

UNIVERSITY OF BAHRAIN

College of Information Technology



**Toward a More Accurate
Speech-to-Text Transcription Services
for the Bahraini Dialect**

A Thesis Submitted in Partial Fulfilment of the Requirements for the M.Sc
in Machine Learning and Computational Intelligence

Submitted By

Abdulla Almahmood

202100222

Supervised By

Dr. Hesham Al Ammal

(Assistant Professor)

Co-Supervisor

Dr. Fatema AlBalooshi

(Assistant Professor)

University of Bahrain

August 2024

DECLARATION

I declare that this work is the result of my own investigation and that it has not already been accepted in substance for any degree, nor is it currently submitted for any degree.

Signed: *Abdulla* Date: 01-May-2024

Abdulla Basil Ebrahim Al-Mahmood (Candidate)

The Thesis defense committee considers the thesis titled:

Toward a More Accurate Speech-to-Text Transcription Services for the Bahraini Dialect

Submitted by **Abdulla Basil Ebrahim Al-Mahmood** to the College of Information Technology is satisfactory and acceptable for the M.Sc in Machine Learning and Computational Intelligence

The Defense Committee:

Dr. Hesham Al Ammal

(Supervisor)

Department of Computer Science
University of Bahrain

Dr. Fatema AlBalooshi

(Co-Supervisor)

Department of Computer Engineering
University of Bahrain

ABSTRACT

This study explores innovative techniques to optimize the accuracy of Automated Speech Recognition (ASR) systems for the Bahraini dialect. Through the creation of custom audio datasets tailored for Bahraini speech, we investigated various advanced techniques to fine-tune the OpenAI Whisper model. Whisper, a multilingual speech recognition model, showed significant improvement in processing the Bahraini dialect after fine-tuning. We successfully reduced the Word Error Rate (WER) from 181.608% to 13.54% by utilizing these datasets that were specifically developed for this research.

Contents

APPROVAL	A
DECLARATION	A
ABSTRACT	B
TABLE OF CONTENTS	C
LIST OF FIGURES	E
LIST OF TABLES	F
ACKNOWLEDGMENTS	G
DEDICATION	H
LIST OF ABBREVIATIONS	I
1 Introduction	1
1.1 Overview	1
1.2 Research Problem	2
1.3 Research Objectives	2
1.4 Contribution to the Field	3
1.5 Thesis Outline	4
2 Literature Review	5
2.1 Literature Review	5
2.1.1 Historical Overview of ASR Development	6
2.1.2 Enhancing Speech Recognition: The Crucial Role of Mel Frequency Cepstral Coefficients	8

2.2	Dialectal and Accent Variations in Automatic Speech Recognition	10
2.2.1	The Variability of the Arabic Language	11
2.3	Advancements in Automatic Speech Recognition	12
2.4	Comparison of ASR Models for Dialectal Arabic	15
3	Datasets	17
3.1	Creating a Dialect Specific Speech Dataset	17
3.1.1	Background on Current Available Datasets	17
3.1.2	Sanaa Al Saad Manual Transcriptions	18
3.1.3	Manual Transcription Process	19
3.1.4	Expanded Dataset	19
3.1.5	Data Augmentation	23
3.2	Methodology	24
3.2.1	Hyperparamater Optimization	24
3.2.2	Whisper Fine-Tuning	27
4	Experimental Results	32
4.0.1	Transfer Learning Strategy	33
4.1	Experimental Results	34
4.1.1	Performance Metric	40
4.2	Comparison with Commercial Speech-to-Text Models	41
5	Discussion, Limitations, and Conclusion	43
5.1	Summary	43
5.2	Limitations	44
5.3	Recommendations and Future Work	45
5.4	Conclusion	46
	References	48
	Abstract (in Arabic)	

List of Figures

2.1	Whisper architecture	13
3.1	Proportion of audio files in the dataset based on source . . .	21
3.2	Word Cloud Showing the Prevalence of Important Terms in Arabic Transcriptions: This visual aid highlights the essential linguistic components in the dataset by displaying the frequency and prominence of the most frequently used terms	22
3.3	List of Figures Entry	30
3.4	An outline of the major steps of the methodology	31
4.1	Optuna optimization for the Whisper model on the Sanaa Alsaad dataset	37
4.2	WER improvement is notable in Trial 5 during Optuna optimization. Total compute hours is 45.3557 hours	38
4.3	The graph shows training loss decreasing over epochs. . . .	40
4.4	The graph represents WER decreasing over epochs, illustrating the model's improved generalization.	41

List of Tables

3.1	Optuna’s Hyperparameter Search Space	25
3.2	Optimal Hyperparameters Identified by Optuna for Whisper Model Training	28
4.1	Comparison of Initial and Fine-Tuned Whisper Model Per- formance	34
4.2	Performance Metrics of the Whisper-Medium Base Model .	36
4.3	Performance metrics of the Whisper-medium model after fine-tuning.	39
4.4	Performance Comparison of Fine-tuned Whisper Model and Commercial ASR Solutions	42

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my Advisor, Dr. Hesham Al-Ammal and my Co-Supervisor, Dr. Fatima Al-Balooshi for their great support, help, and motivation. Their guidance helped me in all the time of my research and during the writing of this thesis.

Special thanks to Dr. Dana Abdulrahim and Dr. Latifa, together with their hardworking group of volunteers, for their invaluable contributions to the dataset's compilation.

I am also grateful to the Benefit Advanced AI Lab and Benefit Bahrain for their valuable support by providing the resources necessary to conduct this study. Their commitment to supporting research in ML and AI technology is sincerely appreciated.

DEDICATION

To my family for their encouragements.

LIST OF ABBREVIATIONS

ASR	Automated Speech Recognition
MSA	Modern Standard Arabic
WER	Word Error Rate
DID	Dialect Identification
CSV	Comma-Separated Values
GPU	Graphics Processing Unit
NLP	Natural Language Processing
MFCC	Mel Frequency Cepstral Coefficients
DNN	Deep Neural Networks
CNN	Convolutional Neural Networks
RNN	Recurrent Neural Networks
HMM	Hidden Markov Models
LVCSR	Large Vocabulary Continuous Speech Recognition
TPE	Tree-structured Parzen Estimator
DL	Deep Learning
ML	Machine Learning

Chapter 1

Introduction

1.1 Overview

The demand for accurate Automated Speech Recognition (ASR) systems is higher than ever. However, the development of such systems is challenging due to the poor performance of ASR models when confronted with dialect variations, particularly in languages with rich regional diversity like Arabic (Chowdhury, Hussein, Abdelali, & Ali, 2021). This thesis focuses on developing a speech-to-text model specifically tailored for accurately transcribing the Bahraini dialect—a variant that has received minimal attention in natural language processing research. The goal is to fine-tune the OpenAI Whisper model to enhance its transcription accuracy for the Bahraini dialect. This study aims to significantly reduce the Word Error Rate (WER) and improve the usability of ASR systems for Bahraini Arabic speakers across various applications, from personal assistants to automated customer service platforms.

Smart agents that enable seamless human-machine interactions—especially in customer service and assistive technologies—are largely dependent on the incorporation of ASR technology (Karlos, Fazakis, Karanikola, Kotsiantis, & Sgarbas, 2016). We can improve the interaction between technology and local populations by creating ASR systems that can accurately

transcribe local dialects, which will help in making services more accessible and customized to meet specific linguistic needs.

Furthermore, correct transcription and understanding of spoken language are crucial for processing both audio and video data, which is a major catalyst for the development of specialist datasets for regional dialects. Along with helping to create more reliable ASR models, these datasets advance our knowledge of the subtle differences in language across various populations (A. Ali et al., 2015).

1.2 Research Problem

The primary issue this thesis attempts to address are the high word error rate (WER) of current ASR models when addressing the Bahraini dialect and the severe lack of customized datasets that appropriately capture the dialect's features. Due to these difficulties, ASR models' recognition accuracy is generally low, which restricts the usefulness of ASR systems for Bahraini dialect speakers.

1.3 Research Objectives

The following is an outline of this thesis's objectives:

1. To create an audio dataset customized for the Bahraini dialect.
2. To greatly increase the OpenAI Whisper model's accuracy in transcribing the Bahraini dialect by fine-tuning it with this dataset.
3. To show that these fine-tuning methods are effective in lowering the WER compared to the initial benchmarks.

1.4 Contribution to the Field

This thesis contributes significantly to the field of automated speech recognition by adapting current technologies to accurately recognize low-resource languages and dialect, such as the Bahraini dialect. Such adaptations help to strengthen the linguistic research carried out and the implementation field of Natural Language Processing. It also helps in handling particular technical challenges related to dialect variability. The major contributions of this work include:

1. The research has effectively illustrated how a general-purpose ASR system can be adapted for a regional dialect that has historically been underrepresented in speech recognition research. This was accomplished by fine-tuning the OpenAI Whisper model for the Bahraini dialect. Extensive testing was done to optimize training methods and hyperparameter tuning for the unique linguistic and acoustic characteristics of the Bahraini dialect.
2. Building a Specialized Audio Dataset: A significant obstacle facing dialectal ASR is the severe lack of high quality, annotated speech datasets. In order to fill this void, this thesis has constructed two datasets containing a variety of Bahraini speech samples, from official media broadcasts to informal talks. This dataset serves as a cornerstone for this research and all upcoming ASR researches and developments.
3. Decrease in Word Error Rate (WER): The study's methodological fine-tuning resulted in a significant decrease in WER, improving the model's accuracy over its initial benchmarks. This reduction is an important milestone in developing accurate voice-activated systems and transcription services for the Bahraini dialect.

4. Contributions to Linguistic Diversity in ASR: This thesis advances the main objective of linguistic inclusivity in technology by focusing on the Bahraini dialect. Other Arabic dialects and potentially other languages that are similarly underrepresented in speech technology research can benefit from the findings and methodologies developed here.
5. Possibility for Wider Impact: The fine-tuned ASR model has potential for a number of uses, such as accessibility tools, educational technology, and improved customer service in Bahrain and other comparable areas.

Overall, this thesis contributes to the field of computational linguistics by showing how targeted fine-tuning might address the difficulties posed by dialectal variation.

1.5 Thesis Outline

This thesis is organized as follows. In Chapter 2, we examine the existing literature on Automated Speech Recognition (ASR), focusing on dialectal variations within Arabic. Chapter 3: Dataset Development describes the production of a customized audio dataset for the Bahraini dialect, including collecting, annotation, and preprocessing. Chapter 4 discusses the methodologies used to fine-tune the OpenAI Whisper model using the two developed datasets. Chapter 5: Results and Discussion, highlights the results of the fine-tuning process, such as increased accuracy and lower Word Error Rate (WER). Chapter 6: Conclusion and Future Work summarizes the findings and explores potential future paths for research in ASR systems that are tailored to regional dialects.

Chapter 2

Literature Review

2.1 Literature Review

This section gives a detailed assessment of the literature on Automated Speech Recognition (ASR) technologies. The chapter is divided into three major sections. The first section examines the historical perspective on the evolution of ASR and its technological milestones. The second section discusses the ongoing challenges in the field such as handling dialectal and accent variations. The third section showcases the advancements in ASR technology and the latest research and achievements in this field.

Before beginning our research, it is critical to do a thorough review of related papers. Reviewing existing literature helps to discover current trends and patterns in ASR technology, particularly when dealing with dialectal variances. This helps in determining the strengths and limitations of current approaches, and identify areas that require more exploration.

