

UNIVERSITY OF BAHRAIN

College of Information Technology



An Investigation in Bahraini Dialect Text to Speech Synthesis Models and Datasets

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

Submitted By

Mohamed Almahmood

202100223

Supervised By

Dr. Fatema Albalooshi

(Assistant Professor)

Co-Supervisor

Dr. Hesham M. Al-Ammal

(Assistant Professor)

University of Bahrain

May 2024



Dissertation Deposit Form

Author Name:	Mohamed Basel Almahmood		
Email:	202100223@stu.uob.edu.bh	Contact No:	32205455
College:	Information Technology		
Department:	Computer Science		
Thesis Title:	An Investigation in Bahraini Dialect Text to Speech Synthesis Models and Datasets		
Year:	2024		

Author's Declaration: I agree to the following conditions:

The above thesis will be available (in the University of Bahrain library and externally) and reproduced as necessary at the discretion of the University of Bahrain. It may also be digitized by the University of Bahrain and made available via the university's repository or via other parties on the Internet.

This condition shall apply to ALL copies including electronic copies.

The above thesis has been provided on the understanding that it is copyright protected material and that no quotation from this thesis may be published without proper acknowledgement.

Signature of the Author:	Mohamed Basel Almahmood	Date:	05-May-24
--------------------------	-------------------------	-------	-----------

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:  Date: 01-May-2024

Mohamed Basil Ebrahim Al-Mahmood (Candidate)

The Thesis defense committee considers the thesis titled:

An Investigation in Bahraini Dialect Text to Speech Synthesis Models and Datasets

Submitted by **Mohamed 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. Fatema AlBalooshi

(Supervisor)

Department of Computer Science
University of Bahrain

Dr. Hesham Al Ammal

(Co-Supervisor)

Department of Computer Engineering
University of Bahrain

ABSTRACT

In the past few years, there has been an accelerated development in the field of Artificial Intelligence (AI) and Machine Learning (ML). Various Text To Speech (TTS) modules have been developed, and successfully implemented in many languages. However, implementations of such systems for the Arabic language has been minimal, with implementations mostly relying on makeshift solutions such as using diacritics to transliterate the Arabic text to English. This study presents the development of a Text To Speech (TTS) system for the Bahraini dialect. The proposed approach uses the Forward Tacotron model integrated with HiFi-GAN vocoder, and Optuna for hyperparameter optimization. Our approach provides a thorough understanding of the requirements for synthesizing natural and expressive Bahraini Arabic speech, and investigates the possibility of training such a system without the use of diacritics or phonetization tools. For the purpose of this study we created a dataset in the Bahraini dialect, containing 9 hours of recordings collected from Sanaa Alsaad's YouTube channel. However, due to the small dataset size, data augmentation techniques were used expand the dataset to 36 hours. Another version of the dataset was created spanning 8 hours of augmented data and involving only higher quality recordings. The smaller dataset was then AI enhanced using Adobe Podcasts. The sample rate of the recorded data is 22050 Hz, as per the model's requirements. Our system was first trained on the 36 hours dataset, and then transfer learning was used to fine tune the training on the 8 hours dataset. The system was assessed using metrics such as Attention Score, Mel Loss, Duration Loss, Pitch Loss, Word Error Rate (WER), and Mean Opinion Score (MOS). The model demonstrated promising capabilities in replicating the local dialect with an MOS of 3.4, and a WER of 34.

Contents

APPROVAL	A
DECLARATION	B
ABSTRACT	C
TABLE OF CONTENTS	D
LIST OF FIGURES	G
LIST OF TABLES	I
ACKNOWLEDGMENTS	J
DEDICATION	K
LIST OF ABBREVIATIONS	L
1 Introduction	1
1.1 Overview	1
1.2 Research Problem	1
1.3 Research Objectives	2
1.4 Contribution to the Field	2
1.5 Thesis Outline	4
2 Literature Review	5
2.1 State of The Art TTS Models	6
2.1.1 Tacotron	6
2.1.2 Tacotron 2	7
2.1.3 Fast-Speech	8
2.1.4 One TTS Alignment To Rule Them All	9

2.1.5	Forward Tacotron	11
2.1.6	HifiGan	12
2.2	Hyperparameter Optimization	13
2.2.1	Optuna	13
2.3	Arabic TTS Systems	14
2.4	Non-Arabic TTS Systems	17
3	An Investigation in Bahraini Dialect Text to Speech Synthesis	
	Models and Datasets	20
3.1	Methodology	20
3.1.1	The Dataset	21
3.1.1.1	Data Collection	21
3.1.1.2	Data Processing	22
3.1.1.3	Data Transcription	22
3.1.1.4	Data Augmentation	22
3.1.1.5	Data Enhancement	23
3.1.2	The Speech Synthesis Model	24
4	Experimental Results	27
4.1	Performance Metrics	27
4.1.1	Attention Score	28
4.1.2	CTC Loss	28
4.1.3	DIA Loss	28
4.1.4	Mel Loss	29
4.1.5	Duration Loss	29
4.1.6	Energy Loss	30
4.1.7	Pitch Loss	30
4.1.8	Word Error Rate	30
4.1.9	Mean Opinion Score	31
4.2	Results	32

4.2.1	Attention Score	32
4.2.2	CTC Loss	33
4.2.3	DIA Loss	34
4.2.4	Mel Loss	35
4.2.5	Duration Loss	36
4.2.6	Pitch Loss	37
4.2.7	Energy Loss	38
4.2.8	Word Error Rate	39
4.2.9	Mean Opinion Score	40
5	Discussion, Challenges, and Conclusion	41
5.1	Discussion	41
5.2	Challenges	42
5.3	Conclusion	45
	References	47
	Abstract (in Arabic)	

List of Figures

2.1	The architecture of the Tacotron model (Wang et al., 2017) .	7
2.2	The architecture of the Tacotron 2 model (Shen et al., 2017)	8
2.3	The architecture of the FastSpeech model (Ren et al., 2019) .	9
2.4	The CER of different modules with and without the alignment framework (Badlani et al., 2021)	10
2.5	The architecture of Forward Tacotron (Schafer, 2023)	12
3.1	A simplified view of the dataset creation process	24
3.2	A simplified view of the model’s training process	26
4.1	The Attention Score using the original dataset vs the attention score with transfer learning using the enhanced dataset.	32
4.2	The CTC Loss of the Tacotron step using the original dataset vs the CTC Loss with transfer learning using the enhanced dataset.	34
4.3	The DIA Loss of the Tacotron step using the original dataset vs the DIA Loss with transfer learning using the enhanced dataset.	35
4.4	The Mel Loss of the Tacotron step using the original dataset vs the Mel Loss with transfer learning using the enhanced dataset.	36

4.5	The Mel Loss of the Forward step using the original dataset vs the Mel Loss with transfer learning using the enhanced dataset.	36
4.6	The Duration Loss of the Forward step using the original dataset vs the Duration Loss with transfer learning using the enhanced dataset.	37
4.7	The Pitch Loss of the Forward step using the original dataset vs the Pitch Loss with transfer learning using the enhanced dataset.	38
4.8	The Energy Loss of the Forward step using the original dataset vs the Energy Loss with transfer learning using the enhanced dataset.	39

List of Tables

2.1	Comparison of the MOS and speed of the top performing publicly accessible models (Kong, Kim, & Bae, 2020) . . .	12
2.2	A comparison of hyperparameter optimization frameworks (Akiba, Sano, Yanase, Ohta, & Koyama, 2019)	13
2.3	A summary of datasets used by the discussed Arabic TTS systems.	17
2.4	The models and results of the Arabic TTS systems.	17
2.5	The models and results of the non-Arabic TTS systems. . . .	19
4.1	The WER of the current TTS system	40
4.2	The MOS of the current TTS system	40

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Fatema Albalooshi and my Co-Supervisor, Dr. Hesham Al-Ammal for their great support, help, and motivation. I also express my gratitude to Dr. Dana Abdulla, Dr. Latifa Shamsan from the English department, and the team of volunteers for helping transcribe the dataset. Their guidance helped me in all my research and during the writing of this thesis. Benefit Bahrain Company and Benefit AI Lab at University of Bahrain.

DEDICATION

To my parents for their encouragements.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ML	Machine Learning
TTS	Text To Speech
RNN	Recurrent-Neural-Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
MSA	Modern Standard Arabic
ACR	Absolute Category Rating
CER	Character Error Rate
WER	Word Error Rate
MOS	Mean Opinion Score
CTC	Connectionist Temporal Classification
DIA	Diagonal Loss
SR	Success Rate
MOS	Mean Opinion Score

Chapter 1

Introduction

1.1 Overview

The context of this thesis is to investigate the training of a Text To Speech (TTS) system in the Bahraini dialect. TTS algorithms have made significant advancements in recent years due to the development of deep learning techniques and innovative architectures. Several modules such as Tacotron, Tacotron 2, Fast-Speech, HifiGan, and Forward-Tacotron have contributed to the field, demonstrating remarkable improvements in naturalness, expressiveness, and speaker adaptation. Such systems have rarely been implemented for the Arabic language, and no implementations exist for the Bahraini dialect. This research will investigate the possibility of developing a TTS system using modern TTS algorithms with the help of public datasets and other Machine Learning (ML) tools.

1.2 Research Problem

Most Arabic TTS systems are developed using diacritized text, phonetization tools, and transliteration from Arabic to English. Furthermore, there are few studies focusing on TTS systems for specific Arabic dialects, and there is a lack of datasets for the Bahraini dialect. In addition, no existing

Arabic studies have employed the Forward Tacotron model. Moreover, the effectiveness of Optuna in optimizing TTS systems and the effectiveness of data augmentation for such systems have not been thoroughly explored.

1.3 Research Objectives

The Objectives of the work in our thesis can be summarized as follows:

1. To develop a dataset in the Bahraini dialect.
2. To develop a TTS system using Forward Tacotron with minimal data, and without the use of diacritics, transliteration, and phonetization tools.
3. To adapt the Forward Tacotron code to implement Optuna for hyper-parameter optimization.
4. To explore the effectiveness of data augmentation on TTS systems.
5. To explore the effectiveness of using high quality recordings vs low quality recordings on TTS system.
6. To explore the effectiveness of enhancing poorly recorded audio on a TTS system.
7. To explore the effectiveness of transfer learning on Forward Tacotron.

1.4 Contribution to the Field

First, developing a dataset for the Bahraini dialect. This involves gathering audio recordings from Sanaa Alsaad's YouTube channel, pre-processing it, and transcribing the collected audio recordings. This is the first Arabic dataset in the Bahraini dialect, and such a dataset is critical for training a TTS model that aim to reproduce Bahraini Arabic.

Second, is to create a TTS system using the Forward Tacotron model, with minimal data and without the use of diacritics, transliteration, or phonetization tools. This approach allows the model to train on the Arabic alphabet natively and derive the correct pronunciations without the use of diacritics just like native human speakers, potentially allowing for a more straightforward and robust system.

Third, is to modify the Forward Tacotron code to integrate Optuna for hyperparameter optimization. By implementing Optuna, the system can automatically optimize each layer of the model and find the best parameters for our specific application. This could lead to improved speech synthesis quality and efficiency.

Fourth, is to investigate the effectiveness of data augmentation techniques on TTS systems. Data augmentation can help improve a TTS system when data is sparse, and when acquiring more data is not feasible. The data can be expanded by modifying various aspects of the audio such as speed and pitch.

