing*. Chapman-Hall, 1993.

[75] H. A. Hassan, N. H. Nasrudin, M. N. M. Khalid, A. Zabidi, and A. I. Yassin. Pattern classification in recognizing qalqalah kubra pronunciation using multilayer perceptrons. In *2012 IEEE Symposium on Computer Applications and Industrial Electronics (ISCAIE)*, pages 325–329, 2012.

[76] Mahmoud Al-Ayyoub, Nour Alhuda Damer, and Ismail Hmeidi. Using deep learning for automatically determining correct application of basic quranic recitation rules. *The International Arab Journal of Information Technology*, 15(3A), 2018. Special Issue.

[77] M. S. Khorsheed and A. M. Al-Thubaity. Automatic quranic recitation recognition using hidden markov models. In *2013 International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–6, 2013.

[78] Ali M. Alagrami and Maged M. Eljazzar. Smartajweed: Automatic recognition of arabic quranic recitation rules. *arXiv preprint arXiv:2101.04200*, 2019.

[79] Teddy S. Gunawan and colleagues. Development of quranic reciter identification system using mfcc and gmm classifier. *International Journal of Electrical and Computer Engineering (IJECE)*, 8(1):70–76, 2018.

[80] S. A. Chowdhury, H. Mubarak, A. Abdelali, S. Jung, B. J. Jansen, and J. Salminen. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6203–6212, 2020.

[81] N. Farra, K. McKeown, and N. Habash. Annotating targets of opinions in arabic using crowdsourcing. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 89–98, 2015.

[82] W. Zaghouani and K. Dukes. Can crowdsourcing be used for effective annotation of arabic? In *LREC*, pages 224–228, 2014.

[83] S. Bougrine, A. Chorana, A. Lakhdari, and H. Cherroun. Toward a web-based speech corpus for algerian dialectal arabic varieties. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 138–146, 2017.

[84] A. Abid, A. Abid, and A. Abdalla. Tarteel initiative white paper. `https://www.scribd.com/document/406489298/Tarteel-Whitepaper`, 2019. SCRIBD.

[85] H. M. Osman, B. S. Mustafa, and Y. Faisal. Qdat: a data set for reciting the quran. *International Journal on Islamic Applications in Computer Science And Technology*, 9(1):1–9, 2021.

[86] R. Salameh, M. Al Mdfaa, N. Askarbekuly, and M. Mazzara. Quranic audio dataset: Crowdsourced and labeled recitation from non-arabic speakers. *arXiv preprint arXiv:2405.02675v1*, 2024.

[87] Anne Aldahi. Quran recitation dataset (qdat). `https://www.kaggle.com/datasets/annealdahi/quran-recitation`, 2023. Accessed: 2025-04-13.