2.1.1 Historical Overview of ASR Development

Since its foundation, Automated Speech Recognition (ASR) has undergone substantial advancement, driven by both broader advances in technology and the continuous evolution in signal processing and computational linguistics. This in-depth historical narrative outlines the turning points and significant changes in the evolution of ASR systems.

Bell Laboratories created the Audrey system, which could identify single digits spoken by a restricted range of voices. This early ASR system laid the groundwork for more sophisticated speech recognition technologies by using basic acoustic models based on sound wave frequency. Audrey was significant because it showed that, although in a restricted way, electronic systems could understand spoken language. This early achievement opened the door for additional studies and advancements in the area (Gold, Morgan, & Ellis, 2011).

A major breakthrough in ASR occurred in the late 1970s and early 1980s with the emergence of Hidden Markov Models (HMM). By offering a strong statistical foundation for modeling speech sequences, HMMs improved the system's capacity to identify continuous, natural speech more effectively. During this time, ASR made the shift from simple digit recognition to increasingly difficult speech recognition tasks (Espy-Wilson, 2005).

In the 1980s and 90s, as neural networks were introduced into ASR systems, they started to tackle some of the nonlinear complexities of speech that had proven difficult for traditional models. Early neural network integration helped ASR systems become more adaptive to speech variations by enhancing their capacity to generalize from training data (Helander, Moody, & Joost, 1988).

Large vocabulary continuous speech recognition (LVCSR) systems were developed by the 1990s, which signaled the start of another important period. These systems were the first to accomplish speaker independence and

could recognize thousands of words, exposing the possibility of using ASR in non-controlled applications. The commercialization of speech recognition technologies was heavily affected by this milestone (O'Shaughnessy, 2008).

The increasing deployment of ASR systems in noisy and diverse environments led to a significant focus on enhancing noise robustness. Advanced acoustic modeling and noise reduction algorithms are two techniques that have been developed to improve ASR accuracy in suboptimal auditory environments. These developments made it possible to use ASR in practical applications like public service announcements and automotive systems. (Ranjeet, Prakash, Amruta, & Monali, 2016).

With the introduction of deep learning techniques in the 2000s, ASR underwent its most revolutionary period. The accuracy of ASR has been significantly increased by these methods, which include the use of Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). A thorough assessment of the literature covering the noteworthy advancements in ASR brought about by deep learning is presented by (Alharbi et al., 2021). Their analysis demonstrates how these cutting-edge computational methods have made it possible for ASR systems to handle multi-accent and complex speech patterns with never achieved before accuracy (Alharbi et al., 2021).

Modern ASR systems still have a lot of work ahead of them, especially when it comes to handling conversational language, multiple speakers, and loud background noise. (Padmanabhan & Johnson, 2015) survey explores these issues and discusses potential directions for future work. Their research highlights the need for more context-aware and adaptive ASR systems in order to increase accuracy and dependability in a variety of dynamic contexts (Padmanabhan & Johnson, 2015).

This ASR's historical development illustrates how technology advance-

ment and the ever-evolving demands of human-computer interaction interact. Thus, every study mentioned gives a richer understanding about the origin of ASR and its possible direction in the future. In an effort to create systems that are as adaptable and reliable as human communication, continuous research and development efforts are pushing the limits of speech recognition technology.

2.1.2 Enhancing Speech Recognition: The Crucial Role of Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) have been influential in the advancement of Automatic Speech Recognition (ASR) technology, as demonstrated by a number of studies that have investigated both primitive and cutting-edge applications. (Karlos et al., 2016) made a valuable contribution by creating a strong classification model that combines spectrogram-based Content-Based Image Retrieval (CBIR) features with MFCCs. By leveraging both audio and visual representations of speech signals, this technique focuses on improving classification accuracy and demonstrates an all in one approach to feature extraction in ASR systems.

The work by (Ahadi et al., 2003) further improved the efficiency of ASR systems by proposing a novel front-end that was proved to be more efficient than conventional MFCC-based systems. Because of its improved architecture over the conventional MFCC extraction method, this front-end is specifically made for low-resource environments. It contains changes that enable almost flawless speech signal reconstruction, which is crucial for speech enhancement tasks prior to recognition. This study demonstrates how the overall efficacy and resource efficiency of ASR systems can be greatly enhanced by optimizing the early phases of signal processing.

The efficacy of Mel Frequency Cepstral Coefficients (MFCCs) and Dy-

dynamic Time Warping (DTW) in ASR systems is demonstrated by (Mohan & Babu, 2014)'s study. This combination performs exceptionally well at capturing speech's characteristics, which makes it ideal for applications like security and healthcare that need accurate and effective speech recognition. This technology improves biometric verification in security by precisely matching voice prints, even when speech variations are present. It helps with condition monitoring and hands-free communication in the healthcare industry, which is important in busy environments. By controlling speech rate and pronunciation variances, the system becomes much more useful in dynamic environments, demonstrating the potential of the combined approach to improve service flexibility and dependability in vital industries.

(Chechi, 2013) conducted a comparative study between MFCC and Linear Predictive Coding (LPC) methods in ASR systems. The main focus was on the computational algorithms crucial for extracting speech features. LPC offers a complementary method by estimating the vocal tract configuration, while MFCCs are utilized to capture the cepstral properties of speech. This research's dual subsystems, each processing a single feature set, enable a complete analysis of speech inputs which have the potential of reducing error rates during recognition.

(H. Ali et al., 2013) conducted a study that examined how well MFCCs performed in ASR frameworks when paired with DWT features. The study looked at how these various feature extraction techniques impact system accuracy using an isolated words corpus in Urdu. While DWT offers an alternative method by concentrating on time-scale characteristics, MFCCs offer a time-frequency analysis that is well-suited to capturing the subtleties of speech. Furthermore, in order to examine the advantages and disadvantages of each feature set in speech recognition tasks, the comparative study applies Linear Discriminant Analysis (LDA) for classification.

2.2 Dialectal and Accent Variations in Automatic Speech Recognition

For Automatic Speech Recognition (ASR) systems, understanding and modeling the individual variation in spoken language poses significant challenges. The design of robust, speaker-independent ASR systems is challenged by variations caused by dialects and accents, which are influenced by various factors such as geographical location and social background.

Disparities in phonetic articulation, vocabulary use, and syntactic structures among speakers of the same language are examples of dialectal variations. These variations can lower the accuracy of ASR models. In order to improve these models, (Hirschberg & Biadsy, 2011) investigate automatic dialect and accent recognition as a strategy to enhance these models, which involves modifying systems to identify and process the distinctive features of regional speech. Their work shows how differentiating between dialects can be achieved by utilizing a variety of phonetic and acoustic features, and by integrating dialect-specific models. By doing so, they have significantly improved ASR accuracy by incorporating dialect-specific models (Hirschberg & Biadsy, 2011).

Since accents differ from the standard speech models commonly used in ASR systems, they present a significant challenge, especially for non-native speakers of a given language. (Zhang et al., 2006) address these issues by creating a Pronunciation Variation Model (PVM) for Mandarin, which demonstrate a notable decreases in error rates in accent speech recognition. This method emphasizes how crucial it is to modify ASR systems in order to properly handle accent variations (Zhang et al., 2006).

Recent studies have increasingly focused on creating techniques that allow ASR models to handle different dialect and accents. The study of (Winata et al., 2020) introduced a new approach to enhance ASR systems'

ability to adapt to different accents using Model-Agnostic Meta-Learning (MAML). Using a number of tasks for training, this method allows quick fine-tuning with little data from new accents, enabling ASR systems to quickly adapt to new accents. The ability for the ASR system to rapidly adapt new accents is achieved by using an accent-agnostic approach. This approach reduces the system’s reliance on accent features of the training data.

2.2.1 The Variability of the Arabic Language

Training an ASR model on the Arabic language presents a large number of challenges. For example, most written Arabic datasets lack diacritics; this makes it challenging for ASR models to comprehend vowel sounds, which are essential for accurate transcription (Abdou & Moussa, 2019). Furthermore, the authors explain how Arabic is a morphologically rich language, with a large number of prefixes and suffixes which create a vast array of possible word forms. Hence, ASR models will face significant challenges when asked to transcribe spoken Arabic sentences.

Another challenge for ASR models in Arabic is the lack of consistent writing system for dialectal speech (A. Ali et al., 2019). While MSA has a standardized writing system, dialects in Arabic often lack uniform spelling rules. The lack of such rules can increase the difficulty for transcription and alignment algorithms used in ASR systems (A. Ali et al., 2019).

In the field of automatic speech recognition, most research done focused on MSA, while only a small number of studies focused on different Arabic dialects. This is primarily due to the scarcity of datasets which are specifically tailored for those dialects (A. Ali et al., 2015), (Wray & Ali, 2015). For example, a study by (Elmahdy et al., 2014) which created a voice recognition system especially designed to identify Qatari Arabic in television shows. Furthermore, another challenge with the Bahraini dialect is the code-

switching, which is a phenomena of the speaker alternating between the Bahraini dialect and the Modern Standard Arabic (MSA) (Albirini, 2011).

For ASR systems that target the Bahraini dialect, they need to solve several challenges to be deemed useful in real-world applications. First, these ASR systems must be robust in term of their ability to accommodate the large number of variation within the Bahraini dialect. Second, these systems should be able to transcribe speech into a written standard that accurately represent the linguistic diversity in Bahrain.

2.3 Advancements in Automatic Speech Recognition

The advancements in ASR technology is linked by improved model robustness and error handling capabilities (Martucci et al., 2021). The inclusion of reliable noise-cancellation algorithms and the ability for the models to adapt to a wide range of dialects have significantly improved the usability of ASR systems (Vu et al., 2014). Also, with the introduction of deep learning models, the need for traditional linguistic models is no longer the case (Dubey & Shah, 2022). These enhancements increased the speed and accuracy of transcriptions, which in term improved the usefulness of ASR systems in multilingual contexts, demonstrating their increasing significance in international low resource research (Hassan et al., 2022).