Fifth, is to assess the impact of using high-quality versus low-quality recordings in a TTS system. This can give us insights into the practical applications of TTS technology, and helps us decide the viability of developing datasets using sources not intended for TTS applications.

Sixth, is to assess the impact of enhancing poorly recorded audio on a TTS system. Specifically, trimming silences, volume normalization, and audio quality enhancement using AI tools.

Finally, is to assess the effectiveness of applying transfer learning techniques to the Forward Tacotron model. Mainly, training on a dirty dataset, then using transfer learning to continue training on a cleaner, AI enhanced dataset.

1.5 Thesis Outline

This thesis is structured as follows. In chapter 2, the literature review of Arabic and English TTS systems are discussed. Chapter 3 introduces the components of the Forward Tacotron model, the process of collecting the dataset, along with the process of training the model. Furthermore, chapter 3 shows and discusses the results of the study, along with the challenges faced while doing the study. Chapter 4 provides a summary of the study and presents recommendations and suggestions for future work.

Chapter 2

Literature Review

Modern Standard Arabic (MSA) and dialectal Arabic are two different forms of the Arabic language, each serving a specific purpose in the Arabic world. MSA is the standardized and formal version of Arabic that is used in formal communication, speeches, and media. It is standard and does not change through all regions of the Arab world. On the other hand, dialectal Arabic varies significantly from region to region in the Arab world, and it is used in daily informal communication. These dialects are significantly different in terms of pronunciation, vocabulary, and grammar.

Arabic dialect transcription presents many difficulties especially with the dialect spoken in Bahrain. For example, there are phonemes that do not have a direct equivalent in Arabic writing like the "ch" sound. Furthermore, there is no widely recognized transcription technique, which means that different people may write the same sentence in different ways. In addition, in Bahrain, there are over 10 distinct dialects spoken in different cities and villages, each with its own unique characteristics. The speaker chosen for

our dataset speaks in a combination of Riffai and Muharraqi dialects.

2.1 State of The Art TTS Models

TTS algorithms have made significant advancements in recent years due to the development of deep learning techniques, and innovative architectures. Several studies such as Tacotron (Wang et al., 2017), Tacotron 2 (Shen et al., 2017), Fast-Speech (Ren et al., 2019), and Forward-Tacotron (Schäfer, 2023) have contributed to the field, demonstrating remarkable improvements in naturalness, expressiveness, and speaker adaptation.

2.1.1 Tacotron

Tacotron is an auto-regressive TTS model produced by Google in 2016 [1]. It takes text as input, then converts it to a set of linguistic features using techniques such as text normalization, text tokenization, vocabulary creation, token embedding, linguistic feature extraction, sequence padding and length normalization. It uses an encoder, an attention mechanism, and a decoder as part of its framework. The encoder is a Recurrent-Neural-Network (RNN) that takes the linguistic features and learns to extract information about the words, their relationships, and the overall context. An attention mechanism assigns weights to different parts of the encoder's output. The decoder uses those weights to focus on the corresponding encoded text during speech synthesis. The system essentially works by aligning the decoder's generated mel-spectrogram with the encoder's target mel-spectrogram using the attention mechanism. Figure 2.1 shows Tacotron's architecture in which characters are given as input and raw spectrogram is generated as output. The raw spectrogram is then fed to Griffin-Lim vocoder to generate the corresponding wav file (Wang et al., 2017).

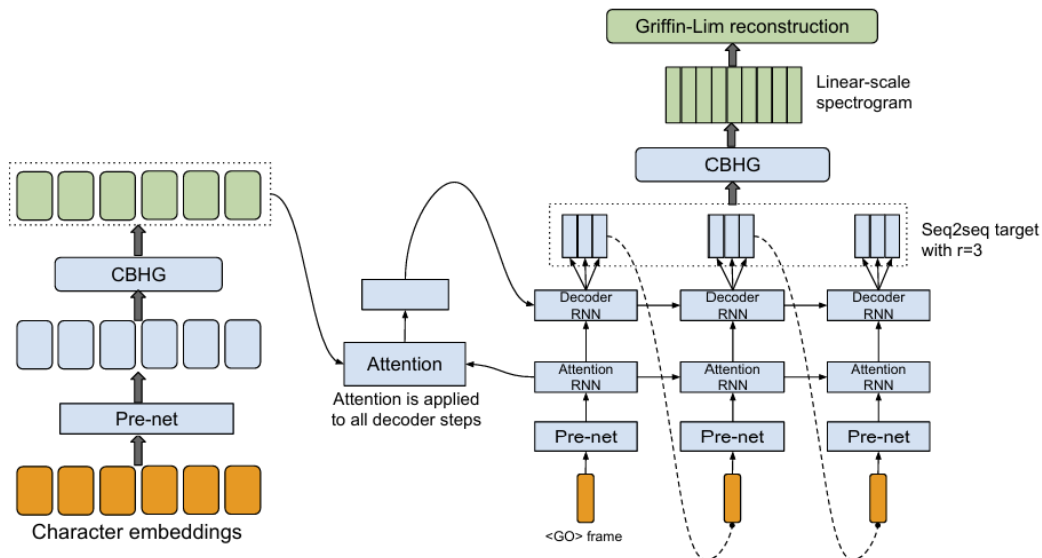


Figure 2.1: The architecture of the Tacotron model (Wang et al., 2017)

2.1.2 Tacotron 2

Tacotron 2 was developed by Google in 2017 as an attempt to overcome some of its predecessor’s drawbacks [2]. The original Tacotron uses a combination of convolutional layers and Gated Recurrent Units (GRU) recurrent layers, while Tacotron 2 utilizes a stack of convolutional layers followed by a single bi-directional Long Short-Term Memory (LSTM) layer [2]. The convolutional layers on Tacotron 2 helps the model learn important features and patterns, while the LSTM layer being bi-directional, can capture information from both past and future contexts, improving the model’s understanding of the input sequence. This makes Tacotron 2 potentially faster to train and run. Tacotron 1 uses a reduction factor, meaning each decoder step corresponds to multiple spectrogram frames [1]. This increases the speech smoothness but decreases the speech clarity. Tacotron 2 takes a different approach, with each decoder step generating a single spectrogram frame allowing the model to better capture fine-grained details of the speech and potentially making speech clearer [2]. Tacotron 1 utilizes additive atten-

tion, while Tacotron 2 uses location-sensitive attention. Additive attention computes the attention weights by combining the previous attention context and the current decoder state, while location-sensitive attention, takes into account the position of the attention window within the spectrogram. This allows Tacotron 2 to align the attention more accurately with the input text during the decoding process, and potentially improving speech quality. Although Tacotron 2 improved on the original Tacotron, it still suffers from some of its main drawbacks. Figure 2.2 shows the Tacotron 2 architecture in which characters are given as input and raw spectrogram is generated as output. The raw spectrogram is then fed to WaveNet vocoder to generate the corresponding wav file. (Shen et al., 2017)

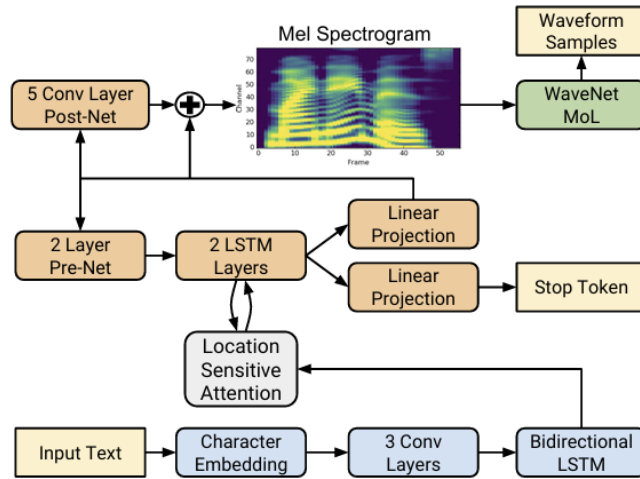


Figure 2.2: The architecture of the Tacotron 2 model (Shen et al., 2017)

2.1.3 Fast-Speech

Fast Speech is a non-auto-regressive model that offers a faster and a more efficient solution than the Tacotron models. Unlike Tacotron which uses an encoder-decoder attention mechanism to generate mel-spectrograms sequentially, Fast-Speech utilizes a feed-forward Transformer architecture to generate mel-spectrograms in parallel, making it significantly faster than Tacotron [3]. Furthermore, Fast-Speech incorporates a phoneme duration

predictor. The duration predictor processes each phoneme individually and predicts the duration of each phoneme in terms of the number of frames or time steps required to generate that phoneme. This prediction happens before the acoustic frames are generated. This allows for simultaneous generation of all acoustic frames, resulting in faster inference compared to Tacotron’s sequential generation. In Tacotron, the acoustic frames are generated one-by-one, conditioned on the previously generated frames, which can be computationally slower. Fast-Speech’s approach reduces common Tacotron issues like word skipping and repeating. This alignment mechanism enhances the overall quality and robustness of the synthesized speech. Figure 2.3 shows the architecture of the FastSpeech model consisting of 4 main steps. (a) Feed Forward Transformer, (b) FFT Block, (c) Length Regulator, (d) Duration Predictor, MSE loss stands for the difference between the actual duration vs the predicted duration. (Ren et al., 2019).

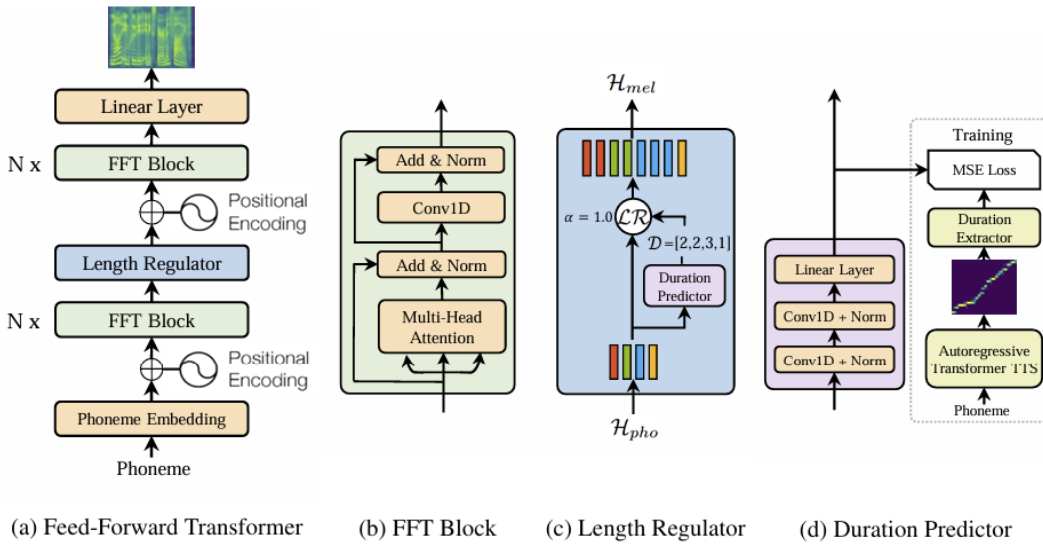


Figure 2.3: The architecture of the FastSpeech model (Ren et al., 2019)

2.1.4 One TTS Alignment To Rule Them All

Tacotron’s attention mechanism can have issues aligning speech with text, especially with long utterances resulting in missing or repeating words [1][2],

while Fast-Speech relies on phoneme duration prediction [3]. Nvidia’s One TTS Alignment To Rule Them All framework introduces an alignment learning framework that can be used by both auto-regressive and non-auto-regressive modules. It builds on their mechanism, enhancing the alignment convergence speed, simplifying the training pipelines, and improving the speech synthesis quality [4]. The framework leverages unsupervised alignment learning objectives and static alignment priors to accelerate the learning process. It eliminates the need for forced aligners or alignments obtained from pre-trained models which are expensive to use and often not readily available for certain languages. Results demonstrate improved alignment convergence, stability on long sequences, and overall speech quality based on human evaluation. Figure 2.4 shows the character error rate of different modules with and without the alignment framework. Models using the framework make fewer mistakes and can generate longer sentences. (Badlani et al., 2021).

