



Cairo University
Faculty of Engineering
Department of Computer Engineering

O2Doublage

Speech Synthesis



A Graduation Project Report Submitted
to
Faculty of Engineering, Cairo University
in Partial Fulfillment of the requirements of the degree
of
Bachelor of Science in Computer Engineering.

Presented by
Rawaa Ahmed Said Hanafy

Supervised by
Dr. Micheal Nashaat
July 11, 2023

All rights reserved. This report may not be reproduced in whole or in part, by photocopying or other means, without the permission of the authors/department.

ACKNOWLEDGMENT

First of all we want to thank Allah for giving us the ability to learn and improve ourselves and ask him to perpetuate gifts for us to do more in our life. Then we would like to express our deepest appreciation and special thanks to Dr. Micheal Nashaat because he helped us a lot and provided us the possibility to complete this Project. We also want to thank all professors through our educational years who gave us knowledge and finally big thanks for our parents because without them we are nothing.

Table of Contents

| | |
|---|-----------|
| ACKNOWLEDGMENT..... | 2 |
| Table of Contents..... | 3 |
| List of Figures..... | 4 |
| List of Tables..... | 4 |
| List of Abbreviation..... | 5 |
| Chapter 1: Introduction..... | 1 |
| Chapter 2: Model Design and Architecture..... | 1 |
| 2.1. Overview and Assumptions..... | 1 |
| 2.2. Model Architecture..... | 1 |
| 2.5.1. Functional Description..... | 1 |
| 4.5.1.1. The main stages in the module:..... | 2 |
| 4.5.1.2. FastSpeech2 detailed functional description:..... | 2 |
| 4.5.1.3. Variance adaptor components functional description:..... | 3 |
| 4.5.1.4. HiFi-GAN components functional description:..... | 4 |
| 4.5.2. Modular Decomposition..... | 4 |
| 2.5.3. Design Constraints..... | 6 |
| Chapter 3: System Testing and Verification..... | 6 |
| 3.2. Integration Testing..... | 12 |
| 3.3. Comparative Results to Previous Work..... | 16 |
| References..... | 17 |

List of Figures

- Figure 2.1: Block diagram indicates the model function
- Figure 2.2: Block diagram indicates the main stages in the model
- Figure 2.3: Indicates the main components of FastSpeech2
- Figure 2.4: The architecture of the variance adaptor
- Figure 2.5: The Hifi-GAN generator architecture
- Figure 2.6: The predictors components
- Figure 3.1: Mel loss on training and validation data vs number of iterations
- Figure 3.2: Pitch loss of the model on training and validation data vs number of iterations
- Figure 3.3: Energy loss on training and validation data vs number of iterations
- Figure 3.4: Duration loss on training and validation data vs number of iterations
- Figure 3.5: Post-net loss on training and validation data vs number of iterations
- Figure 3.6: Total loss on training and validation data vs number of iterations

List of Tables

- Table 3.1: Comparison between synthesized and ground truth mel-spectrograms on some samples.
- Table 3.2. The model output on unseen text.
- Table 3.3. MOS of previous models

List of Abbreviation

| | |
|------|---------------------------------|
| GAN | Generative Adversarial Networks |
| HMM | Hidden Markov Model |
| HiFi | High Fidelity |
| MOS | Mean Opinion Score |
| MPD | Multi-Period Discriminators |
| MSD | Multi-Scale Discriminators |
| STFT | Short-Time Fourier Transform |
| TTS | Text to speech |

This page is left intentionally empty



Chapter 1: Introduction

TTS is the final module in the pipeline. It converts Arabic text to speech represented by waveforms. It is a very important model used in many applications. Virtual assistants such as Apple's Siri and Microsoft's Cortana use TTS models to be able to have conversations with the users. In this project, Arabic TTS is used to convert translated text into Arabic speech. The work idea is to convert Arabic text to phones and convert these phones to mel-spectrogram representations. Finally, we use a vocoder to generate waveforms from mel-spectrograms.