One of the recent and promising ASR models is Whisper which was introduced by OpenAI. In their paper, (Radford et al., 2023) suggest that weakly supervised pre-training has received relatively little attention in the literature. The authors draw attention to the potential of overfitting data to a training set during the unsupervised learning process of fine-tuning. Although this method is good at identifying patterns in the training data, these might be brittle patterns that do not apply in general cases. However,

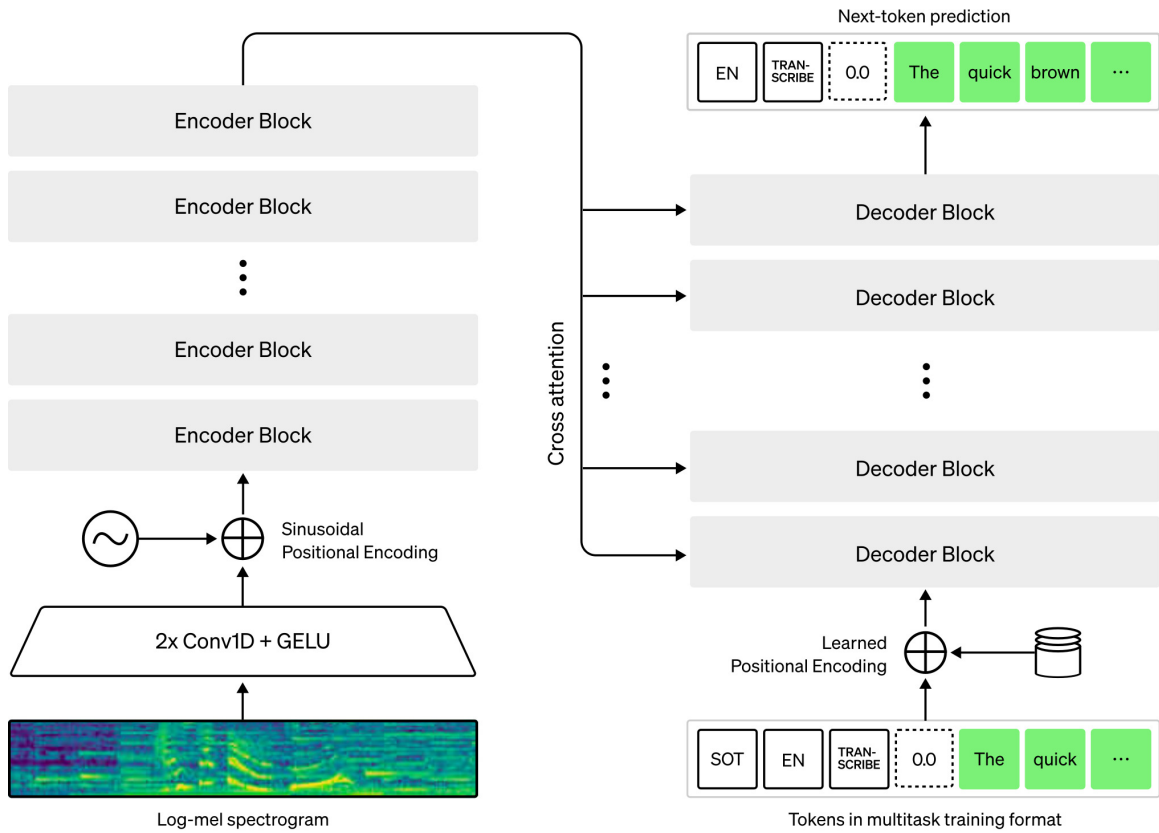


Figure 2.1: Whisper architecture
(Radford et al., 2023)

Whisper stands out for its robust design that allows it to transcribe a wide range of languages and acoustic environments.

As shown in Figure 2.1 (adapted from Radford et al., 2023), the architecture of Whisper is a sequence-to-sequence model that consists of a log-mel spectrogram where audio signals are converted into their correspondent log-mel spectrogram, which is a visual representation of the frequency of audio over a specific period. Then, the transformed spectrogram passes through two 1-dimensional convolutional layers where each have a GELU activation function. These layers help in extracting the low-level audio features from the log-mel spectrogram. Since transformers are not sensitive to the sequence order, the Sinusoidal Positional Encoding is implemented to provide

the model with information about the order of the input frames. There are several identical blocks that make up the transformer's encoder. Each block includes Multi-head self-attention and feed-forward neural network layers. The self-attention mechanism enables the model to assign different weights to distinct segments of the input sequence. While producing the output, the cross attention component of the decoder blocks uses the encoder's output to help it in focusing on specific segments of the input sequence. Like the encoder, the decoder consists of several identical blocks. By using cross-attention, each block in the decoder can attend to all positions in the input sequence, and by using masked self-attention, it can attend to every position in the output sequence. To provide the model with information about the position of each token in the sequence, learned positional encoding is appended to the input embeddings at the base of the decoder stack. In contrast to the encoder's sinusoidal positional encoding, this one is learned, meaning that the model fine-tunes it while it is being trained. The output of the model then predicts the next token in the sequence and the process continues until the model produces an end-of-sequence token (Radford et al., 2023).

(Gris et al., 2023) shows the robustness of Whisper in handling different challenges such as punctuation prediction in the Portuguese language. The study claims that the Whisper outperformed previous models in predicting punctuation marks. Whisper also perform exceptionally well in domain-specific applications as demonstrated in the study by (Oh et al., 2023). In this study, the authors have fine tuned Whisper on a specialized Korean dataset, showing a significant improvement in Korean speech recognition when compared with the baseline model (Oh et al., 2023). Each of these examples highlights how Whisper may transform the field of speech recognition by offering a scalable and adaptable solution for a wide range of speech-related applications.

Recent development in automatic speech recognition (ASR) has given a

strong incentive to the necessity of building systems to handle multilingual inputs and cross-accented speech. The authors of this article, (Winata et al., 2020), were the first to develop a model-agnostic meta-learning approach for ASR, which surprisingly showed good adaptability on unseen accents while outperforming the conventional joint training methods. This shows that ASR systems can be easily fine-tuned to new linguistic setups, hence become adaptable to be more effective for real-world applications. In addition, the work by (Khandelwal, Jyothi, Awasthi, & Sarawagi, 2020) shows some innovative ways to adapt black-box, cloud-based ASR to target accents, hence merging outputs with local models which reduced word error rates tremendously for Indian and Australian accents.

2.4 Comparison of ASR Models for Dialectal Arabic

Several Automatic Speech Recognition (ASR) models have recently demonstrated promise in dealing with dialectal Arabic. This section compares the Whisper model to other popular ASR models, highlighting their advantages and disadvantages in terms of Bahraini Arabic transcription.

DeepSpeech, created by Mozilla, employs a recurrent neural network (RNN) architecture. While it is effective in many languages, its performance in dialectal Arabic is limited because it is based on a language model that does not adequately capture the nuances of Arabic dialects (Hannun et al., 2014; A. Ali et al., 2019; Khurana & Ali, 2016).

Wav2Vec, introduced by Facebook AI, performs self-supervised learning on unlabeled audio data (Baevski, Zhou, Mohamed, & Auli, 2020; Conneau, Baevski, Collobert, Mohamed, & Auli, 2020). This approach has produced positive results for low-resource languages, making it a strong contender for dialectal Arabic. However, because of a lack of dialect-specific

pre-training data, its performance may vary across Arabic dialects.

Jasper, an end-to-end convolutional neural network for speech recognition, has achieved high accuracy in English and other well-resourced languages (Li et al., 2019). However, its use in dialectal Arabic has been limited, and it may require significant adaptation to handle the unique phonetic and lexical features of Bahraini Arabic (A. Ali et al., 2019). The difficulties in adapting Jasper to Bahraini Arabic stem from the dialect's distinct linguistic characteristics, including its phonetic inventory, lexical variations, and morphological complexity (A. Ali et al., 2019).

OpenAI's Whisper model stands out for its multilingual capabilities and strong performance in a variety of acoustic environments. Its transformer-based architecture and training on a wide range of audio data make it ideal for adapting to dialectal differences (Radford et al., 2023).

Chapter 3

Datasets

3.1 Creating a Dialect Specific Speech Dataset

Our model’s training procedure relied on two datasets, each uniquely contributing to the learning process of the model. However, before discussing these datasets, it is important to understand to explore the current available audio datasets and the pressing need for localized, dialect-specific datasets.

3.1.1 Background on Current Available Datasets

The field of ASR has historically been dominated by datasets the are extensive in quantity but limited in its linguistic diversity. Popular datasets such as the Common Voice and the LibriSpeech are mainly focused on English-language data with representation from dominant dialects, usually American, British and Australian English (Panayotov, Chen, Povey, & Khudanpur, 2015; Ardila et al., 2019).

These widely used datasets are essential for training models that can understand General American or any other sufficiently represented language or dialect; however, using such datasets to train models is insufficient for poorly represented languages or dialects.

There is an urgent need for custom datasets for areas like Bahrain, where

the local dialect displays distinctive linguistic features not captured by popular ASR models. There is frequently a lack of representation for dialect-specific characteristics like intonation, pronunciation, and the use of local vocabulary. For instance, the phonetic and lexical differences between the Bahraini dialect and other Arabic dialects and Modern Standard Arabic are significant. According to (A. Ali et al., 2015), this variation may result in high transcription error rates when using models trained on non-local datasets, which could lower the efficacy and user acceptability of ASR technology in regional applications.

Additionally, by creating and utilizing localized datasets, technologies like voice-activated assistants, customer service bots, and educational tools can be improved and made more accessible and functional for the local population. This, in turn, promotes inclusivity in technology and enhances user interaction.

3.1.2 Sanaa Al Saad Manual Transcriptions

The first dataset used in our training process contains manually transcribed audio. The audio was carefully transcribed by our team to ensure high accuracy transcriptions. Spanning 569.9 minutes, this dataset represents a comprehensive collection of spoken content presented by Sanaa Alsaad, who is a famous Bahraini TV host.

The show was acquired from Youtube in an audio only format (Alsaad, 2023). Audio from each show was segmented into chunks based on natural pauses between sentences. In addition, to ensure high quality training data, background noise was carefully removed and the loudness of each audio was appropriately adjusted using a specialized software called Audacity. These pre-processing steps are essential to ensure the best learning environment for the model.

3.1.3 Manual Transcription Process

The manual transcription process was carried out by a team of six, four undergraduate students and two graduate students all from University of Bahrain. All team members were native speakers of Bahraini Arabic and had received a brief training prior to starting the transcription process.

During the transcription process, issues with fast speech, background noise in certain recordings, and variances in colloquial expressions were encountered. Team discussions and, when needed, consultations with linguistic experts were used to address these issues.

3.1.4 Expanded Dataset

In addition to the carefully transcribed Sanaa Al Saad dataset, our training procedure was augmented with an expanded dataset that was developed by us. The dataset was created using diverse audio sources. It encompasses a rich range of content, including Bahraini podcasts, sessions from the Bahraini Parliament, and episodes from the Bahraini TV show "Shams Al Bahrain" (BahrainTV, 2018). Unlike the manually transcribed Sanaa Al Saad transcriptions dataset, the expanded dataset includes a diverse range of audio content. The loosely transcribed audio data from diverse domains exposes our model to real-world scenarios effectively, including audio data that contain background music, background chatter, ambience noise, and multiple speakers. While the transcription accuracy within the expanded dataset may fluctuate, this dataset serves to enrich the model with real-world speech patterns and improve how the model adapt to diverse linguistic contexts.

The dataset was then created using a custom script that automatically generated the text file. This text file is part of the dataset used for training the Whisper model, containing audio file names paired with their respective

transcriptions. Each line in the file represents an audio chunk, identified by a filename, and its corresponding transcription. Here’s how the data is structured:

- **Filename:** This column contains the name of the audio file, indicating the speaker, date, and segment number. For example, "SanaaAlsaad31-01.wav" refers to the first audio segment from a session recorded on January 31st by speaker Sanaa Alsaad.
- **Transcription:** The Arabic text next to each filename is the direct transcription of what is spoken in the audio file. This transcription captures the dialectal nuances and typical expressions of the Bahraini dialect, providing rich linguistic data for the ASR system.

For instance, the file "SanaaAlsaad31-01.wav" is transcribed as سلام الله عليكم معطر بذكر الرحمن وحفظه, each subsequent file contains phrases or sentences that form a continuous narrative, which is typical in conversational speech datasets.

The distribution of audio files by source, as seen in Figure 1, gives insight on the variety and size of the training datasets that were used to fine-tune the ASR model. The majority of files in the dataset (36.59%) begin with "SanaaAlsaad," indicating the substantial contribution of content from the well-known Bahraini TV personality Sanaa Alsaad. This significant percentage is a reflection of the dataset’s reliance on manually transcribed, high-quality audio, which is essential for providing accurate linguistic data to train the model.