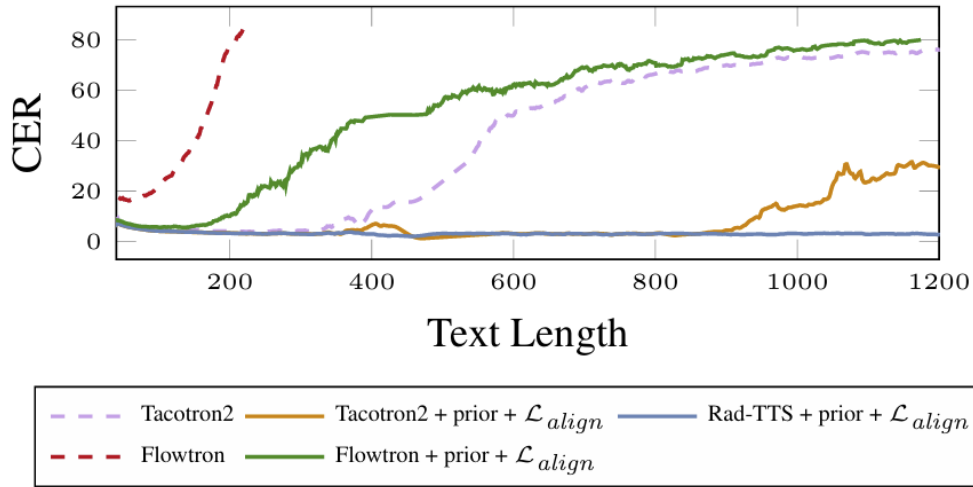


Figure 2.4: The CER of different modules with and without the alignment framework (Badlani et al., 2021)

2.1.5 Forward Tacotron

Forward Tacotron is a non auto-regressive encoder decoder TTS model that incorporates an attention mechanism. It modifies the attention mechanism of Tacotron by using a forward attention approach called Teacher-Forced Attention. This type of attention uses the ground truth mel spectrogram from the previous time step instead of relying on the decoder’s predictions to compute the attention score, making the attention mechanism more robust and accurate. In addition, Forward Tacotron uses a combination of convolutional and recurrent layers to generate mel spectrogram frames and duration values, which are then combined to form the final output audio signal. The duration prediction component helps determine the length of time that each mel spectrogram frame should be played for. Moreover, Forward Tacotron incorporates alignment conditioning based on the ”One TTS Alignment To Rule Them All” framework which improves alignment convergence, stability and speech quality. Thus, Forward Tacotron combines some of the positive aspects of Tacotron, Fast Speech and ”One Alignment To Rule Them All” frameworks. Figure 2.5 shows the architecture of the Forward Tacotron model. The duration predictor aligns text input with the generated mel spectrograms (Schäfer, 2023).

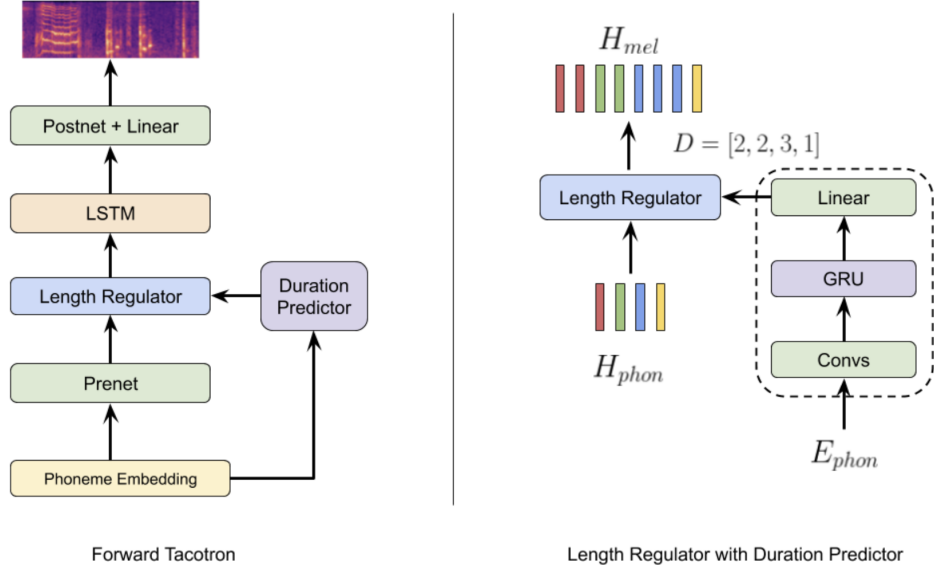


Figure 2.5: The architecture of Forward Tacotron (Schafer, 2023)

2.1.6 HifiGan

HifiGan is a vocoder developed to synthesis high-fidelity speech sounds using inputs from mel-spectrograms. With the use of Generative Adversarial Networks, HiFiGAN generates audio that sounds realistic and natural, in a fast and efficient manner. This makes it especially useful for speech synthesis applications where realism and intelligibility are essential. Table 2.1 shows the model generating superior speech than the top-performing publicly accessible models (Kong, Kim, & Bae, 2020).

Model	MOS	Speed on GPU (kHz)
Ground Truth	4.45	-
WaveNet	4.02	0.07
WaveGlow	3.81	501
MelGAN	3.79	14,238
HifiGAN V3	4.05	26,169

Table 2.1: Comparison of the MOS and speed of the top performing publicly accessible models (Kong, Kim, & Bae, 2020)

2.2 Hyperparameter Optimization

2.2.1 Optuna

Optuna is a hyperparameter optimization framework that can speed up the process of finding the optimal hyperparameters for a given module and dataset. Manual optimization of these parameters would be very tedious and time-consuming due to their vast number and the complexity. Each parameter has a significant influence on the model’s performance, and manually determining the ideal collection of parameters is impractical. Optuna automates this process by using advanced algorithms to select the appropriate parameters more effectively than standard grid or random search approaches. It employs techniques like as Bayesian optimization, gradient methods, and evolutionary algorithms to intelligently explore the parameter space, prioritizing combinations that are more likely to provide better outcomes. This intelligent search feature shortens the time and processing resources required to identify the optimal hyper-parameters. Table 2.2 shows a comparison of hyperparameter optimization frameworks with Optuna being superior than all other frameworks (Akiba, Sano, Yanase, Ohta, & Koyama, 2019).

Framework	API Style	Pruning	Lightweight	Distributed	Dashboard	OSS
SMAC [3]	define-and-run	✗	✓	✗	✗	✓
GPyOpt	define-and-run	✗	✓	✗	✗	✓
Spearmint [2]	define-and-run	✗	✓	✓	✗	✓
Hyperopt [1]	define-and-run	✗	✓	✓	✗	✓
Autotune [4]	define-and-run	✓	✗	✓	✓	✗
Vizier [5]	define-and-run	✓	✗	✓	✓	✗
Katib	define-and-run	✓	✗	✓	✓	✓
Tune [7]	define-and-run	✓	✗	✓	✓	✓
Optuna (this work)	define-by-run	✓	✓	✓	✓	✓

Table 2.2: A comparison of hyperparameter optimization frameworks (Akiba, Sano, Yanase, Ohta, & Koyama, 2019) .

2.3 Arabic TTS Systems

Our work of constructing a dataset in the Bahraini dialect and training the Forward Tacotron model on such a dataset without the use of diacritics and phonetization tools, along with the use of Optuna for hyperparameter optimization and data augmentation techniques, have never been explored before, as demonstrated by the following studies:

A study by (Sawalha & Al-Radhi, 2023) was conducted to improve the naturalness of neural-based text-to-speech systems that were trained on a small dataset. To efficiently achieve high-quality speech synthesis, the researchers used the FastSpeech2 model in conjunction with three vocoders: Ahocoder Vocoder, Continuous Vocoder, and World Vocoder (Sawalha & Al-Radhi, 2023). They used a dataset of about 3.7 hours in Modern Standard Arabic (MSA) that had been phoneme-aligned to incorporate the Arabic language into the FastSpeech2 model (Sawalha & Al-Radhi, 2023). The model's parameters were modified by the researchers to preserve natural and high-quality speech synthesis while matching different speech characteristics, such as speaking tone and speed. An Absolute Category Rating (ACR) scale was used to evaluate the study with 0 being the worst score and 5 being the best score. The results of the study show that the Continuous Vocoder performed the best, with Ahocoder Vocoder being a close second, while World Vocoder performing the worst. In addition, the training results of the algorithm showed a Mel Loss of 0.473, a Pitch Loss of 0.331, and an Energy Loss of 0.08. The results show that the synthesized speech using FastSpeech2 was not comparable to the original sound.

Hanane & Hamadouche (2023) developed a system of five agents working together to synthesize speech, each having a specific goal. The study used Nawar Halabi's Arabic dataset, consisting of 2.41 hours of studio recorded audio. Furthermore, the dataset contains diacritics and is written in MSA. The model was tested by synthesizing random sentences from

the speech corpus. The sentences were evaluated by 10 participants where the number of well pronounced phrases is divided by the number of tested phrases to get the success rate of the model. The results show a success rate of 86%.

Azab and his colleagues (2023) developed a TTS system for Egyptian Arabic using Tacotron2 and the HiFi-GAN Vocoder (Azab, Zaky, Ogawa, & Gomaa, 2023). The dataset named EGYARA-23, consisted of 20.5 hours recordings by a male speaker narrating news and general writings. Furthermore, it was divided into 32,716 audio files with an average duration of 8 seconds. The dataset was transcribed in Egyptian Arabic to reflect the actual dialect, and each segment was accompanied by its respective transcript. The dataset included diacritics to accurately represent spoken Arabic, and the preprocessing phase used Buckwalter transliteration for converting text to phonetic representations. The results of the study were evaluated using Character Error Rate (CER), Word Error Rate (WER), and Mean Opinion Score (MOS). The study achieved a CER of 7.3, a WER of 22.3, and a MOS of 4.48, indicating the system's performance in close proximity to English and Modern Standard Arabic systems.

In their work (Hamdy, Magdy, Alfawzy, Ghaly, & Abbas, 2021) developed an Arabic TTS system using Tacotron 1 paired with Griffin-Lim vocoder and Tacotron 2 paired with Wavenet vocoder. The study used Nawar Halabi's diacritised MSA dataset. The researchers evaluated the models using 20 participants to calculate the MOS. The results show a MOS of 4.01 for Tacotron 1, and an MOS of 4.38 for Tacotron 2.