Chapter 2: Model Design and Architecture

In this chapter, we will discuss the design of the system, the architecture, the implemented modules and the theory behind it in detail.

2.1. Overview and Assumptions

In this chapter, I will talk about all the details of the model including basic architecture, main stages and components, dataset and all the implementations.

2.2. Model Architecture

As a mel-spectrogram can represent time, frequency and amplitudes of each frequency component, then it represents all the information needed to reconstruct the waveform data. After predicting mel-spectrograms, we use the Hi-Fi GAN as a vocoder to reconstruct the signal.

2.2.1. Functional Description



Figure 2.1. Block diagram indicates the model function

The purpose of this module is to convert text to speech to be combined with the voice-off video. We try to produce correct, pure, and high-quality speech which could be understood well by humans.

2.2.1.1. The main stages in the module:

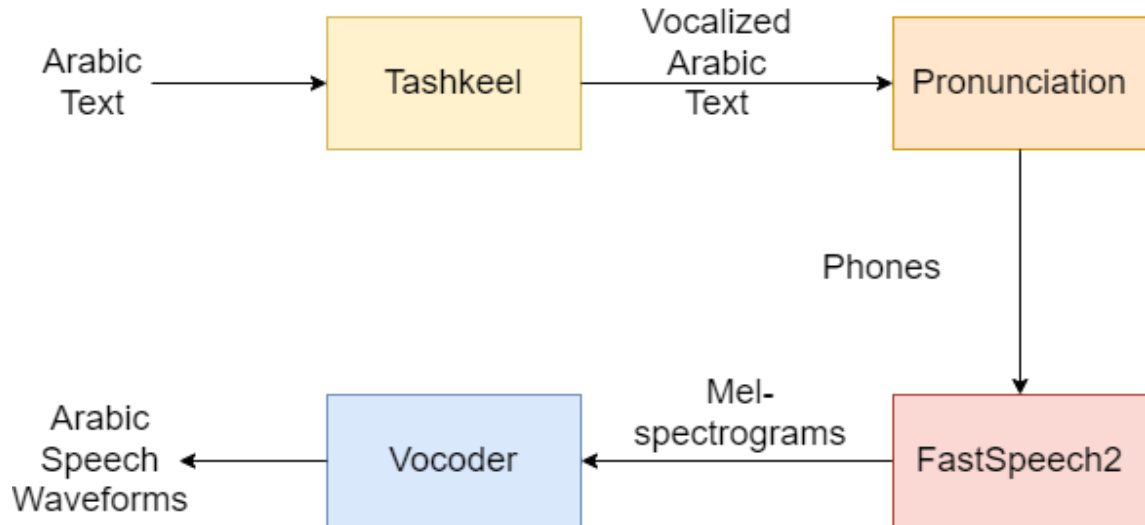


Figure 2.2. Block diagram indicates the main stages in the model

There are four main stages in the module:

1. **Tashkeel:** vocalizing Arabic text to get the desired pronunciation to increase the accuracy of the speech.
2. **Pronunciation:** converts vocalized text to phoneme sequence.
3. **FastSpeech2:** converts the phoneme to mel-spectrograms.
4. **HiFi-GAN vocoder:** generates final waveforms from mel-spectrograms.

2.2.1.2. FastSpeech2 detailed functional description:

1. **Phoneme embedding layer:** converts phonemes into phoneme embeddings.
2. **Encoder:** converts phoneme embeddings to phoneme hidden sequence.
3. **Variance adaptor:** adds variance information (pitch, energy, and duration) into the phoneme hidden sequence.
4. **Mel-spectrogram Decoder:** converts phoneme hidden sequence to mel-spectrograms.
5. **Post-net:** to convert mel-spectrograms to linear spectrograms with a higher resolution before the vocoder stage

Note: Post-net is not included in the actual implementation but we added it in this implementation to improve performance of the model.