18.88% of the audio recordings are from parliament sessions (BahrainTV, 2019), showing a significant amount of formal, structured speech, which is useful for teaching the model to identify more formal variations of the Bahraini dialect. The model is exposed to a wider range of spoken material types, from media presentations to casual conversations, thanks to

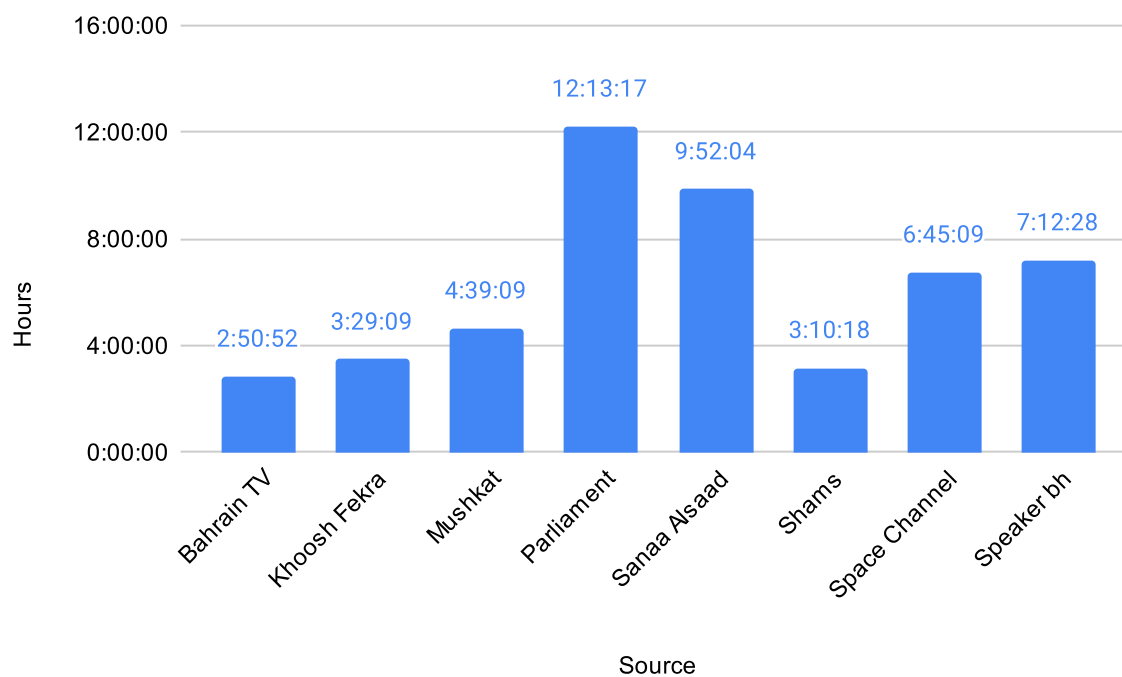


Figure 3.1: Proportion of audio files in the dataset based on source

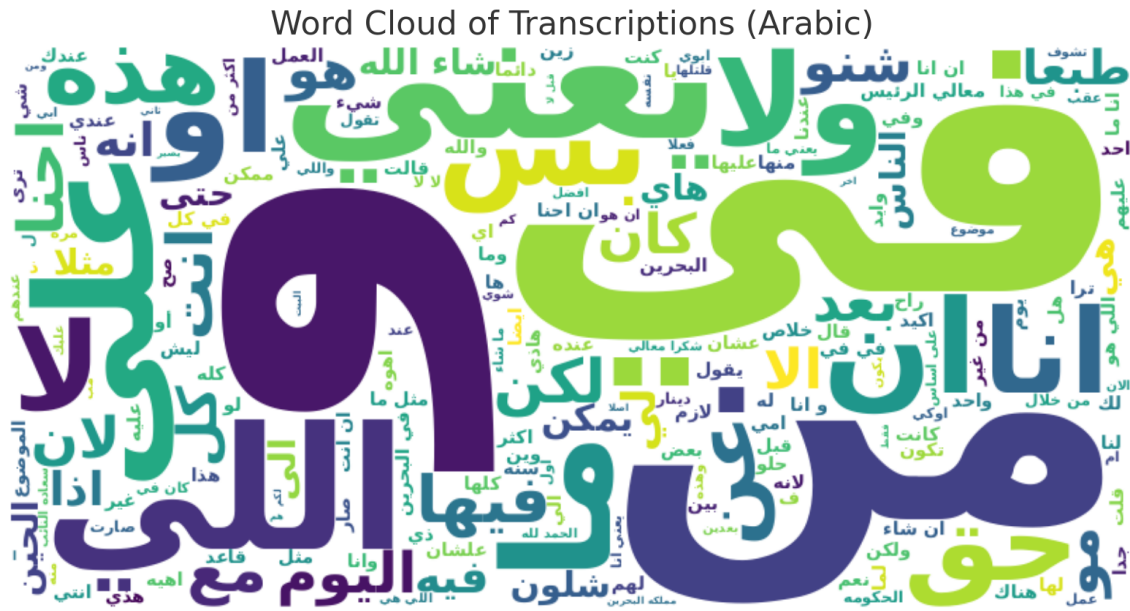


Figure 3.2: Word Cloud Showing the Prevalence of Important Terms in Arabic Transcriptions: This visual aid highlights the essential linguistic components in the dataset by displaying the frequency and prominence of the most frequently used terms

the addition of content from "Mushkat" (pod.mishkah, 2023) and "shams" (BahrainTV, 2018), which account for 7.57% and 4.79% of the total.

The training dataset’s varied audio sources and careful curation contribute significantly to the model’s increased accuracy and adaptability, which are essential for its effective use in identifying and processing the unique traits of the Bahraini dialect.

The most common words in the audio transcription collection are visually represented in the Word Cloud provided in Figure 2. The key phrases and concepts that appear frequently in the dataset are especially well-represented in this visualization. Larger fonts are used to indicate the frequency of usage of words like **يعني** and **أحنا**, suggesting common grammatical structures in conversational Bahraini Arabic. The Word Cloud is a useful tool for quickly finding important words and phrases, but it also helps to grasp the linguistic and thematic emphasis in the transcriptions that have been collected.

3.1.5 Data Augmentation

Augmentation techniques were used to improve the dataset’s variety. These techniques include pitch shifting, which is a technique that involves modifying the audio pitch without altering the speed or duration (Arakawa, Takamichi, & Saruwatari, 2019). Masking frequency, this method replicates the distortion or background noise that is frequently present in audio situations in the real world (Alex, Wang, Gastaldo, & Cavallaro, 2023). Frequency masking increases the model’s resilience to noisy situations by exposing it to fluctuations in audio clarity and fidelity (Alex et al., 2023). Synthetic noise addition, by adding artificial noise to audio signals, noise augmentation mimics environmental disruptions or interference that are frequently present in natural recordings (Zhou, Chen, & Chien, 2022). This improves the model’s performance in noisy situations by strengthening its capacity to distinguish relevant speech signals from background noise (Zhou et al., 2022). Time Stretch addition, which is done by adjusting the speech pace and tempo without compromising essential pitch characteristics (Arakawa et al., 2019). It does this by adjusting the duration of audio segments while retaining pitch. By simulating changes in speech rate, this method improves the model’s capacity to accommodate fluctuations in speaking speed that occur in the actual world (Arakawa et al., 2019).

Following the augmentation procedure, the data was converted into a standardized format that is compatible with HuggingFace, a popular platform for machine learning tasks. This transformation required transforming the augmented data into CSV files with two columns: "sentence" and "path." The "sentence" column contains the transcript of each audio clip, whereas the "path" column specifies the file path to the associated audio file. This standardized format allows for smooth integration with HuggingFace’s training pipeline, which streamlines the model training and evaluation process.

3.2 Methodology

Figure 3.3 illustrates a flowchart summary of the research methodology used in this study. This diagram visualizes the steps we took from the initial data collection to the final model evaluation stage. Each column represents a major phase of the process, including data preparation, model training, and results. A simpler outline can be found in Figure 3.4.

3.2.1 Hyperparameter Optimization

Fine-tuning the Whisper model for the Bahraini dialect required careful consideration of hyperparameter optimization. The learning rate, number of training epochs, seed, and weight decay were the four main hyperparameters that we concentrated on. Optuna, an effective hyperparameter optimization framework, was used to carry out the optimization.

We had a number of difficulties when optimizing. Resolving the trade-off between computing resources and model performance was the most important. While lower learning rates necessitated longer training times, higher learning rates frequently resulted in faster convergence but carried the danger of overshooting optimal values. In order to overcome this, we implemented a learning rate scheduler that modifies the rate dynamically while training.

Finding the ideal number of training epochs was another difficulty. An excessive number of epochs caused over-fitting and higher computing costs, whereas an insufficient number of epochs led to under-fitting. In order to address this issue, we introduced early stopping with a three-epoch tolerance.

Considering our comparatively small dataset, weight decay optimization was essential in avoiding over-fitting. We discovered that the best trade-off between model generalization and performance was offered by a moderate

weight decay value of 0.01.

These optimization efforts resulted in significant improvements in the model’s performance, reducing the Word Error Rate (WER) from 181.61

Several factors have lead to the decision of using Optuna for hyperparameter optimization. The first factor is that Optuna uses a highly sophisticated approaches for handling large search spaces like Tree-structured Parzen Estimator (TPE) models. Second, it has automatic pruning to early cut off poorly performing trials. For our project, the primary parameters that we have explored are learning rate, number of training epochs, seed, and weight decay. Table 3.1 provides the search space for the hyperparameter tuning.

Hyperparameter	Description
Learning Rate	Varied logarithmically between 1×10^{-5} and 1×10^{-3} . This range optimizes the model’s learning rate without overshooting by providing fine-grained control over the learning process.
Number of Training Epochs	varied between 1 and 5. This parameter, which balances between underfitting and overfitting, indicates how many times the model should run through the complete training dataset.
Seed	Ranged from 1 to 10. Using different seeds ensures that the results are robust and not due to a lucky initialization of model weights.
Weight Decay	Varied logarithmically between 1×10^{-3} and 1×10^{-1} . Larger weights are penalized during the model’s learning process, which helps avoid overfitting.

Table 3.1: Optuna’s Hyperparameter Search Space

Optuna was used to find the best set of hyperparameters for fine-tuning our Whisper model (Yadavalli, Mirishkar, & Vuppala, 2022). Optuna is an open source hyperparameter optimization framework that is designed

to automate finding the best hyperparameters for a given machine learning model. It uses advanced algorithms such as tree-structured Parzen estimators (TPE) and Bayesian optimization algorithms that are able to efficiently explore the hyperparameter space. The framework highly adaptable and can be easily integrated with a range of machine learning libraries, such as Scikit-learn, TensorFlow, and PyTorch.

Optuna’s Core Components: The two components that optuna operates on are the study and the trial. The study component is the overall optimization task, and the trial component is the evaluation of a set of hyperparameters.

As can be seen in Equation 3.1, Optuna attempts to minimize a user-defined objective function and returns a scalar score to be optimized.

$$\min_x f(x) \quad (3.1)$$

The optimization process in Optuna requires an objective function that is used to guide the search towards the optimal parameters. The performance of a value-based model on specified hyperparameters is calculated by Equation 3.2.

$$EI(x) = E \max(f(x^*) - f(x), 0) \quad (3.2)$$

The expression inside the expected value $\mathbb{E}$ calculated the maximum of the difference $f(x^*) - f(x)$ and zero. By doing so, we effectively ignore any negative values that result from this difference. This is because a negative values means that $f(x^*) - f(x)$ is higher which indicates no improvement. After that, the expectation E averages these maximum values while taking into account the probability distribution of x , which may have been previously evaluated or known. After that, the objective function is used to train the model using the hyperparameters provided by the user and outputs a performance measure. Then, by employing an effective sampling

technique, Optuna will suggest the next set of hyperparameters by looking for values that could indicate certain regions within a hyperparameter space that would result in improved performance as shown in the Tree-structured Parzen Estimator(TPE) algorithm presented in Equation 3.3.

$$p(x | y) = \frac{p(y | x)p(x)}{p(y)} \quad (3.3)$$

Optuna has a number of advanced features that increase the effectiveness of ML tasks. One of the main characteristics is automated pruning, which is a method that reduces the length of some trials to conserve computing power and optimize the speed of finding optimal hyperparameters. The ability to parallelize trials is another important characteristic that helps in distributed optimization. This allows trials to run concurrently on multiple machines, which speeds up the process of finding hyperparameters. Additionally, Optuna offers a wide range of tools for analyzing the optimization process and drawing conclusions from research results. These capabilities include plotting the optimization history, contour charts, and parallel coordinate plots.