Fahmy, Abbas, & Khalil (2022) developed a Transfer Learning TTS system in Modern Standard Arabic. The researchers used a pre-trained Tacotron2 model in the English language with Waveglow Vocoder. Furthermore, they used Nawar Halabi's Arabic Dataset, consisting of 2.41 hours of Arabic speech and diacritised text in MSA. The researchers were able

to use transfer learning by transforming diacritic Arabic words into English characters using Buckwalter Arabic Morphological Analyzer (Fahmy, Abbas, & Khalil, 2022). The study was successfully able to generate speech from diacritised Arabic input with an MOS of 4.21 (Fahmy et al., 2022).

In their work (Abdelali et al., 2022), the authors introduced an end-to-end Text-to-Speech (TTS) system for MSA, named the NatiQ system. They developed their own dataset consisting of high quality studio recordings of a male voice and a female voice with its corresponding diacritised text (Abdelali et al., 2022). The female voice named Amina consisted of 3,964 audio files, each audio file having an average duration of 7 seconds and totaling 50,714 words. The male voice named Hamza consisted of 6,005 segments, each audio file having an average duration of 7 seconds and totaling 80,409 words. The female voice is expressive, while the male voice is monotonic. The study used WaveRNN Vocoder with Tacotron1, WaveGlow Vocoder with Tacotron2, and Parallel WaveGan Vocoder with ESPnet. The results for Amina showed a WER of 40.42 for ESPnet2, 43.98 for Tacotron 1, and 64.8 for Tacotron 2. For Hamza, the results showed a WER of 24.87 for ESPnet2, 46.12 for Tacotron 1, and 93.62 for Tacotron 2. Regarding the MOS, Amina achieved a score of 3.57 for ESPnet2, 4.21 for Tacotron 1, and 3.49 for Tacotron 2. Hamza achieved a MOS of 4.40 for ESPnet2, 4.38 for Tacotron 1, and 2.34 for Tacotron 2. This shows that some models excel at handling expressive voices, while others excel at handling monotone voices.

In conclusion, there are not many recent studies available on Arabic text-to-speech (TTS) systems. In order to improve voice synthesis, a large number of these studies make use of phonetization and diacritics to transliterate Arabic text into English. Most of these studies focus on Modern Standard Arabic, a formal variant that is commonly used in literature and media. Most of the current experiments use common TTS models such as Tacotron and Fast Speech. These experiments have had mixed results, despite the state-of-

the-art technology; some have good results with low word error rates, while others have higher inaccuracy with a large word error rate. This illustrates the variability and difficulties in creating efficient Arabic TTS systems. Table 2.3 shows the dialect used in the discussed studies and whether diacritics and phonetization tools were used or not. Table 2.4 shows the models used in these studies along with the results.

Study Name	Dataset Dialect	Diacritics Used?
Sawalha & Al-Radhi, 2023	MSA	Yes
Hanane & Hamadouche, 2022	MSA	Yes
Azab, Zaky, Ogawa, & Gomaa, 2023	Egyptian Arabic	Yes
Hamdy, Magdy, Alfawzy, Ghaly, & Abbas, 2021	MSA	Yes
Fahmy, Abbas, & Khalil, 2022	MSA	Yes
Abdelali et al., 2022	MSA	Yes

Table 2.3: A summary of datasets used by the discussed Arabic TTS systems.

Study Name	Model Name	Results
Sawalha & Al-Radhi, 2023	FastSpeech2	Mel Loss: 0.473
Hanane & Hamadouche, 2022	5 Custom Agents	SR: 86%
Azab, Zaky, Ogawa, & Gomaa, 2023	Tacotron 2	WER: 22.3
Hamdy, Magdy, Alfawzy, Ghaly, & Abbas, 2021)	Tacotron 1	MOS: 4.01
	Tacotron 2	MOS: 4.38
Fahmy, Abbas, & Khalil, 2022	Tacotron 2	MOS: 4.21
Abdelali et al., 2022	Tacotron 1	WER: 46.12
	Tacotron 2	WER: 93.62
	ESPnet	WER: 24.87

Table 2.4: The models and results of the Arabic TTS systems.

2.4 Non-Arabic TTS Systems

A study by (Aziz, Tiun, & Ruslan, 2023) aimed to create a TTS system in the Malay language using Tacotron and Tacotron2. The researchers created

a dataset in the Malay language using a single male speaker and totaling 14.29 hours of speech. For pre-processing the dataset underwent normalization such as converting numbers to words, lowercasing, tokenization, and character embedding. The study found that the Tacotron model requires a smaller batch size compared to Tacotron 2. The study found that Tacotron synthesized speech more naturally and intelligibly than Tacotron2. However, both models did not reach the naturalness and intelligibility levels of human speech with an MOS of 3.8 for Tacotron and 3.12 for Tacotron2.

In a study by (Tran, 2021), the researchers developed a TTS system for the Vietnamese language based on Tacotron2. They used a publicly available dataset containing 30 hours of recorded data. The researchers used Mozilla's implementation of Tacotron2, and found that the synthesized speech was fairly clear and natural but not on par to human speech. with an MOS of 3.

The researchers (Gopalakrishnan, Imam, & Aggarwal, 2022) conducted a study to train a TTS system on a small dataset in a low resource environment. The researchers compared the performance of Tacotron2, FastSpeech2 and DeepVoice3 using a subset of the LJSpeech dataset and using the Tesla T4 GPU. The study found that Tacotron2 produced the most natural speech with an MOS of 4.25.

A study by (Koçak & Büyükzincir, 2023) aimed to develop a TTS system in the Turkish language. The researchers developed their own dataset spanning 63 hours of studio recorded audio using a professional speaker with 28 years of work experience. The study used the Tacotron and the FastSpeech2 models for training and found that the FastSpeech2 model scored an MOS of 4.1, while the Tacotron model scored an MOS of 3.72.

To conclude, studies that got a lower MOS seem to be studies with inferior datasets. For example (Tran, 2021) used a publicly available dataset with a variety of recording qualities and got a low MOS. While, (Koçak &

Büyükcincir, 2023) used 63 hours of professionally recorded audio and got a near human speech synthesis system. Refer to table 2.5 for a summary of the discussed research papers.

Study Name	Model Name	Results
Aziz, Tiun, & Ruslan, 2023	Tacotron	MOS: 3.8
	Tacotron 2	MOS: 3.12
Tran, 2021	Tacotron 2	MOS: 3
Gopalakrishnan, Imam, & Aggarwal, 2022	Tacotron 2	MOS: 4.25
Koc ak & B uy ukzincir, 2023	FastSpeech2	MOS: 4.1
	Tacotron	MOS: 3.72

Table 2.5: The models and results of the non-Arabic TTS systems.

Chapter 3

An Investigation in Bahraini Dialect Text to Speech Synthesis Models and Datasets

3.1 Methodology

Pre-trained TTS models are readily available for the English language due to the simplicity of the language, and the availability of high quality datasets. However, pre-trained models on Arabic are very few, and such models are usually trained on standard Arabic and use diacritics. Furthermore, there is no dataset available in the Bahraini dialect. In this study, we created a limited dataset in the Bahraini dialect without the use of diacritics and investigated the possibility of training Forward Tacotron on such a dataset. Furthermore, we used Optuna for hyper-parameter optimization in order to minimize the error that can be caused by using inefficient hyper-parameters. Moreover, we investigated whether data augmentation helps improve a TTS system when data is sparse.

The experiment was conducted on the High Performance Computing facilities at the Benefit Advanced AI Lab. The hardware comprised of two Dell R640 master nodes operating synchronously. Furthermore, eight Dell R640 nodes equipped with Nvidia Tesla T4 GPUs. Moreover, two Dell R7525 nodes having two Nvidia A100 GPUs and two AMD CPUs each. The performance of the cluster was benchmarked at 14.25 TeraFLOPS for the compute nodes and 35.25 TeraFLOPS for the GPU nodes. Such a fast and efficient setup was crucial in training the model and maintaining results in such a limited time.

This section is divided into two subsections. The first subsection discusses the process of creating the dataset used in this study. Figure 3.1 shows a simplistic view of the dataset creation process. The second subsection discusses the modifications done to the Forward Tacotron model as well as the training process. Figure 3.2 shows a simplistic view of the module training process.

3.1.1 The Dataset

Sanaa Alsaad is a famous Bahraini media personality that has produced a lot of shows and radio programs. Her Youtube channel (SanaaAlsaad, 2017) has many of her content uploaded. We chose to create the dataset using her YouTube content due to her authentic Bahraini dialect and the availability of the large amount of audio files recorded in only her voice. This section will discuss the process of creating the dataset and its features.

3.1.1.1 Data Collection

Python's *youtube-dl* library was used to extract and download the audio files from SanaaAlsaad's YouTube channel with a sampling frequency of 44.1 KHz, and in wav format. A total of 80 videos were processed, amounting to

9 hours of recorded audio.

3.1.1.2 Data Processing

The audio files were processed using *htdemucs* Python library to remove background music and isolate the speaker’s voice. Each audio file was analyzed and split using Audacity into 1 to 10 seconds segments based on the flow of the sentence being spoken. This resulted in 12,209 segments. The sample rate was then converted from 44100 Hz to 22050 Hz, as per the requirements of the model.

3.1.1.3 Data Transcription

The audio files were manually transcribed in the Bahraini dialect. Departing from conventional practices, diacritics were deliberately omitted from the transcription process. This departure was based on the fact that diacritics are rarely used in real-life written communication, and human readers can comprehend text without their inclusion. However, this decision made it impossible to use phonetization tools for converting the transcribed text into phonetic representations, as these tools typically rely on transliterating diacritised Arabic text into English. Such tools serve as a makeshift solution that may compromise accuracy and fidelity.

3.1.1.4 Data Augmentation

Most of the common datasets used for TTS training are vast and professionally recorded. For example, Forward Tacotron provides pretrained models on LJ Speech Dataset and Thorstenmueller Dataset. Both the LJ Speech Dataset and the Thorstenmueller Dataset comprise of about 24 hours of clean single speaker recordings that are segmented to 1 to 10 seconds clips, and are recorded 22050 Hz. Another common dataset is the Kokoro Speech Dataset comprising around 59 hours of professionally recorded data from

a single speaker. Thus, the dataset developed in this study needs to be expanded to a minimum of 24 hours, especially with the complexity of the Arabic language compared to other languages. Since acquiring more data is not feasible, data augmentation techniques were utilised to expand the dataset.

The segments acquired in the previous steps were processed using Audacity to triple the dataset's size. This involved altering the pitch and speed of the audio files. The initial subset was generated by accelerating the audio files by 1 percent while adjusting the pitch correspondingly. Subsequently, a second subset was formed by accelerating the audio files by 3 percent, with corresponding pitch adjustments. Finally, a third subset was derived by decelerating the audio files by 5 percent, accompanied by pitch modifications. These adjustments subtly altered the waveforms of the audio files, resulting in an expanded dataset totaling 36 hours.

3.1.1.5 Data Enhancement

Obtaining audio from YouTube videos for research purposes has revealed a range of issues related to quality. Notably, differences in microphone proximity, where in some cases the speaker is placed far from the mic, while others the speaker is close. This causes disparities that have a direct impact on audio clarity. This variety is reflected in the recordings as well, which range in quality from subpar to satisfactory. Another layer of complexity is added by the speed at which the speaker presents her information. In multiple cases the speaker speaks too fast resulting in mispronunciations and even comprehension problems. These findings emphasize how complex it is to use YouTube audio as a resource and how strong methods are needed to reduce inconsistent quality and guarantee accurate data extraction.