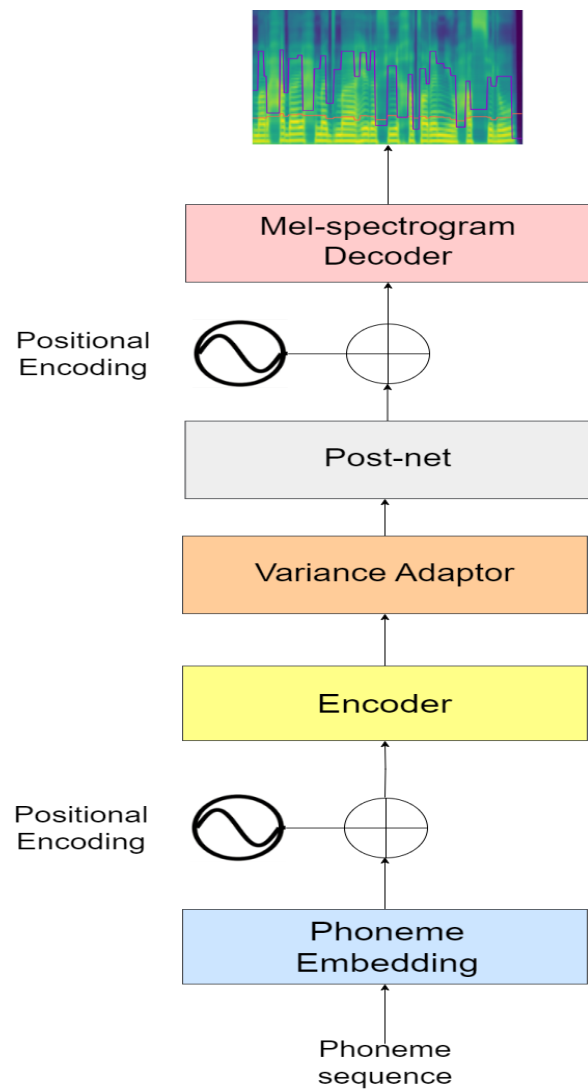


Figure 2.3. Indicates the main components of FastSpeech2

2.2.1.3. Variance adaptor components functional description:

1. **Pitch predictor:** takes the phoneme hidden sequence as input predicts the pitch spectrogram.
2. **Energy predictor:** predict the original values of energy instead of the quantized values.
3. **Duration predictor:** takes the phoneme hidden sequence as input and predicts the duration of each phoneme (the number of mel-spectrograms that each phoneme corresponds to), which represents how many mel frames correspond to this phoneme, and is converted into logarithmic domain for ease of prediction.
4. **Length regulator:** up-samples the phoneme sequence according to the phoneme duration to match the length of the mel-spectrogram sequence.