3.2.2 Whisper Fine-Tuning

Whisper utilizes a specific Transformer architecture that is called an encoder-decoder Transformer architecture (Charola, Kachhi, & Patil, 2023), (Huang, Xing, & Liu, n.d.). This architecture processes the audio file by dividing it into 30-seconds chunks, resampling the chunks to 16,000 Hz and finally converting the chunks into 80-channel Mel spectrogram (Mulfari & Villari, 2024). An 80-channel Mel spectrogram is a graphical representation of audio signals that divides the frequency range into 80 logarithmically divided channels. These channels divide the frequency range based on how humans perceive sound. This process efficiently captures detailed frequency information, which makes it useful for ASR related tasks (Charola et al., 2023).

Hyperparameter	Value
Learning Rate	0.00010357075845237331
Num Train Epochs	5
Seed	8
Weight Decay	0.0031325225870988543

Table 3.2: Optimal Hyperparameters Identified by Optuna for Whisper Model Training

Then, after converting the audio chunks into 80-channel Mel spectrogram, a feature normalizer is applied followed by a two layer convolutional encoder that uses GELU as the activation function. After that, sinusoidal position embeddings followed by transformer blocks are applied (Huang et al., n.d.).

For our fine-tuning process of the Whisper model, we have leveraged the capabilities of the HuggingFace Transformers library (Wolf et al., 2020). We have also adapted a pre-configured WhisperFeatureExtractor class with a sampling rate of 16kHz and feature size of 80. The audio that is used to train this model was divided into 30 seconds slices with a hop length of 160 samples. We have used the Whisper medium checkpoint, which is a publicly available version that is trained on 769 M parameters (Radford et al., 2023). Our approach consists of altering some hyper parameters and fine-tuning techniques. We have utilized a hyperparameter optimization framework called Optuna to find the best hyperparameters for our model’s fine-tuning procedure. This process spanned for around 326 hours using the computational power of Nvidia A100 GPUs which is hosted on the Benefit Advanced AI and Computing Lab cloud platform. Through 21 trials, Optuna explored the hyperparameter space and identified the ultimate settings that minimizes the Word Error Rate (WER) for the model Table 3.2.

The training process for our model consists of two phases. First, the model was fine-tuned on a larger dataset comprising 66,662 chunks of loosely transcribed audio sourced from Bahraini podcasts, Bahraini TV shows and Bahraini Parliament sessions. The purpose of training the model on a larger

and loosely transcribed dataset is to allow it to capture the diverse linguistic variation of the Bahraini Arabic dialect. Thus, the model will be able to understand the broad speech patterns and contextual variations present in real-world Bahraini audio data. Afterward, the model underwent a second phase of training on a carefully transcribed and clean dataset. This dataset consists of 24,402 chunks of accurately and manually transcribed audio data. The purpose of this phase is to decrease the WER of the model by exposing it to a high-quality and error free transcriptions. By undergoing such sequential training on both datasets, our model was able to transcribe complex Bahraini speech patterns.

An encoder-decoder transformer architecture is used by the OpenAI Whisper model, which is essential to its success in speech recognition tasks. After processing the audio input, the encoder turns the unprocessed waveform into a series of feature vectors that represent the speech's acoustic characteristics. It examines the connections between various input sequence components using self-attention mechanisms. In contrast, the decoder uses these encoded features to produce the text output. In order to align the generated text with the pertinent segments of the input audio, it employs both self-attention on the output sequence and cross-attention to the encoder's output.

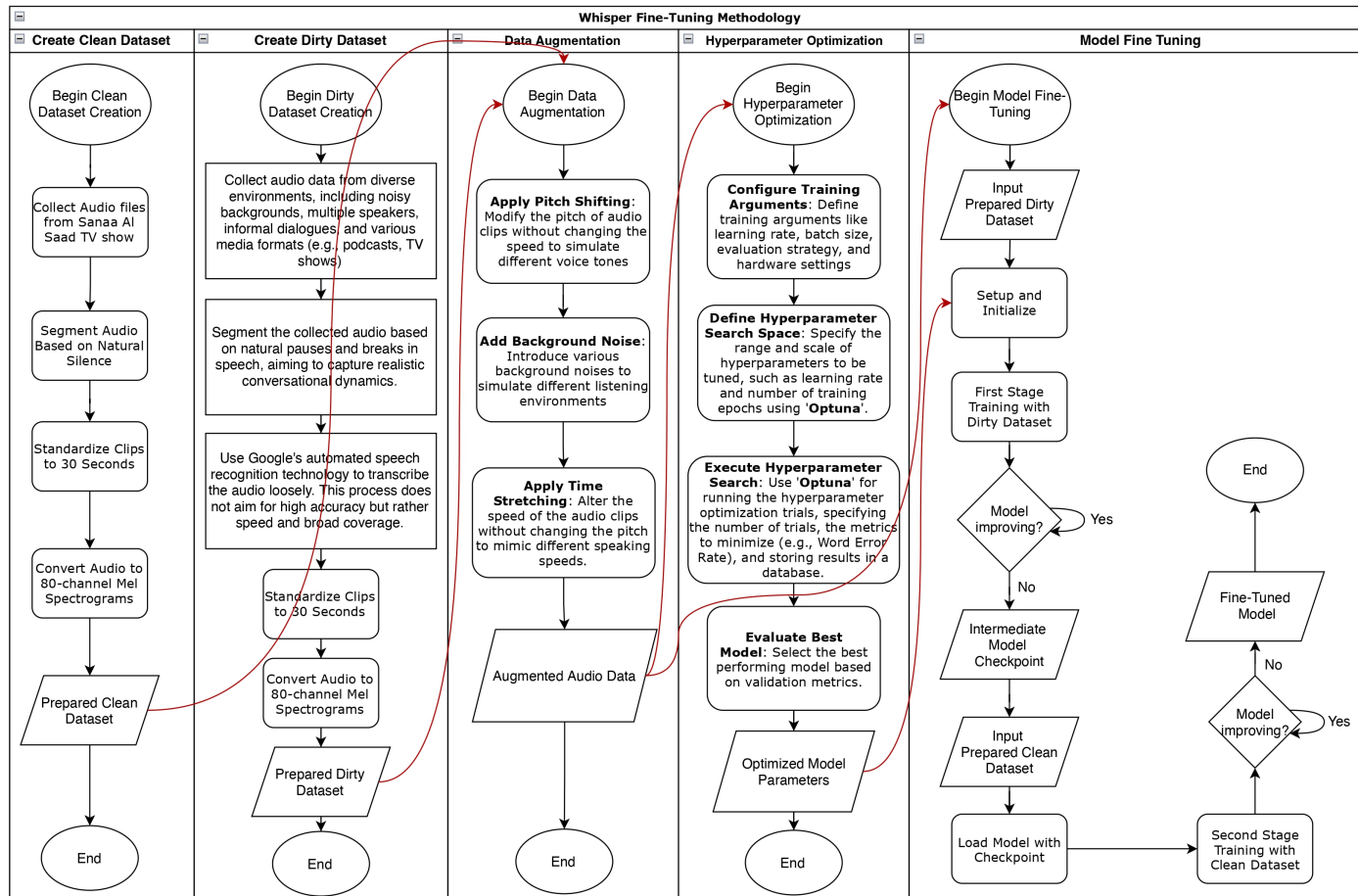


Figure 3.3: Flowchart of the Research Methodology. This diagram outlines the sequential stages of this study

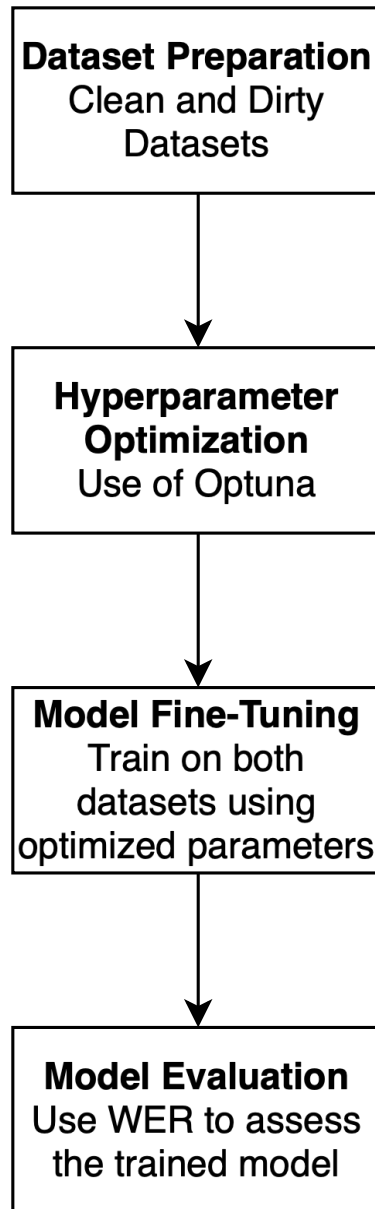


Figure 3.4: An outline of the major steps of the methodology

Chapter 4

Experimental Results

In this chapter, we present the results of the conducted experiment and the improvement of the Whisper model's WER after fine-tuning with the Bahraini dialect. The purpose of the experimental process was to investigate the hypothesis that developing dialect-specific dataset, certain data augmentation techniques and hyperparameter tweaks would lead to a considerable increase in the accuracy and efficiency of the model. A comprehensive summary of the actions taken from the first setup to the model evaluation is given in the next paragraph.

The Benefit Advanced AI Lab's high-performance computing (HPC) resources were used for the experiments detailed in this thesis. Among the hardware configurations is GPU Nodes which is a node that is used for intensive model training and data processing tasks. This consists of two Dell R7525 nodes, each equipped with two Nvidia A100 GPUs and AMD CPUs. This configuration offered a strong infrastructure that could manage the massive computations needed for deep learning models. Benchmarking the cluster's performance revealed that the compute nodes could handle 14.25 TeraFLOPS and the GPU nodes³² could handle 35.25 TeraFLOPS.

To improve learning and accurately test the model's performance, the dataset was selectively partitioned into subsets for the purpose of training and evaluating the Whisper-medium model:

- **Training Set:** The model was trained using 60% of the available data. The volume of samples in this subset was sufficient for the model to pick up on the subtleties of the Bahraini dialect.
- **Validation Set:** 15% of the data was reserved for validation.
- **Testing Set:** After the training and validation stages was over, the remaining 25% was utilized as the testing set, which was only for assessing the model's performance.

This division ensured that the model was tested against unseen data, providing a reliable measure of its real-world efficacy.