In order to address this problem, a rigorous screening procedure was put in place to identify recordings in which the speaker speaks clearly. After

that, all of the audio’s beginning and ending silences were eliminated, along with prolonged silences within the recording. In addition, Adobe Podcast was used to significantly enhance the audio quality through AI processing. Thus, a different version of the dataset with 2 hours of high quality recordings was produced. The dataset was then expanded using the previously discussed augmentation techniques, resulting in eight hours of recordings in total. Figure 3.1 shows a simplified view of the dataset creation process.

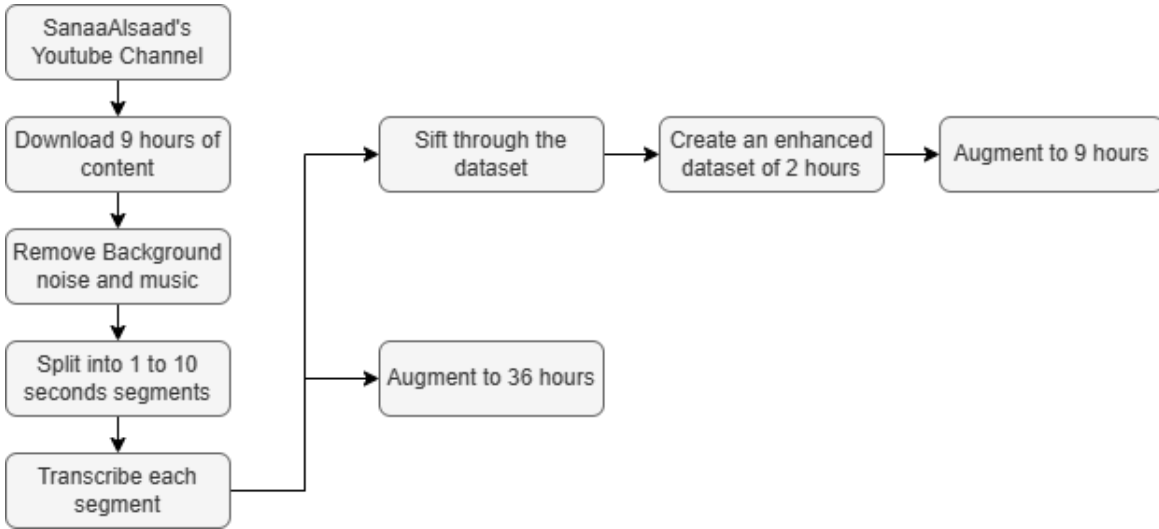


Figure 3.1: A simplified view of the dataset creation process

3.1.2 The Speech Synthesis Model

The Forward Tacotron system comprises three integral sub-modules: the Tacotron module, the Forward Tacotron module, and the HifiGan Vocoder module. To enhance the efficiency and performance of these modules, Optuna was employed for hyperparameter optimization which significantly increased the model’s capabilities. The code for the Tacotron and Forward Tacotron modules was modified to properly integrate Optuna, enabling rigorous parameter adjustment across layers. Each layer in each module was optimized along with parameters like learning rate, r value, and batch size. For the Tacotron module, the optimization goal was to decrease the vali-

validation mel-loss. This metric served as a crucial indicator of the module’s proficiency in generating high-quality mel-spectrograms from text inputs. In contrast, the optimization strategy for the Forward Tacotron module was more extensive, addressing multiple objectives simultaneously. These include minimizing the validation mel-loss, the validation duration loss and the validation pitch loss. Incorporating these additional objectives aimed to enhance the overall quality and naturalness of the synthesized speech output.

The model’s documentation recommended using *espeak-ng* for the text-to-phoneme conversion process, highlighting its support for the Arabic language. However, the error rate for the module was extremely high due to the use of *espeak-ng* with non-diacritized text. Thus, the module’s code was modified to discard the *espeak-ng* implementation.

In the original implementation, the code utilized the griffin-lim Vocoder to transform synthesized spectrograms into audible signals. However, this approach yielded subpar results, characterized by a robotic-sounding voice and poorly articulated speech. Given that the module generates the requisite mel files for compatibility with alternative vocoders, the state-of-the-art HiFi-GAN vocoder was implemented to replace the default vocoder, utilizing a pretrained universal module. Figure 3.2 shows a simplified view of the model’s training process.

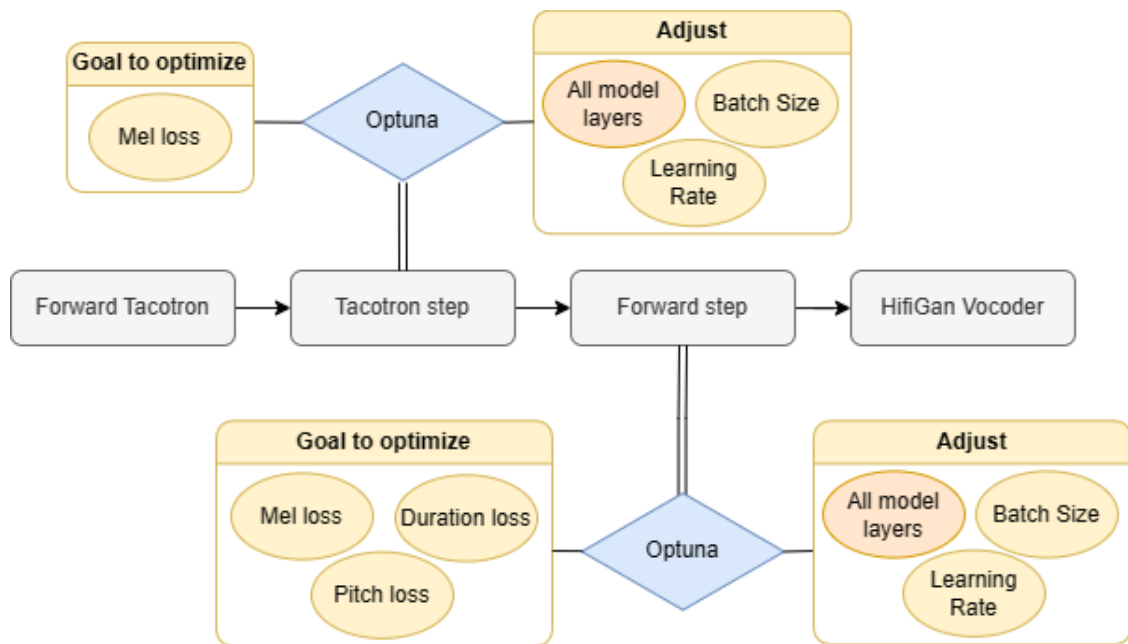


Figure 3.2: A simplified view of the model's training process

Chapter 4

Experimental Results

TensorBoard was used to monitor the training process and access the evaluation metrics. It is visualization tool developed by Google as part of the TensorFlow ecosystem. It makes machine learning workflows easier to comprehend. Furthermore, it makes it easier to compare various runs and assess the efficacy of model modifications by offering a clear graphical depiction of model runs. As discussed previously, the system was trained on two datasets consecutively, the original 36 hours augmented dataset, and an 8 hours enhanced and augmented dataset containing. The original dataset was split into 47429 audio files for training, and 400 audio files for testing. The enhanced dataset was split into 3506 audio files for training, and 400 audio files for testing.

4.1 Performance Metrics

The evaluation metrics for the initial phase of training the Tacotron speech synthesis model involve several key measures. These include: Attention

Score, Connectionist Temporal Classification (CTC) loss, Durational Intonation-Aware (DIA) Loss, and Mel Loss. In the second stage of training, which involves the Forward Tacotron module, the evaluation includes: Mel Loss, Duration Loss, Energy Loss, and Pitch Loss. These metrics are explained below.

4.1.1 Attention Score

This metric measures how well the model’s attention mechanism is focusing on relevant parts of the input data to predict the output correctly. The score may be calculated using equation 4.1 where Q represents the query metric, K is the key metric, V represents the value metric, d_k is the dimensionality of the keys and queries, softmax is applied to convert the scores into probabilities (Henriquez, 2020).

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4.1)$$

4.1.2 CTC Loss

This metric measures the alignment between the predicted and actual sequences without requiring a pre-defined alignment between them. It may be calculated using equation 4.2 where y is the label, X is the split spectrogram, p is the probability (Yeo, 2020).

$$\text{CTC Loss} = -\log(p(y|X)) \quad (4.2)$$

4.1.3 DIA Loss

The Diagonal Loss helps guide the attention mechanism of the neural network. It is calculated to encourage the attention matrix of the model to focus near the diagonal, making the alignment between the text and audio output

more accurate. It is calculated using equation 4.3 where `dia_mat` represents the guided attention matrix, `attention` is the actual attention matrix generated by the model, `*` is the element-wise multiplication, and `.mean()` computes the mean of all elements in the resultant matrix. This equation is derived from the source code found in (Schäfer, 2023).

$$\text{DIA Loss} = ((1 - \text{dia_mat}) * \text{attention}).\text{mean}() \quad (4.3)$$

4.1.4 Mel Loss

A Mel spectrogram is a representation of the audio’s frequency spectrum over time. This metric measures the error between the predicted Mel spectrogram and the target Mel spectrogram. It may be calculated using equation 4.4 where $\hat{m}1$ is the initial predicted mel-spectrogram, $\hat{m}2$ is the post-processed mel-spectrogram after additional refining (postnet output), $L1_loss$ represents the L1 loss function, and mel is the target mel-spectrogram from the dataset. This equation is derived from the source code found in (Schäfer, 2023).

$$\text{Mel Loss} = L1_loss(\hat{m}1, mel) + L1_loss(\hat{m}2, mel) \quad (4.4)$$

4.1.5 Duration Loss

This measures how accurately the model predicts the duration of each phoneme or sound unit in the speech, which is critical for natural-sounding speech synthesis. It may be calculated using equation 4.5 where $\hat{d}_i$ is the predicted duration, d_i is the actual duration of the i_{th} phoneme, and N is the total number of phonemes in the batch. This equation is derived from the source code found in (Schäfer, 2023).

$$\text{DurLoss} = \frac{1}{N} \sum_{i=1}^N |\hat{d}_i - d_i| \quad (4.5)$$

4.1.6 Energy Loss

This measures the model’s ability to predict the energy or loudness of the speech, impacting the expressiveness of the synthesized voice. It may be calculated equation 4.6 where, $\hat{E}_n$ is the predicted energy value at frame n , and E_n is the target energy values at frame n , and N is the total number of phonemes in the batch. This equation is derived from the source code found in (Schäfer, 2023),

$$\text{Energy Loss} = \frac{1}{N} \sum_{n=1}^N |(\hat{E}_n - E_n)| \quad (4.6)$$

4.1.7 Pitch Loss

This measures how accurately the model predicts the pitch patterns of the speech, which are essential for conveying emotions and intentions in spoken language. It may be calculated using equation 4.7 where $\hat{P}_n$ is the predicted pitch value at frame n , and P_n is the target pitch value at frame n , and N is the total number of phonemes in the batch. This equation is derived from the source code found in (Schäfer, 2023).

$$\text{Pitch Loss} = \frac{1}{N} \sum_{n=1}^N |\hat{P}_n - P_n| \quad (4.7)$$

4.1.8 Word Error Rate

One popular statistic for assessing Text-to-Speech (TTS) system performance is Word Error Rate (WER). By comparing the spoken output with

the accurate text transcript, it calculates the amount of errors in the synthesized speech. The whole amount of insertions, deletions, and substitutions needed to convert the synthesized speech into the reference text is divided by the reference's word count to determine WER. This measure is presented as a percentage, and a lower WER means greater accuracy and, hence, a better TTS system. WER is an essential technique for evaluating intelligibility and the naturalness of synthesized speech because it directly shows a TTS system's ability to replicate precise phrases and sentences. It can be calculated using equation 4.8 where S refers to number of substitutes, D refers to number of deletions, I refers to the number of insertion needed to convert the synthesized speech into the reference text (Greyling, 2022).