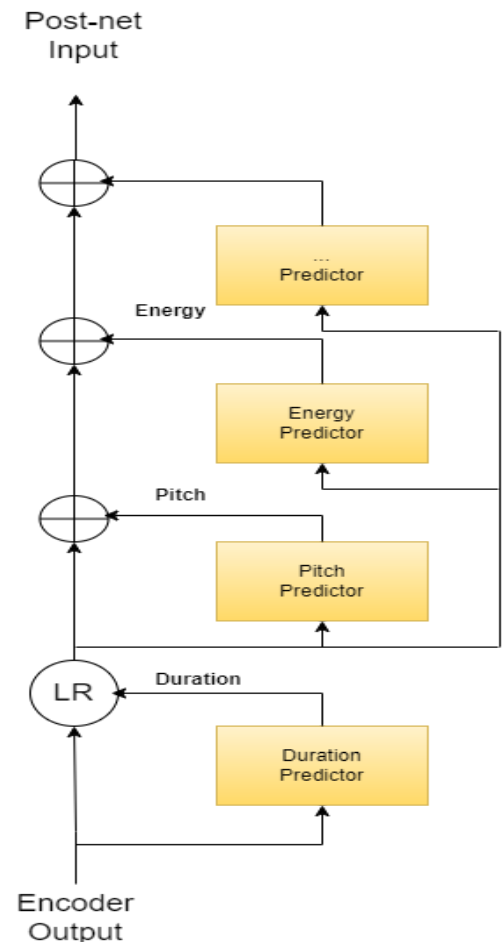


Figure 2.4. The architecture of the variance adaptor

2.2.1.4. HiFi-GAN components functional description:

1. **One generator:** gets the mel-spectrogram input and up-samples it until its length matches the temporal resolution of the waveform.
2. **Multi-period discriminators (MPD):** consists of multiple sub-discriminators each of which handles a portion of the input waveform.
3. **Multi-scale discriminators (MSD):** evaluates the audio samples at different levels.

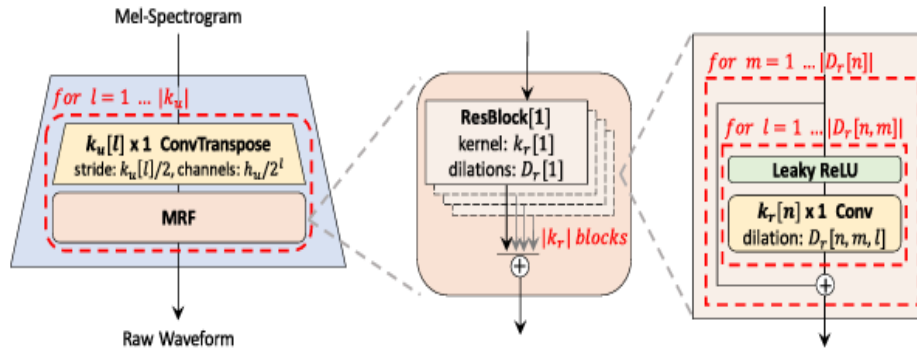


Figure 2.5. The HiFi-GAN generator architecture

2.2.2. Modular Decomposition

2.2.2.1. Dataset: Arabic Speech Corpus

It is a Modern Standard Arabic Speech corpus built specifically for speech synthesis but is used also for building Arabic HMM based voices. It contains:

- 1813 .lab files which contain Arabic text phonetics. Annotations include stress marks on individual phonemes.
- 1813 .wav files which are the spoken speech of the text. The total time of the records is more than 3.7 hours.
- 1813 .textGrid files which contain information on the .wav files on the phoneme level.

This corpus is a part of a doctoral project by **Nawar Halabi** at **University of Southampton**. The corpus is available for strictly non-commercial purposes through the official Arabic Speech Corpus website.

2.2.2.2. Preprocessing:

In the preprocessing stage, we compute mel-spectrograms, durations, pitch, and energy for all the 1813 .wav files. Some statistics are also computed for pitch and energy. We compute minimum, maximum, mean and standard deviation which are used for data normalization. Preprocessing takes about 19-39 minutes. Pitch and energy are computed on phoneme level which is more accurate than on frame level and gives very good results. Also, we remove outliers from data. The four features above are then stored in .npy files for training. The phonemes are aligned with corresponding .wav names and split into training and validation.

2.2.2.3. FastSpeech2 components:

1. **Encoder:** embedding layer, positional encoding layer for training, and a stack of self attention layer and 1D convolution layer.
2. **Decoder:** positional encoding layer for training and a stack of self attention layer and 1D convolution layer.
3. **Post-net:** five 1D convolution with 512 channels and kernel size 5.
4. **Variance adaptor:** three variance predictors for pitch, energy, and duration information and a length regulator.
5. **Variance predictor:**
 - 1D convolution layer followed by ReLU activation function.
 - Normalization layer.
 - Dropout layer.
 - 1D convolution layer followed by ReLU
 - Normalization layer.
 - Dropout layer.
 - Linear layer.
6. **Length regulator.**