The datasets were created as discussed in Figure 3.3 and Figure 3.4. Training of the models is achieved through many epochs with wide training and iteration of hyperparameters in order to reduce the validation loss. The robust evaluation process combines quantitative metrics like WER with the qualitative reviews of the transcriptions that reconfirm the results against baselines. Improvements are then assessed against baseline performances, and statistical tests are confirmed. Training of the models is achieved through many epochs with wide training and iteration of hyperparameters in order to reduce the validation loss. The robust evaluation process combines quantitative metrics like WER with the qualitative reviews of the transcriptions that reconfirm the results against baselines. Improvements are then assessed against baseline performances, and statistical tests are confirmed.

4.0.1 Transfer Learning Strategy

We used a two-phase transfer learning strategy in our training process, going from a larger, poorly transcribed dataset to a smaller, meticulously transcribed one. This strategy has a number of benefits:

1. **Broad Feature Learning:** By using the larger, "noisy" dataset for initial training, the model is able to pick up on a variety of linguistic traits and dialect variations found in the Bahraini language.
2. **Robustness:** The model becomes more resilient to transcription errors and real-world speech variations after being exposed to flawed transcriptions during the initial phase of training.
3. **finely divided Accuracy:** The model's predictions are improved and accuracy on high-quality transcriptions is increased through the subsequent fine-tuning on the clean dataset.
4. **Efficient Use of Limited Data:** This method maximizes the utility of the smaller, high-quality dataset by building upon the broader knowledge gained from the larger dataset.
5. **Better Generalization:** Across a range of Bahraini dialect speech contexts, the combination of extensive exposure and precise tuning may result in improved generalization.

4.1 Experimental Results

Metric	Initial Results	Fine-Tuned Results
Evaluation Loss	3.8142	0.107
Word Error Rate (WER)	181.61%	13.54%
Runtime (seconds)	3902.27	3595.27
Samples Per Second	1.564	1.698
Steps Per Second	0.196	0.213

Table 4.1: Comparison of Initial and Fine-Tuned Whisper Model Performance

Table 4.2 provides an overview of how we initially assessed the Whisper-medium model's performance in terms of fine-tuning. We tested the model

with our manually transcribed Sanaa Al Saad dataset and achieved the following results: The difference between the model's predictions and the actual dataset values is shown by the evaluation loss of 3.8142 that the model achieved. The percentage of incorrectly transcribed words by the model was determined to be 181.61%, which is known as the Word Error Rate (WER). This metric establishes a baseline performance measure for the Whisper-medium model and forms the basis of our subsequent discussions on strategies to fine-tune the model in order to improve its transcription accuracy for the Bahraini dialect. Table 4.1 provides a summary of our initial fine-tuning evaluation of the Whisper-medium model's performance. Using our manually transcribed Sanaa Al Saad dataset, we tested the model and got the following outcomes: The model achieved a 3.8142 evaluation loss, which illustrates the discrepancy between the model's predictions and the actual dataset values. The Word Error Rate (WER), which measures the WER, was found to be 181.61%. This metric provides the Whisper-medium model with a baseline performance measure, and it serves as the foundation for our discussions on how to optimize the model going forward to increase its transcription accuracy for the Bahraini dialect. Following fine-tuning, notable gains were seen in every performance metric, as shown by the data in Table 4.1. The significant decrease in WER and evaluation loss, in particular, demonstrate how successful the optimization techniques used were. These findings highlight how ASR model performance can be greatly improved through fine-tuning, especially when the model is customized to the distinct features of the Bahraini dialect.

Our work of fine-tuning a Whisper model for the Bahraini dialect transcription. Results from this study suggest that our proposed process lead to significantly increased Bahraini transcription accuracy and, hence, proves to be effective in dialect-based voice recognition studies. A fine-tuned model of Whisper continually outperformed its baseline equivalent while offering

Metric	Value
Evaluation Loss	3.8142
Word Error Rate (WER)	181.61%
Runtime (seconds)	3902.27
Samples Per Second	1.564
Steps Per Second	0.196

Table 4.2: Performance Metrics of the Whisper-Medium Base Model

enhanced accuracy for Bahraini dialectal speech transcriptions, as proposed in Figures 4.3 and 4.4 below. Tweaks on hyperparameters and training data bring substantial improvements to the transcription quality. The fine-tuned Whisper model demonstrates improved performance metrics, demonstrating that it can adapt to a wide range of language situations and data complexity.

The normalizing and standardization of outputs, particularly in dialectal speech which is known to have variation of pronunciation, vocabulary, and syntax, which creates challenges in assessing the model. While our analysis uses minor normalization, especially for punctuation and capitalization, more standardization efforts may improve comparability across models and datasets.

Overall, the findings demonstrate the usefulness of fine-tuning procedures in improving the transcription accuracy of Whisper models for dialectal voice recognition tasks. The findings highlight the significance of model adaptation to individual linguistic variances and data features, opening the way for more accurate and contextually relevant speech recognition systems adapted to regional dialects like the Bahraini dialect.

The Optimization History Plot provides vital insights into the Whisper model’s iterative fine-tuning process on the SanaaAlsaad dataset shown in Figure 4. Initial trials (0-5) revealed a rather high Word Error Rate (WER) of about 110, suggesting unsatisfactory performance and the need for additional refining. However, experiment 5 demonstrated a dramatic improve-

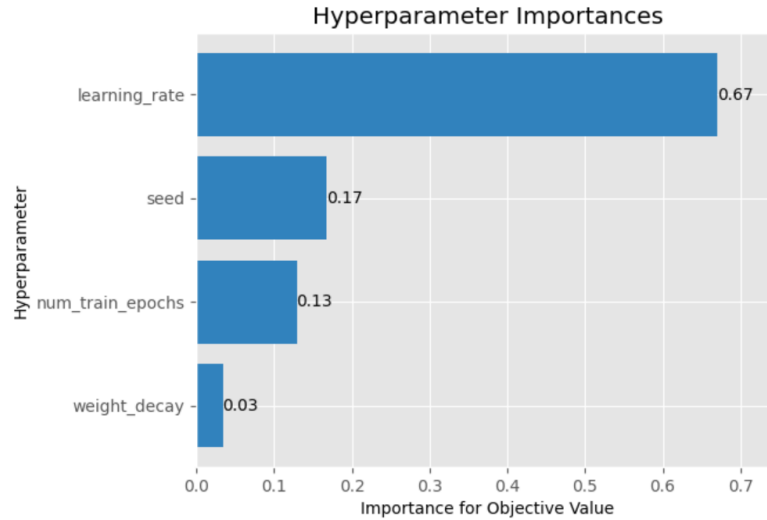


Figure 4.1: Optuna optimization for the Whisper model on the Sanaa Alsaad dataset

ment, with the WER lowering rapidly to around 50, indicating that successful changes were made following the initial trials. Subsequent trials (6-20) showed consistently low WERs, demonstrating steady performance and the efficacy of continued optimization efforts. This iterative optimization strategy, depicted in Figure 4, shows the relevance of focused tweaks in improving transcription accuracy for the Bahraini dialect, underlining the need for extensive optimization procedures to construct strong speech recognition models adapted to various linguistic settings.

Figure 4.1 shows the hyperparameter importance plot, which provides information about the relative significance of several hyperparameters throughout the hyperparameter search using Optuna. The diagram focuses on four important hyperparameters: learning rate, seed, number of training epochs, and weight decay. Among them, learning rate stands out as the most important, with an importance value of 0.67 implying a significant influence on model performance. Following learning rate, the seed hyperparameter has a value of 0.17, which is also a good hyperparameter that has a relatively good influence on the WER. The number of training epochs parameter, with an importance of 0.13, influences training, whereas weight decay, with an

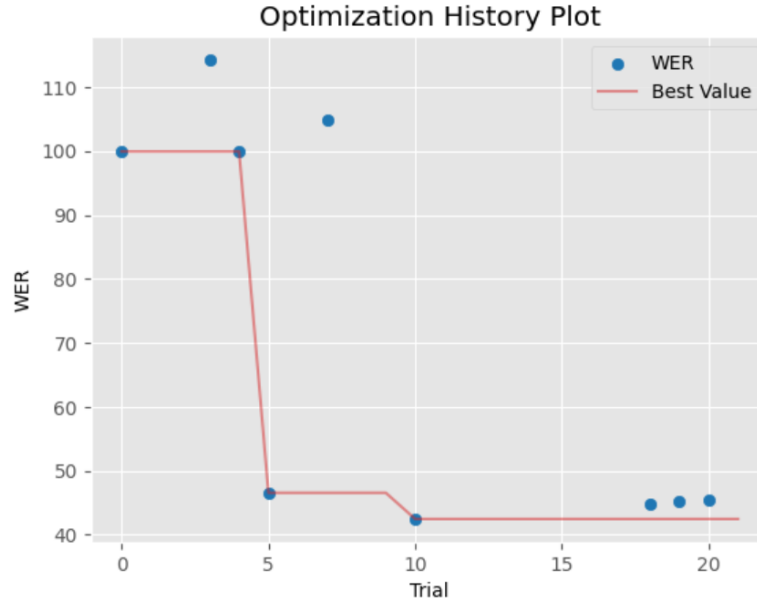


Figure 4.2: WER improvement is notable in Trial 5 during Optuna optimization. Total compute hours is 45.3557 hours

importance of 0.03, has only a modest impact on the fine-tuning process.

Through the epochs, the model is able to continuously decrease the loss as shown in Figure 4.3. That trend would mean that this model is gradually improving its ability to predict target outputs while learning from training data. The reduction in loss is an indication of successful adaptation of the model that learns from the space of features responsible for obtaining high performance. This graph clearly depicts the decreasing error rates during internal parameter tuning and also the effectiveness of the used learning rate during training.

Furthermore, the evaluation on the validation dataset, shown in Figure 4.2, demonstrates the model's generalization capabilities. The downwards trend in WER is a good indicator that the model is not only fitting on the training data but generalizing on new unseen data. The small variations in WER point to difficulties in the generalization of the model, which may be caused by differences in the input data's complexity or by the model meeting uncommon or confusing situations.

Metric	Value
Evaluation Loss	0.107
Word Error Rate (WER)	13.54%
Runtime (seconds)	3595.27
Samples Per Second	1.698
Steps Per Second	0.213

Table 4.3: Performance metrics of the Whisper-medium model after fine-tuning.

Notable performance gains were seen after the Whisper- medium model was fine-tuned using our two datasets Table 4.4. The evaluation loss of the model was reduced significantly to 0.107 from the pre-fine-tuning evaluation loss of 3.814. The Word Error Rate (WER) also showed a remarkable improvement in transcription accuracy, falling from 181.61to 13.54%. With a runtime of 3595.27 seconds, the fine-tuned model showed effective dataset processing; it processed 1.698 samples per second and performed 0.213 steps per second, indicating improved processing and learning speed after fine-tuning.