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \quad (4.8)$$

4.1.9 Mean Opinion Score

One of the most important metrics for evaluating the quality of synthetic speech is the Mean Opinion Score (MOS). It measures how pleasant, understandable, and natural the synthesized voice sounds to others. Performing the test involves participants scoring speech samples produced by a TTS system on a standard 5-point rating scale, with 1 representing poor speech quality and 5 representing human speech quality. It can be calculated using equation 4.9 where N is the total number of evaluators, and R_n represents the rating score by the n th evaluator (Arimas, n.d.).

$$\text{MOS} = \frac{\sum_{n=1}^N R_n}{N} \quad (4.9)$$

4.2 Results

4.2.1 Attention Score

Figure 1 shows the attention score of the Forward Tacotron system when trained with the original dataset in the first image, and when training is continued with the enhanced dataset in the second image.

In the first image, representing 500k steps of training on the original dataset, the attention score has high variability, with values frequently peaking near 0.7 and then dropping sharply to below 0.4. This suggests irregularities and a difficulty for the model to acquire a consistent alignment pattern, which is the result of utilizing a dirty or noisy dataset.

The second image shows the attention score after the model continues training on a clean dataset, building upon the first 500k steps. The diagram shows reduced variability and the range of scores is noticeably tighter, with fewer extreme peaks. Because of the higher quality of the dataset, the score stabilizes in a narrower range, primarily between 0.3 and 0.5, suggesting improved model consistency and learning efficiency.

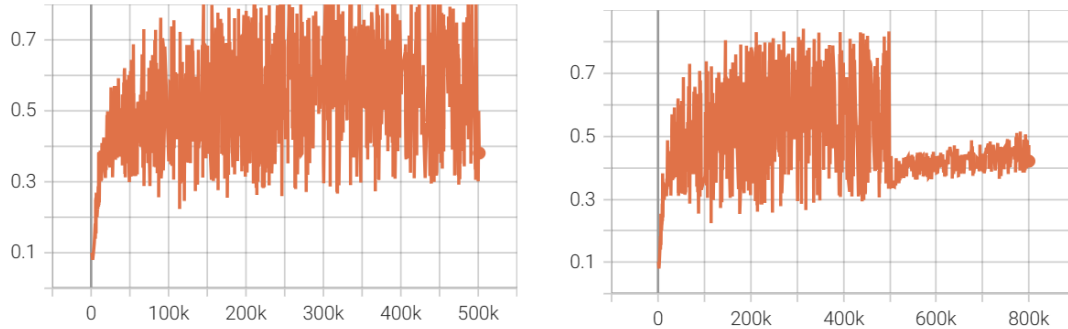


Figure 4.1: The Attention Score using the original dataset vs the attention score with transfer learning using the enhanced dataset.

4.2.2 CTC Loss

CTC loss is a type of loss function used specifically for training sequence learning problems where the alignment between the inputs and the target is not known beforehand. It is commonly applied to speech synthesis applications, where the goal is to train the model to comprehend sequential or temporal data without the need for precise frame-level alignment. It may serve as a gauge for how well the system is lining up the spoken output with the input text. Lower CTC loss values represent better performance as the model makes fewer alignment errors. A visualization of this metric can be seen in Figure 2.

The first graph displays a greater degree of variation in the CTC loss values, with some reaching as high as 4 and the majority varying around 3. This graph represents the model trained on an unclean dataset for 500k steps. This implies that the model had difficulty identifying recurring patterns in the data, most likely as a result of noise and irregularities inherent in a dirty dataset. Higher loss outliers are especially worrisome because they show times when the model's predictions were substantially inaccurate.

The ongoing training on a clean dataset is displayed in the second graph. In this case, the CTC loss values are smaller and show less fluctuation; they primarily stay below 3.5 and trend lower as training goes on. This pattern suggests that the model was able to learn more robustly and accurately with fewer alignment errors after the dataset was cleaned, which produced a more accurate and stable TTS model.

It is clear from comparing the two graphs how dataset quality affects TTS system training. The model learns on a more stable trajectory thanks to the cleaner dataset, which improves CTC loss metrics. Even with the improved dataset, there is still noticeable fluctuation, indicating that although the enhanced dataset is of higher quality, it is still suffering from some of the shortcomings of the original.

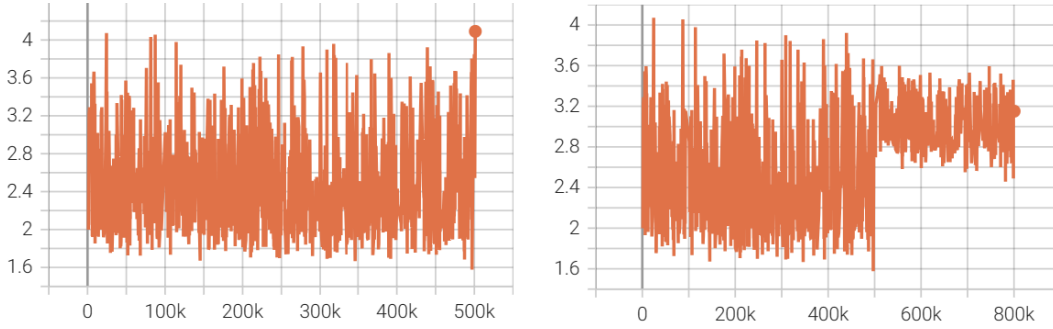


Figure 4.2: The CTC Loss of the Tacotron step using the original dataset vs the CTC Loss with transfer learning using the enhanced dataset.

4.2.3 DIA Loss

The DIA Loss is calculated to encourage the attention matrix of the model to focus near the diagonal, guiding the attention mechanism of the neural network. Better performance is indicated by a lower DIA loss, which shows that the TTS system can predict the rhythm and pitch of speech more precisely. A visualization of this metric can be seen in Figure 3.

The first image shows a high loss peak at the beginning that rapidly tapers off is seen in the first image, which depicts the TTS system trained for 500k steps on the original dataset. Training on poorly recorded or inconsistent audio samples results in the model struggling to identify patterns at first because of the noisy data, but it eventually picks up the ability to weed out unimportant features.

Building on the foundation created by the first 500k steps, the second image depicts the TTS system as it continues to train on a cleaner dataset. The graph seems to have less fluctuations and the loss drops more gradually, indicating that the TTS system is learning and generalizing effectively. The system's capacity to anticipate durations and intonation has been effectively refined by the prolonged training on cleaner data, which may improve the synthesized speech making it more natural and intelligible.

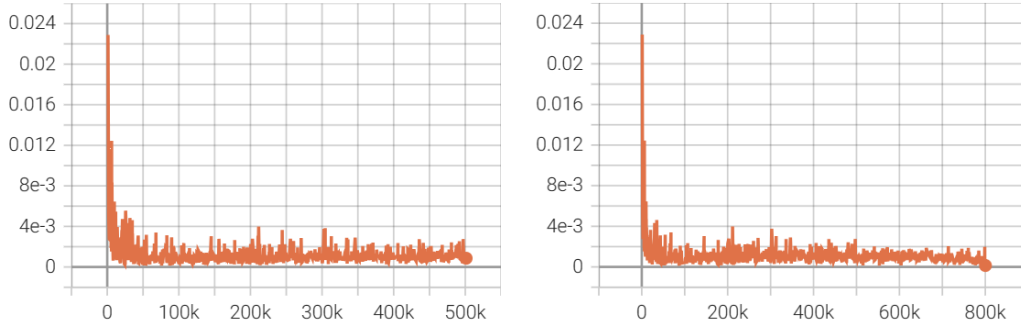


Figure 4.3: The DIA Loss of the Tacotron step using the original dataset vs the DIA Loss with transfer learning using the enhanced dataset.

4.2.4 Mel Loss

The Mel spectrogram loss graphs provided in Figure 2 and 3 illustrate the performance of the Forward Tacotron TTS system over training iterations for the Tacotron step, and for the Forward step respectively.

For Figure 4 and 5, the first image shows the TTS system trained on an unclean dataset for 500k steps, and 600k steps respectively. At first, the Mel loss shows a steep decline, which is followed by a plateau. This implies that although the model gained a lot of initial knowledge from the original dataset, its capacity to advance further was constrained, most likely because of the noise and irregularities in the training set.

In the second image, the system continues but on a clean dataset. The loss continues to decline but more gradually. The eventual leveling off at a lower loss value compared to the first graph indicates that cleaning the dataset allowed the model to extract more useful patterns and relationships from the data, thus achieving a better understanding of the speech generation process. The occasional spikes seen in the loss could be due to the model encountering new, yet-unlearned patterns within the clean data.

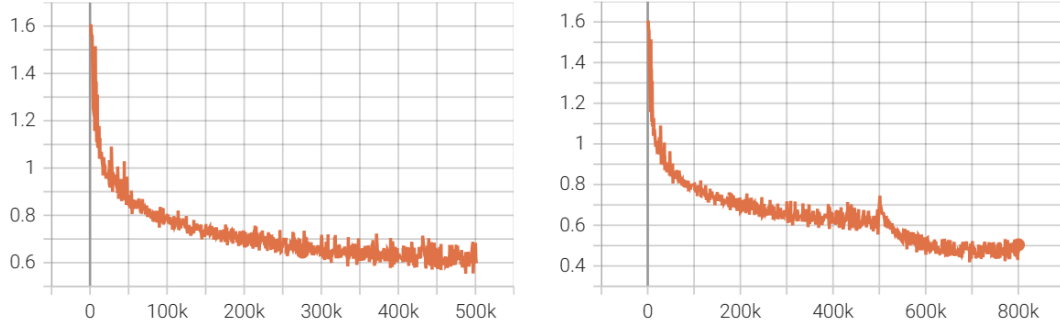


Figure 4.4: The Mel Loss of the Tacotron step using the original dataset vs the Mel Loss with transfer learning using the enhanced dataset.

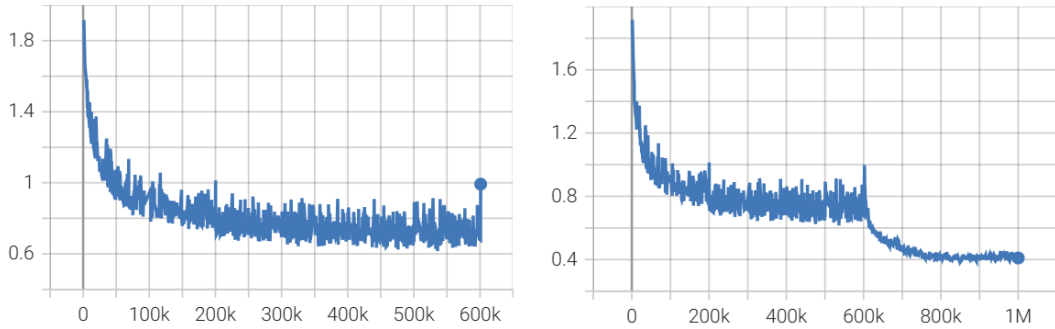


Figure 4.5: The Mel Loss of the Forward step using the original dataset vs the Mel Loss with transfer learning using the enhanced dataset.