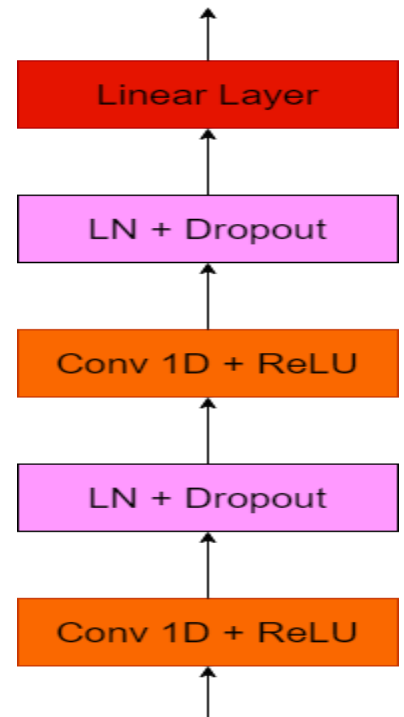


Figure 2.6. The predictors components

2.2.2.4 Training

1. The model was copied on many devices that each of them handles a portion of the input data in the forward pass. But the number of batches should be greater than the number of GPUs.
2. Forward pass: FastSpeech2 gets the inputs and predicts mels, post-net mels, pitch, energy, and duration.
3. Loss: predictions and target values are used to calculate the loss:
 - mels and post-net mels losses are computed using mean absolute error.
 - pitch, energy, and duration losses are computed using mean square error.
4. Backward propagation: the gradients are updated and clipped to keep within some suitable range.
5. Optimization is done to update the model weights.
6. The model is trained for 1,000,000 iterations.
7. To preserve the work, the model was saved every 100,000 steps and sometimes we decrease this value. We may stop training at some point and make the model complete from this step.
8. The model records the losses on training data every 1000 steps.
9. The model is tested on validation data every 10,000 steps.

10. Some log with kept for training and validation results and progress

2.2.2.5 Inference

1. We input normal Arabic text.
2. We may vocalize it if required.
3. We compute the phonetics.
4. The phonetics are fed to FastSpeech2.
5. The output is then given to the vocoder .
6. The vocoder generates the waveform.
7. The mel-spectrogram is saved in the results folder.

2.2.3. Design Constraints

The Arabic text is required to be properly vocalized so we can map it to the correct phoneme sequence.

Chapter 3: System Testing and Verification

The model was trained for 1,000,000 iterations. Through training, 6 losses were computed between predicted and targets on training and validation data:

1. Mel loss:

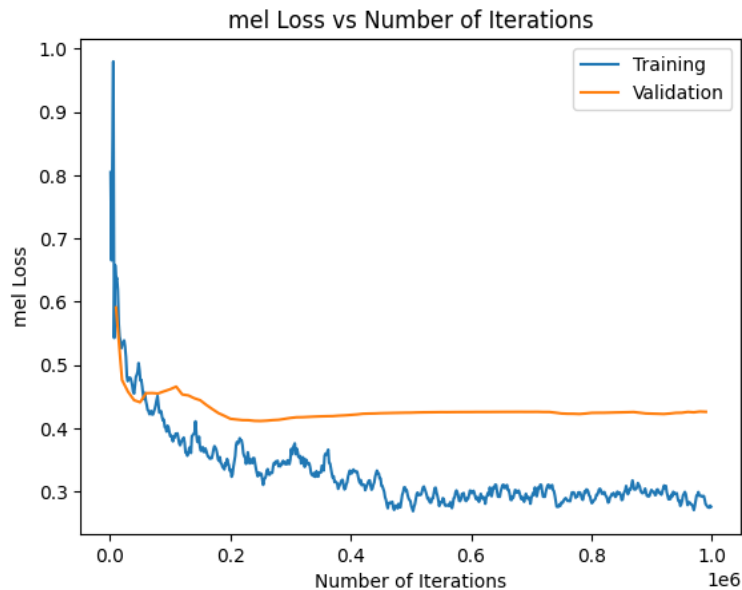


Figure 3.1. Mel loss on training and validation data vs number of iterations

From Figure 4.18., we can see that mel loss decreases for both training and validation data while number of iterations increases.