During the Whisper-medium model’s training phase, the experimental configuration was set up to maximize the model’s performance while maintaining computing efficiency (Liu, Brock, Simonyan, & Le, 2020). The training loss graph shows that the model experienced roughly 1.2 epochs in total. This particular setup was intentionally chosen to use early stopping, which is a method to halt the training process from continuing after a certain number of iterations if no perceivable improvement is shown. This prevents the model from overfitting. This technique protects the model’s capacity to generalize to new data while at the same time, preserves computational resources, which improves the model’s reliability. The validation metrics, which indicate that the model achieved optimal performance without requiring long learning cycles, acknowledge the effectiveness of the training strategy used in this study and support the early training suppression at 1.2

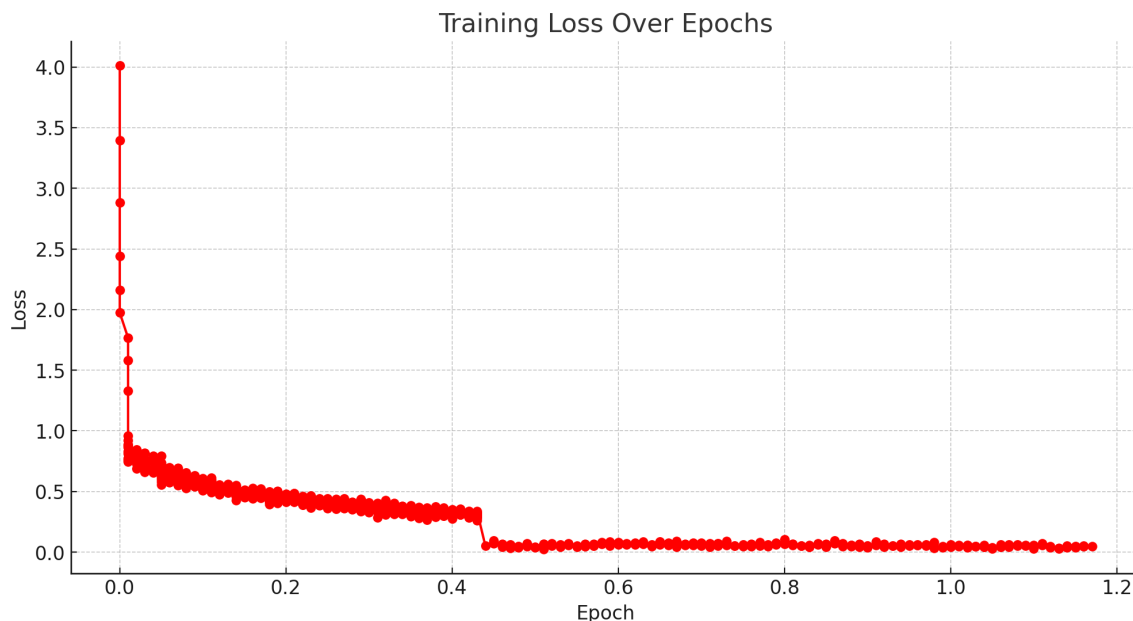


Figure 4.3: The graph shows training loss decreasing over epochs.

epochs.

4.1.1 Performance Metric

The models' performances on the test data can be compared and evaluated by drawing conclusions about the frequency of errors in the models' predictions. A common error metric used to assess speech-to-text and ASR model performance is word error rate (WER) (Santiago, 2023). It compares the recognized words by the model with a reference transcript, then, it calculates the percentage of words that are incorrectly recognized. The calculation formula is depicted in Equation 4.1. A text may contain more than 100% WER but not less than 0%; this is because the variable N represents the total number of words in the transcription that are regarded as true. This means that the computed error rate indicates the portion of the total words in the hypothesis transcription that match the true transcriptions accurately.

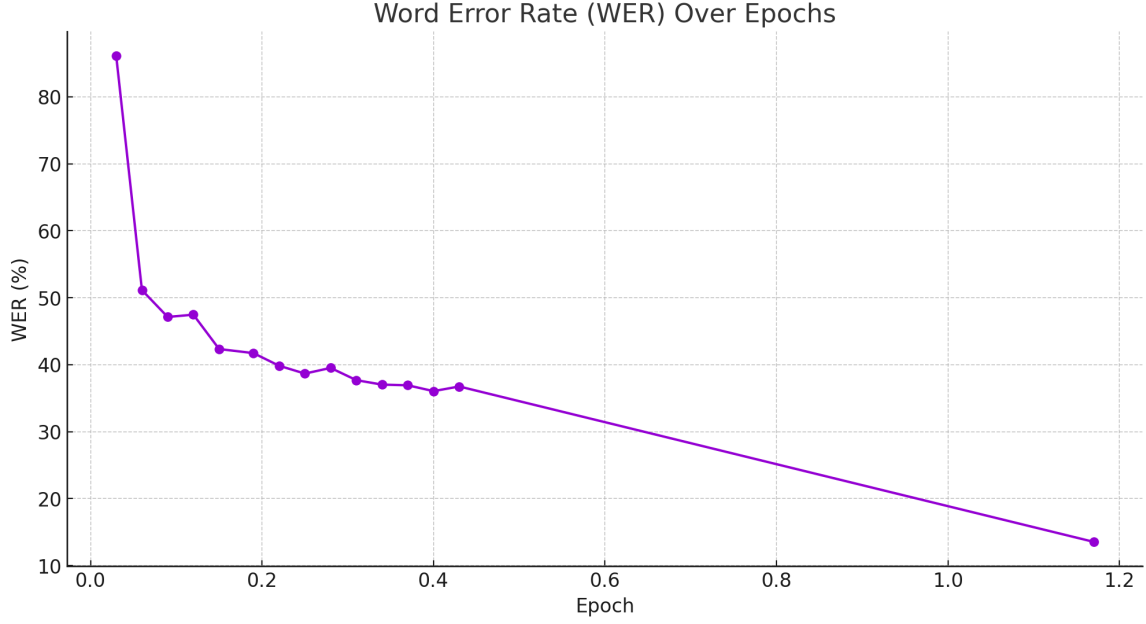


Figure 4.4: The graph represents WER decreasing over epochs, illustrating the model’s improved generalization.

$$\text{WER} = \frac{S + D + I}{N} \quad (4.1)$$

Where:

- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- N is the number of words in the referenced transcript.

4.2 Comparison with Commercial Speech-to-Text Models

We performed a comparative analysis with Google’s Speech-to-Text and Microsoft’s Azure AI Speech services, two of the top commercial ASR so-

lutions, in order to assess the efficacy of our fine-tuned Whisper model. To ensure a fair comparison, we tested all three models using the same Bahraini Arabic dataset.

Model	Word Error Rate (WER)
Fine-tuned Whisper Model	13.54%
Google's Speech-to-Text	22.45%
Azure AI Speech services	52.54%

Table 4.4: Performance Comparison of Fine-tuned Whisper Model and Commercial ASR Solutions

With a Word Error Rate (WER) of 13.54%, our fine-tuned Whisper model outperformed both commercial solutions, as indicated in Table 4.3. On the same dataset, Microsoft's Azure AI Speech services had a WER of 52.54%, whereas Google's Speech-to-Text model recorded a WER of 22.00%.

This comparison shows how well our methodology worked to fine-tune the Whisper model to fit the Bahraini dialect. Considering that these are well-known, commercial-grade solutions, the 8.46 percentage point improvement over Google's model (a relative improvement of about 38.45%) and the 39 percentage point improvement over Azure's model (a relative improvement of about 74.23%) are noteworthy.

The importance of dialect-specific model fine-tuning in ASR tasks is highlighted by this comparison, especially for languages like Arabic that have substantial dialectal variations. It also emphasizes how customized models can perform better in particular linguistic contexts than even well-known, all-purpose solutions.

Chapter 5

Discussion, Limitations, and Conclusion

5.1 Summary

This study focused on improving the accuracy of Automated Speech Recognition (ASR) in transcribing the Bahraini dialect of Arabic, which is a variant that is rarely represented in linguistic research. The primary goal was to fine-tune a Whisper model to better transcribe the Bahraini dialect. We aimed to reduce the Word Error Rate (WER) which will enhance the model's usability and introduce the Bahraini dialect in the natural language processing field.

To achieve the study's goal, we have developed two datasets, one that contained manual transcription of a single Bahraini TV host, and another dataset that contains a collection of Bahraini Arabic audio files that were automatically transcribed using Google's speech to text engine. Then, A variety of machine learning strategies were used to improve the model's performance, particularly for the Bahraini dialect. These strategies included

hyperparameter optimization and sophisticated training procedures.

The results showed a significant improvement in WER, which is an indication that the model has an improved ability to accurately transcribe Bahraini dialect speech. This achievement highlights the possibility of using focused machine learning techniques to address dialectal variance issues in ASR systems. The results point to the possibility of further increasing the accuracy of the model by additionally tweaking the hyperparameters and increase the size and quality of dialect-specific dataset, opening the door to more widespread uses of ASR technology to meet the language requirements of indigenous Arabic speakers.

5.2 Limitations

One of the primary limitations of this study is the size and the accuracy of the dataset available for training and testing the Whisper model. Even though we have produced an extensive dataset specifically designed for the Bahraini dialect, a bigger dataset is required to further improve the model. Dialect-specific ASR research frequently faces the obstacle of finding large-scale, diversified datasets. The model’s generalization across real-world contexts might be improved with larger datasets, which could contain a greater variety of dialectal variants, speaking styles, and contextual usage.

The extent of the fine-tuning and optimization processes that could be carried out depended heavily on the computational resources available. Although a significant amount of computational power (Nvidia A100 GPUs) was used, more computing power might be able to shorten training times, increase the complexity of models that could be tested, and facilitate more thorough hyperparameter tuning. This restriction is especially important considering the high processing requirements of the latest ASR models.

Creating and annotating a high-quality, dialect-specific dataset takes a

significant amount of human labor and expertise. The speed and scope at which the dataset could have been expanded and annotated may have been constrained by the lack of linguists and technicians knowledgeable in the subtleties of the Bahraini dialect. This restriction affects both the volume and quality of the data that the ASR system is trained on.

One major obstacle is the inherent variability of the Bahraini dialect. The challenge of achieving higher transcription accuracy is made more difficult by elements like code-switching, regional accents, and informal versus formal speech, which can differ significantly between speakers and situations.

Word Error Rate (WER) is the primary metric used to assess the ASR system's performance; however, this may not adequately reflect the model's efficacy in practical settings. The semantic accuracy of the transcriptions, which can be essential for understanding user intent and speech context, is not taken into consideration by WER.

5.3 Recommendations and Future Work

We have established a strong basis for the development of ASR systems for the Bahraini language. Extending our dataset is essential to this advancement, as it will improve the understanding and accuracy of our model by integrating a wider range of dialectal expressions, accents, and pronunciation variants from a diverse demography.

Continued optimization of the model is crucial. To guide our advancements, we suggest comparing our ASR system with other cutting-edge models. Further investigating hyperparameters, model topologies, and advanced regularization strategies will support this initiative. Furthermore, the implementation of context-aware models that dynamically adapt to the speaker's surroundings holds significant potential for improving transcription accu-

racy in a variety of real-world scenarios.

5.4 Conclusion

This thesis has investigated how the OpenAI Whisper model can be used to enhance the accuracy of speech-to-text transcriptions for the Bahraini dialect. The study has brought attention to the need for precise, high-quality data that perfectly represents the relevant dialect. A significant factor in the creation of local datasets was the significance placed on the importance of having a systematic and comprehensive methodology for data collection procedures in order to appropriately training of models.

A strong technique for building reliable models for a variety of difficult and demanding real-world audio situations is data augmentation. The model's capacity to generalize beyond the training data was significantly enhanced by operations including noise injection, speed variation, and pitch shifting, which also made the model more accurate in a variety of speech settings.

Hyperparameter tuning is crucial for optimizing the model's performance, which is made possible by cutting-edge optimization tools like Optuna. Adjusting parameters is important in machine learning as it allows for precise adjustments to be made to variables like learning rate, training epoch count, and weight decay. This precise adjustment of hyperparameters is essential, as it directly influences the optimal settings for the training process.