4.2.5 Duration Loss

The Duration loss graphs provided in Figure 6 illustrate the performance of the Forward Tacotron TTS system over training iterations for the Forward step. With a high peak that quickly drops, the first image depicts a system with a considerable degree of loss variability, particularly noticeable in the early phases of training. Because of the noise or inconsistencies in the dataset, it is possible that the system found it difficult to learn from the original data. While there is significant fluctuation throughout the training process, the loss does begin to stabilize as the training goes on, though still displaying severe fluctuations ranging from 0.4 to 1.2, indicating that

the system may not have fully adapted to the original data. However, after 600k steps, the second image shows what happens when training on a clean dataset. A more consistent and noise-free dataset appears to have better patterns for the system to learn, as evidenced by the duration loss, which generally declines with smoother transitions and fewer spikes. TTS systems may become more stable and dependable as a result of the overall loss appearing to drop and the loss's decreased variability when compared to the previous training phase. There is a noticeable outlier spike at the end, which could be caused by a number of things, such as an anomaly in the data. It can also indicate that the enhanced dataset still suffers from some of the shortcomings of the original dataset.

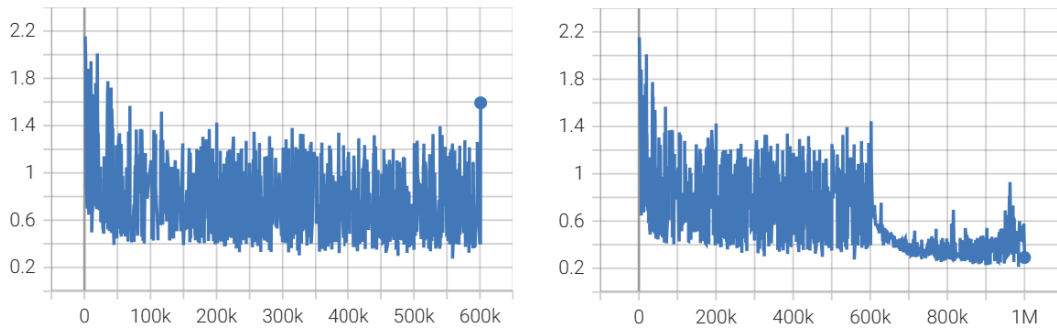


Figure 4.6: The Duration Loss of the Forward step using the original dataset vs the Duration Loss with transfer learning using the enhanced dataset.

4.2.6 Pitch Loss

This metric measures how accurately the model predicts the pitch patterns of the speech, which are essential for conveying emotions and intentions in spoken language. A visualization of this metric can be seen in Figure 8.

The first graph represents the model trained for 600,000 steps on the original dataset. Higher variance in pitch loss is seen in this training stage, which usually means that learning has been inconsistent as a result of noisy or faulty data. The spike at the end can indicate that the model may have

run across outliers or anomalies in the dataset, which could have an impact on the stability and quality of the generated speech.

On the other hand, the second graph shows the ongoing training on a cleaner dataset. The cleaner dataset may have helped the model improve its learning, resulting in more stable pitch predictions, as evidenced by the decrease in pitch loss and the smoother trend line. The model also shows less data irregularities, which usually results in a better-sounding synthesized voice. This is indicated by the decreased variance and the absence of large spikes. Furthermore, the pitch loss decreases notably as the number of steps exceeds 600k, showing the model's improved ability to predict pitch more accurately, likely due to the higher quality of the training data. The spike at the end of the second image might represent an outlier or anomaly that the model encountered later in training in the cleaner dataset. Showing that the enhanced dataset still suffers from some of the shortcomings of the original dataset.

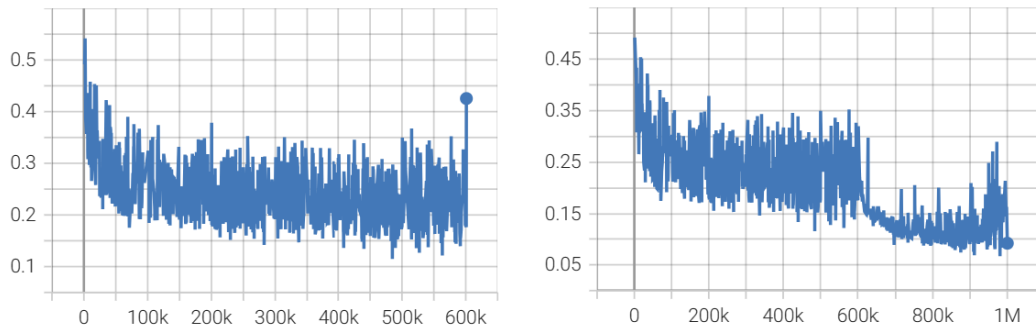


Figure 4.7: The Pitch Loss of the Forward step using the original dataset vs the Pitch Loss with transfer learning using the enhanced dataset.

4.2.7 Energy Loss

The Energy loss depicted in Figure 8 measures the difference between the predicted and actual energy values in synthesized speech, directly impacting the naturalness and intelligibility of the produced audio.

The first image shows the model trained on the original dataset for 600k steps. The graph displays a comparatively higher energy loss with a mean that seems to fluctuate between 0.024 and 0.036 with some peaks. The graph indicates that there may be variations in the energy loss between training instance, which could be caused by the lower quality dataset and lead to inconsistent performance from the model.

On the other hand, the second image shows the model following additional training on a cleaner dataset. The graph shows the energy loss trending downward, with values falling between 0.015 and 0.025. Furthermore, the graph shows a decrease in fluctuations. This indicates that the cleaner data can reduce the variability in the model's learning, resulting in a more dependable performance and possibly producing speech with energy dynamics more closely resembling natural human speech.

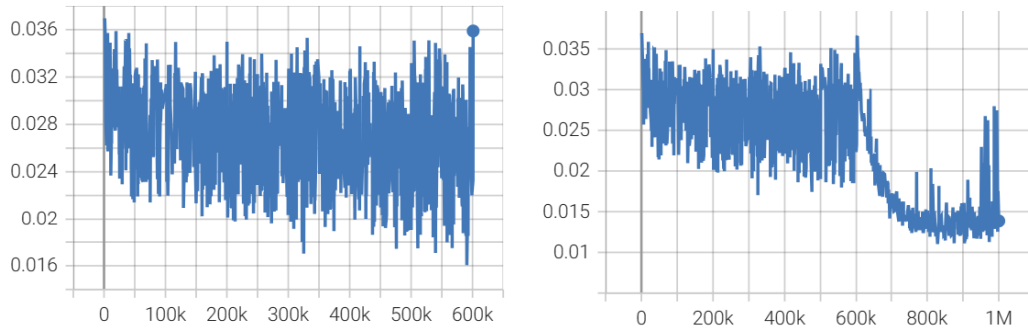


Figure 4.8: The Energy Loss of the Forward step using the original dataset vs the Energy Loss with transfer learning using the enhanced dataset.

4.2.8 Word Error Rate

The Word Error Rate depicted in table 4.1 shows the performance of the TTS system before and after the incorporation of the enhanced dataset. A lower WER means a more accurate TTS system. To illustrate the impact of the enhanced dataset, 18 sentences were generated using both the original model and the model with the enhanced dataset. The results, as shown in

table 4.1, demonstrate a clear improvement in the WER score after using the enhanced dataset. This shows that the quality of the dataset is essential to training a TTS system, and the possibility of using transfer learning on the Forward Tacotron model.

	WER
Original dataset	78
Enhanced dataset	34

Table 4.1: The WER of the current TTS system

4.2.9 Mean Opinion Score

The Mean Opinion Score depicted in table 4.2 shows the performance of the TTS system before and after the incorporation of the enhanced dataset. A higher MOS means a more accurate and natural TTS system. To illustrate the impact of the enhanced dataset, 18 sentences were generated using both the original model and the model with the enhanced dataset. The results, as shown in table 4.2, indicate that using the improved dataset resulted in a significant improvement in the MOS. This demonstrates the need of a high-quality dataset for TTS system training as well as the potential for transfer learning with the Forward Tacotron model.

	MOS
Original dataset	2.2
Enhanced dataset	3.4

Table 4.2: The MOS of the current TTS system

Chapter 5

Discussion, Challenges, and Conclusion

5.1 Discussion

The study found that Forward Tacotron can be used with the Bahraini dialect, particularly without the use of diacritics. This adaption is significant since diacritics provide important vocalization instructions required by speech synthesis systems. Using a dataset of about nine hours is an ambitious strategy, especially when compared to more known datasets, which typically span at least 24 hours and are methodically designed to provide high quality data with consistent recording settings and experienced speakers.

Forward Tacotron was chosen as the synthesis model due to its continual development and stated robustness and flexibility. This conclusion is based on the ongoing improvements in AI and voice synthesis technology, where newer models usually exceed their predecessors. Thus, using a newer model that is frequently being updated should theoretically yield a better TTS sys-

tem.

Despite these advancements, the performance of the Forward Tacotron model trained on the Bahraini dialect dataset was inconsistent. While it managed to correctly emulate the Bahraini dialect and accurately pronounce some sentences as if diacritics were present, it struggled with others. For some sentences, the model failed to articulate words comprehensibly. This inconsistency could be attributed to the limited size of the dataset, which does not provide enough vocabulary and variation to enable the model to learn the language effectively.

Optuna was used to automate the tuning of the hyper-parameters in the Tacotron phase and the Forward phase of our system. Manually tuning around 15 parameters for each phase to find the best combination to use with our dataset and model would be labor intensive and inefficient. Optuna automated this process while also using advanced pruning strategies. This made the training process much easier and less cumbersome.

To improve the model's performance, the dataset was artificially extended to 36 hours using data augmentation techniques. Although this helped to smooth and naturalize the model's output, the basic vocabulary and linguistic features remained the same.

To improve the speech quality, an AI enhanced and augmented dataset of 9 hours was used for Transfer learning. Building on the original dataset, the model showed significant improvement on all matrices and in speech quality.

5.2 Challenges

Although machine learning methods have good achievements for intrusion detection, they still face some challenges.

- 1) **Creating the dataset.** Creating the dataset was extremely challeng-

ing. We had to download each video in Sanaa Alsaad’s YouTube channel. Then we had to go through a tedious and time consuming process of splitting the audio into 1 to 10 seconds sections. We then had to transcribe each section. Overall, the process of creating the dataset took months of work.