In conclusion, this thesis supports the notion that careful hyperparameter optimization, creative data augmentation techniques, and thorough data handling work together can create an accurate ASR system. By investigating more intricate models, larger accurate datasets, and more advanced optimization algorithms, future research can build on these foundations and push the limits of automated speech recognition systems, especially for un-

derrepresented languages and dialects.

References

- Abdou, S. M., & Moussa, A. M. (2019). Arabic speech recognition: Challenges and state of the art. *Computational linguistics, speech and image processing for arabic language*, 1–27.
- Ahadi, S., Sheikhzadeh, H., Brennan, R., & Freeman, G. H. (2003). An efficient front-end for automatic speech recognition. *10th IEEE International Conference on Electronics, Circuits and Systems, 2003. ICECS 2003. Proceedings of the 2003, 1*, 128–131 Vol.1. doi: 10.1109/ICECS.2003.1301993
- Albirini, A. (2011). *The structure and functions of codeswitching between standard arabic and dialectal arabic* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.
- Alex, A., Wang, L., Gastaldo, P., & Cavallaro, A. (2023). Data augmentation for speech separation. *Speech Communication*, 152, 102949.
- Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., ... Almojl, M. (2021). Automatic speech recognition: Systematic literature review. *IEEE Access*, 9, 131858–131876. doi: 10.1109/ACCESS.2021.3112535
- Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., ... Renals, S. (2015). Automatic dialect detection in arabic broadcast speech. *arXiv preprint arXiv:1509.06928*.
- Ali, A., Khalifa, S., & Habash, N. (2019). Towards variability resistant dialectal speech evaluation. In *Interspeech* (pp. 336–340).

- Ali, H., Zhou, X., & Tie, S. (2013). Comparison of mfcc and dwt features for automatic speech recognition of urdu. In *International conference on cyberspace technology (cct 2013)* (pp. 154–158).
- Alsaad, S. (2023). *Sana al saad - bahrain*. <https://www.youtube.com/@sanaalsaad-bahrain6489>. (Accessed: 2023-05-03)
- Arakawa, R., Takamichi, S., & Saruwatari, H. (2019). Implementation of dnn-based real-time voice conversion and its improvements by audio data augmentation and mask-shaped device. *Proc. SSW10*, 93–98.
- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., . . . Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33, 12449–12460.
- BahrainTV. (2018). . <https://www.youtube.com/playlist?list=PL3TTD7BCTQfiF,H>
- BahrainTV. (2019). . <https://www.youtube.com/watch?v=U-dE9Qml33Mlist=PL3TTD7BCTQfiF,H>
- Charola, M., Kachhi, A., & Patil, H. A. (2023). Whisper encoder features for infant cry classification. In *Proc. interspeech* (Vol. 2023, pp. 1773–1777).
- Cechi, R. R. (2013). Performance analysis of mfcc and lpcc techniques in automatic speech recognition. *International journal of engineering research and technology*, 2.
- Chowdhury, S. A., Hussein, A., Abdelali, A., & Ali, A. (2021). Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. *ArXiv, abs/2105.14779*. doi: 10.21437/interspeech.2021-1809
- Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recog-

- nition. *arXiv preprint arXiv:2006.13979*.
- Dubey, P., & Shah, B. (2022). Deep speech based end-to-end automated speech recognition (asr) for indian-english accents. *ArXiv, abs/2204.00977*. doi: 10.48550/arXiv.2204.00977
- Elmahdy, M., Hasegawa-Johnson, M., & Mustafawi, E. (2014). Development of a tv broadcasts speech recognition system for qatari arabic. In *Lrec* (Vol. 14, pp. 3057–3061).
- Espy-Wilson, C. (2005). Automatic speech recognition. *Journal of the Acoustical Society of America*, 117, 2403-2403. doi: 10.1121/1.4786105
- Gold, B., Morgan, N., & Ellis, D. (2011). Automatic speech recognition. , 299-300. doi: 10.1002/9781118142882.PART6
- Gris, L., Marcacini, R., Júnior, A. C., Casanova, E., Soares, A., & Alu'isio, S. (2023). Evaluating openai's whisper asr for punctuation prediction and topic modeling of life histories of the museum of the person. *ArXiv, abs/2305.14580*. doi: 10.48550/arXiv.2305.14580
- Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... others (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.
- Hassan, M. K. A., Rehmat, A., Khan, M. U. G., & Yousaf, M. (2022). Improvement in automatic speech recognition of south asian accent using transfer learning of deepspeech2. *Mathematical Problems in Engineering*. doi: 10.1155/2022/6825555
- Helander, M., Moody, T. S., & Joost, M. (1988). Systems design for automated speech recognition. , 301-319. doi: 10.1016/B978-0-444-70536-5.50019-1
- Hirschberg, J., & Biadys, F. (2011). Automatic dialect and accent recognition and its application to speech recognition. doi: 10.7916/D8M61S68

- Huang, Z., Xing, H., & Liu, M. (n.d.). Adapter integration: Mitigating catastrophic forgetting in multi-language and multi-accent whisper asr model fine-tuning.
- Karlos, S., Fazakis, N., Karanikola, K., Kotsiantis, S., & Sgarbas, K. (2016). Speech recognition combining mfccs and image features. In *Speech and computer: 18th international conference, specom 2016, budapest, hungary, august 23-27, 2016, proceedings 18* (pp. 651–658).
- Khandelwal, K., Jyothi, P., Awasthi, A., & Sarawagi, S. (2020). Black-box adaptation of asr for accented speech. , 1281-1285. doi: 10.21437/interspeech.2020-3162
- Khurana, S., & Ali, A. (2016). Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge. In *2016 ieee spoken language technology workshop (slt)* (pp. 292–298).
- Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., ... Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. *arXiv preprint arXiv:1904.03288*.
- Liu, H., Brock, A., Simonyan, K., & Le, Q. (2020). Evolving normalization-activation layers. *Advances in Neural Information Processing Systems*, 33, 13539–13550.
- Martucci, G., Cettolo, M., Negri, M., & Turchi, M. (2021). Lexical modeling of asr errors for robust speech translation. , 2282-2286. doi: 10.21437/interspeech.2021-265
- Mohan, J., & Babu, R. (2014). Speech recognition using mfcc and dtw. *2014 International Conference on Advances in Electrical Engineering (ICAEE)*, 1-4. doi: 10.1109/ICAEE.2014.6838564
- Mulfari, D., & Villari, M. (2024). A voice user interface on the edge for people with speech impairments. *Electronics*, 13(7), 1389.

- Oh, C., Kim, C., & Park, K. (2023). Building robust korean speech recognition model by fine-tuning large pretrained model*. *Phonetics and Speech Sciences*. doi: 10.13064/ksss.2023.15.3.075
- O'Shaughnessy, D. (2008). Invited paper: Automatic speech recognition: History, methods and challenges. *Pattern Recognit.*, 41, 2965-2979. doi: 10.1016/j.patcog.2008.05.008
- Padmanabhan, J., & Johnson, M. (2015). Machine learning in automatic speech recognition: A survey. *IETE Technical Review*, 32, 240 - 251. doi: 10.1080/02564602.2015.1010611
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5206–5210).
- pod.mishkah. (2023). . <https://www.youtube.com/@pod.mishkah/videos>. (Accessed: 2023-05-03)
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International conference on machine learning* (pp. 28492–28518).
- Ranjeet, P., Prakash, T., Amruta, S., & Monali, S. (2016). Automatic speech recognition system. *Imperial journal of interdisciplinary research*, 2, 165-169.
- Santiago, G. (2023). *A quantitative comparative study of automatic speech recognition (asr) systems using english and lojban languages in the health care industry* (Unpublished doctoral dissertation). Capella University.
- Vu, N. T., Wang, Y., Klose, M., Mihaylova, Z., & Schultz, T. (2014). Improving asr performance on non-native speech using multilingual and crosslingual information. , 11-15. doi: 10.21437/Interspeech.2014-3

- Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., & Fung, P. (2020). Learning fast adaptation on cross-accented speech recognition. , 1276-1280. doi: 10.21437/interspeech.2020-0045
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... others (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations* (pp. 38–45).
- Wray, S., & Ali, A. (2015). Crowdsourcing a little to label a lot: labeling a speech corpus of dialectal arabic. In *Interspeech* (pp. 2824–2828).
- Yadavalli, A., Mirishkar, G. S., & Vuppala, A. (2022). Multi-task end-to-end model for telugu dialect and speech recognition. , 1387-1391. doi: 10.21437/interspeech.2022-10739
- Zhang, C., Wu, J., Xiao, X., & Wang, Z. (2006). Pronunciation variation modeling for mandarin with accent. doi: 10.21437/Interspeech.2006-246
- Zhou, G., Chen, Y., & Chien, C. (2022). On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks. *BMC Medical Informatics and Decision Making*, 22(1), 226.

ملخص الدراسة

اللهجة البحرينية هي من اللهجات التي لم تحظ بالكثير من الاهتمام في الأبحاث المتعلقة بمعالجة اللغات الطبيعية (NLP). تهدف هذه الأطروحة على تطوير نظام لتحويل الكلام المسموع إلى كتابة مخصص بدقة للهجة البحرينية. تم اختيار نموذج Whisper الذي تم تطويره من قبل شركة OpenAI لهذه الدراسة. بالإضافة، تم إنشاء مجموعتين من البيانات الصوتية التي سيتم استخدامها لعمل التحسين الدقيق (Fine-Tuning) لنموذج Whisper.

في بداية هذه الدراسة، قمنا بتطوير مجموعتين من البيانات الصوتية: المجموعة الأولى تحتوي على تسجيلات للمذيعات سناء السعد، هذه التسجيلات قد تم اختيارها بعناية وتمت كتابة التسميات التوضيحية المغلفة يدوياً لضمان أعلى دقة ممكنة. المجموعة الثانية تشمل تسجيلات أكثر تنوعاً مثل حلقات البودكاست، جلسات برلمانية وبرامج تلفزيونية مع الأخذ بعين الاعتبار أن تكون جميع هذه المصادر بحرينية. تم استخدام هذه المجموعات لتدريب نموذج Whisper حيث أظهرت النتائج تحسناً كبيراً في الدقة وإنخفاض في معدل الخطأ في التعرف على الكلمات من 181.608% إلى 13.54%، مما يدل على فعالية النموذج المحسن في التعرف على اللهجة البحرينية.

تسهم هذه الأطروحة بشكل كبير في مجال التعرف الآلي على الكلام، خاصة في تعديل التقنيات الحالية لتناسب مع لهجات غير شائعة مثل اللهجة البحرينية. تمهد هذه الدراسة الطريق لاستخدامات أوسع في البحوث اللغوية وتطبيقات التكنولوجيا في مجال معالجة اللغات الطبيعية، إلى جانب معالجة التحديات التقنية المحددة. بالإضافة إلى ذلك، يمكن للنموذج المحسن أن يكون له تأثير واسع النطاق في توفير أدوات تسهيل الوصول، وتكنولوجيا التعليم، وتحسين خدمة العملاء في البحرين ومناطق أخرى مماثلة.

جامعة البحرين

كلية تقنية المعلومات



نحو التعرف التلقائي للكلام
إلى نص أكثر دقة لهجة البحرينية
أطروحة مقدمة كجزء من متطلبات الحصول على
درجة الماجستير في تعلم الآلة والذكاء الحسائي

إعداد

عبدالله باسل ال محمود

٢٠٢١٠٠٢٢٢

إشراف

ا.د. هشام العمال

(أستاذ مساعد)

د. فاطمة البلوشي

(أستاذ مساعد)

جامعة البحرين

مملكة البحرين

ابريل ٢٠٢٤