- 2) **Dataset not intended for TTS systems.** The dataset used for this study demonstrates a wide range of recording qualities. Many recordings do not meet the needed level of clarity and consistency, which is critical for training robust TTS systems. Such inconsistency in recording quality can introduce a large amount of noise into the data, possibly complicating the training process and reducing the performance of the model. Furthermore, the speaker’s inconsistent articulation and frequent pronunciation errors, such as letter flipping, mispronunciation, and omissions, worsen the problems. These issues are mainly due to the nature of the source material, which consists of YouTube videos that were not designed for TTS applications.
- 3) **Small dataset size.** The dataset used in this study comprises of 9 hours of recorded audio. The most common TTS dataset in the English language is the LJ Speech dataset which is 24 hours long, and is professionally recorded by a skilled speaker. The most common German dataset is the Thorstenmueller dataset which is also 24 hours long, and professionally recorded. Thus, 9 hours of non-diacritised and poorly recorded dataset is not sufficient to train a TTS system, especially given the complexity of the Arabic language compared to other languages.
- 4) **Modifying the model for Optuna implementation.** We had to thoroughly study the Forward Tacotron model and understand how it works in order to modify the code for Optuna implementation. This involved

modifying multiple python files and a lot of trial and error to get it to work. Furthermore, the Forward Tacotron model consists of 2 phases, the Tacotron phase and the Forward phase. Each phase runs for multiple steps including different batch size, learning rate, and r value for each step. For example, for the Tacotron phase, step 1 has an r value of 3, learning rate of 0.0005, batch size of 64. Step 2 has an r value of 2, learning rate of 0.0001, batch size of 32. Step 3 has an r value of 1, learning rate of 0.0001, batch size of 16, and so on. We had to implement Optuna so that the first step predicts the model's layers, while the following steps predicts the best batch size, learning rate, and r value.

- 5) **Modifying the model to accept the Arabic alphabet.** Most Arabic TTS systems use diacritised text and phonetization tools. Those studies define the English transliterations of the Arabic alphabet in the vocabulary part of their code. Thus, finding online resources that have the same approach as our study was challenging. One of the pre-requests of training Forward Tacotron was to use espeak-ng phonetization tool. However, training the model while using the tool yielded high error rates. We did not understand the reason for the high error rate until some time. We realized that espeak-ng relies on diacritised text, and we had to modify the model's code to disable the phonetization tool.
- 6) **Time constraint.** Using Optuna to optimize each phase along with each step of the model was time consuming. Optuna was used for the Tacotron step, and for the Forward step. Each step took around 5 days on the GPU node to find the best hyperparameters. Although it is faster and less tedious than manually going through the process, training still consumed a lot of time.

- 7) **Creating the enhanced dataset.** Creating the enhanced dataset was also time consuming. We had to sift through 9 hours of the original audio to get the segments where the speaker speaks clearly and the recording quality is acceptable. We then had to remove leading and trailing silences along with the long silences in between for each audio file. After that we had to enhance the resulting audio files using Adobe Podcasts, which removes background noise and improves the sound quality using AI. This was a labor intensive manual process.

5.3 Conclusion

Data augmentation was an essential technique for training the TTS model. Expanding the dataset through data augmentation techniques introduced more variability and complexity to the training data, improving the model's robustness and generalization ability. Without data augmentation, the model was overfitting to the limited patterns and features and reached convergence much earlier compared to the original dataset. Furthermore, the study showed that a high quality dataset was critical to the development of a TTS system. This is demonstrated with the introduction of the AI enhanced dataset through transfer learning, the model's loss showed less variability, more consistency and much lower values for all metrics.

Moreover, Optuna was critical to this study since it automated the tuning of the hyper-parameters involved in the Tacotron and Forward phases of the TTS system. Given the challenges presented by the dataset and the complexity of the Forward Tacotron model, extending Optuna's runtime has the potential to greatly improve the model's accuracy. More time may help Optuna to improve its understanding of which parameters have the most effect on the model's performance, resulting in a more precise and potentially a better-performing parameter settings. Furthermore, the dataset must

be expanded, not just in terms of time, but also in quality and consistency. Given the complexity of Arabic relative to English, it is expected that a well-rounded Arabic TTS dataset will be at least 36 hours long, professionally recorded in a consistent recording setting, and contain well articulated words and sentences. In addition, investigating different TTS systems may provide ideas into possibly better frameworks for Arabic. While Forward Tacotron has several benefits, its performance with Arabic and its dialects, such as Bahraini, should be properly compared to other existing systems. Such research might highlight strengths and weaknesses in various models, allowing for better decisions based on the linguistic and dialectal requirements.

References

- Abdelali, A., Durrani, N., Demiroglu, C., Dalvi, F., Mubarak, H., & Darwish, K. (2022, 06). Natiq: An end-to-end text-to-speech system for arabic.
doi: 10.48550/arXiv.2206.07373
- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *CoRR*, *abs/1907.10902*. Retrieved from <http://arxiv.org/abs/1907.10902>
- Arimas. (n.d.). *M.o.s. mean opinion score*.
<https://arimas.com/2017/08/01/mos-mean-opinion-score/>.
(Accessed: 2023-05-05)
- Azab, A., Zaky, A., Ogawa, T., & Gomaa, W. (2023). Masry: A text-to-speech system for the egyptian arabic. In *Proceedings of the 20th international conference on informatics in control, automation and robotics - volume 2: Icinco* (p. 219-226). SciTePress. doi: 10.5220/0012244300003543
- Aziz, A. F. A., Tiun, S., & Ruslan, N. (2023). End to end text to speech synthesis for malay language using tacotron and tacotron 2. *International Journal of Advanced Computer Science and Applications*. Retrieved from <https://api.semanticscholar.org/CorpusID:259311278>
- Badlani, R., Lancucki, A., Shih, K. J., Valle, R., Ping, W., & Catanzaro, B.

- (2021). One TTS alignment to rule them all. *CoRR*, *abs/2108.10447*. Retrieved from <https://arxiv.org/abs/2108.10447>
- Fahmy, F., Abbas, H., & Khalil, M. (2022, 03). Boosting subjective quality of arabic text-to-speech (tts) using end-to-end deep architecture. *International Journal of Speech Technology*, 25, 1-10. doi: 10.1007/s10772-022-09961-0
- Gopalakrishnan, T., Imam, S., & Aggarwal, A. (2022, 07). Fine tuning and comparing tacotron 2, deep voice 3, and fastspeech 2 tts models in a low resource environment. In (p. 1-6). doi: 10.1109/ICD-SIS55133.2022.9915932
- Greyling, C. (2022, 11). *How important is measuring word error rate (wer) for voicebots?* <https://cobusgreyling.medium.com/how-important-is-measuring-word-e> (Accessed: 2023-05-05)
- Hamdy, A., Magdy, M., Alfawzy, M., Ghaly, M., & Abbas, H. (2021, 03). Arabic speech synthesis using deep neural networks. In (p. 1-6). doi: 10.1109/ICCSPA49915.2021.9385731
- Hanane, T., & Hamadouche, M. (2022, 05). Multi-agent based arabic speech synthesis. *International Journal of Speech Technology*, 1-17. doi: 10.1007/s10772-022-09975-8
- Henriquez, A. (2020, 11). *Understanding attention in transformers models*. <https://medium.com/analytics-vidhya/understanding-attention-in-tra> (Accessed: 2023-05-05)
- Kong, J., Kim, J., & Bae, J. (2020). Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis.
- Koçak, T., & Büyükzincir, M. (2023, 06). Building a turkish text-to-speech engine: Addressing linguistic and technical challenges. In (p. 1-4). doi: 10.1109/DSP58604.2023.10167970
- Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., &

- Liu, T. (2019). FastSpeech: Fast, robust and controllable text to speech. *CoRR*, *abs/1905.09263*. Retrieved from <http://arxiv.org/abs/1905.09263>
- SanaaAlsaad. (2017, 06). *Sana al saad - bahrain*. <https://www.youtube.com/channel/UCjJpd0t3E4RNbB4wHLaqaZA/>. (Accessed: 2023-05-05)
- Sawalha, L., & Al-Radhi, M. (2023, 01). Improving naturalness of neural-based tts system trained with limited data. In (p. 71-75). doi: 10.3311/WINS2023-013
- Schäfer, C. (2023, 05). *ForwardTacotron: Implementing tacotron with forward attention*. <https://github.com/as-ideas/ForwardTacotron>. (Accessed: 2023-05-04)
- Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... Wu, Y. (2017). Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, *abs/1712.05884*.
- Tran, D. C. (2021). The first fosd-tacotron-2-based text-to-speech application for vietnamese. *Bulletin of Electrical Engineering and Informatics*. Retrieved from <https://api.semanticscholar.org/CorpusID:224917735>
- Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., ... Saurous, R. A. (2017). Tacotron: A fully end-to-end text-to-speech synthesis model. *CoRR*, *abs/1703.10135*. Retrieved from <http://arxiv.org/abs/1703.10135>
- Yeo, A. (2020, 08). *Speech recognition — connectionist temporal classification (ctc)*. <https://medium.com/machine-learning-evaluation-metrics-precision/s>. (Accessed: 2023-05-05)

ملخص الدراسة

في السنوات القليلة الماضية، شهد مجال الذكاء الاصطناعي (AI) وتعلم الآلة (ML) تطوراً متسارعاً. تم تطوير مختلف وحدات توليد الكلام إلى نص (TTS)، وتنفيذها بنجاح في العديد من اللغات. ومع ذلك، كان تنفيذ مثل هذه الأنظمة للغة العربية قليلة، حيث اعتمد تنفيذها في الغالب على حلول مؤقتة مثل استخدام العلامات التشكيلية لترجمة النص العربي إلى الإنجليزية. تقدم هذه الدراسة تطوير نظام لتوليد الكلام إلى نص (TTS) للهِجَة البحرينية. تعتمد الطريقة المقترحة على نموذج Forward Tacotron مدمج مع مولد صوتي HiFi-GAN، و Optuna لتحسين الهيربارامترات. توفر الطريقة لدينا فهماً شاملاً لمتطلبات توليد الكلام العربي البحريني الطبيعي والتعبيري، وتبحث في إمكانية تدريب مثل هذا النظام دون استخدام العلامات التشكيلية أو أدوات التصويت الصوتي. لأغراض هذه الدراسة، قمنا بإنشاء مجموعة بيانات بالهِجَة البحرينية، تحتوي على 9 ساعات من التسجيلات المأخوذة من قناة يوتيوب لثناء السعد. ومع ذلك، بسبب صغر حجم مجموعة البيانات، تم استخدام تقنيات زيادة البيانات لتوسيع المجموعة إلى 36 ساعة. تم إنشاء نسخة أخرى من مجموعة البيانات تمتد إلى 8 ساعات، وتشمل التسجيلات ذات الجودة الأعلى فقط. ثم تم تحسين مجموعة البيانات الأصغر بواسطة Adobe Podcasts. سُجلت بيانات التسجيل بمعدل 22050 هرتز، وفقاً لمتطلبات النموذج. تم تدريب نظامنا أولاً على مجموعة البيانات 36 ساعة، ثم تم استخدام التعلم النقلي لضبط التدريب على مجموعة البيانات 8 ساعات. تم تقييم النظام باستخدام مقاييس مثل درجة الانتباه، Mel Loss، Duration Loss، و Pitch Loss، ومعدل أخطاء الكلمات (WER)، ومتوسط درجة الرأي (MOS). أظهر النموذج قدرات واعدة في تكرار الهِجَة المحلية مع MOS تبلغ 3.4، و WER تبلغ 34.

جامعة البحرين

كلية تقنية المعلومات



التحقيق في امكانية توليد الكلام بالهجه البحرينية وما يتبعها من مجموعات بيانات و خوارزميات

أطروحة مقدمة كجزء من متطلبات الحصول على
درجة الماجستير في الذكاء الاصطناعي

إعداد

محمد باسل ال محمود

٢٠٢١٠٠٢٢٣

إشراف

أ.د. فاطمه البلوشي

(أستاذ مساعد)

أ.د. هشام العمال

(أستاذ مساعد)

جامعة البحرين

مملكة البحرين

سبتمبر ٢٠٢٢