2. Pitch loss:

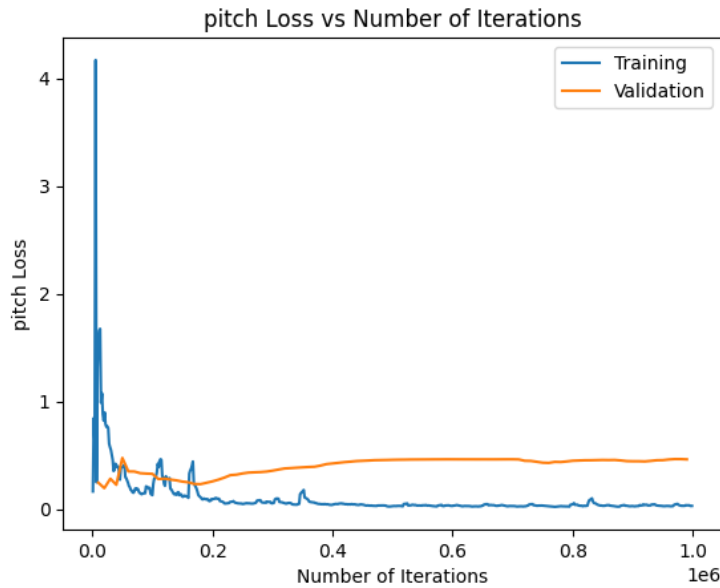


Figure 3.2. Pitch loss of the model on training and validation data vs number of iterations

This figure indicates that the pitch loss of the model on training data decreases when increasing the number of iterations while increases on validation data.

3. Energy loss:

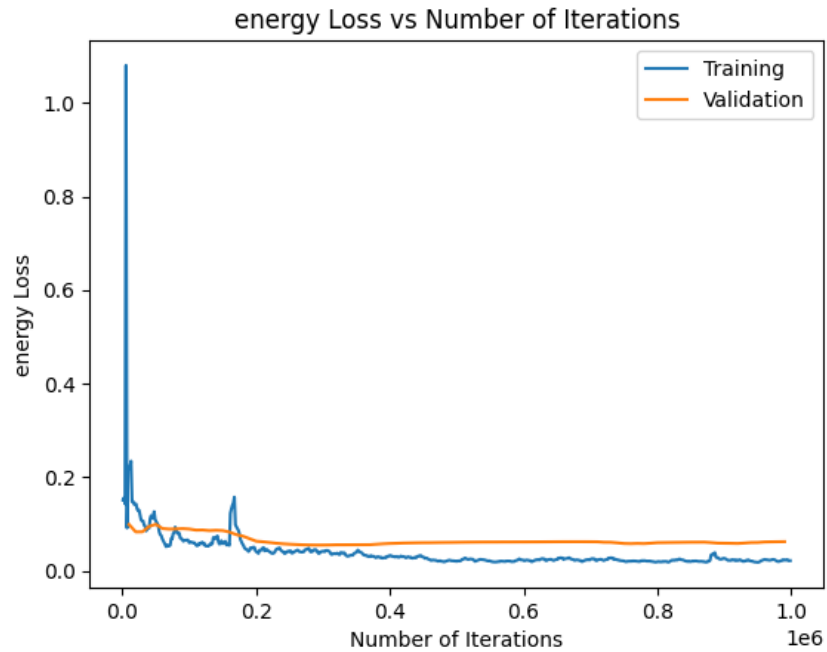


Figure 3.3. Energy loss on training and validation data vs number of iterations

The energy loss on validation and training data decreases when we increase the number of iterations.

4. Duration loss:

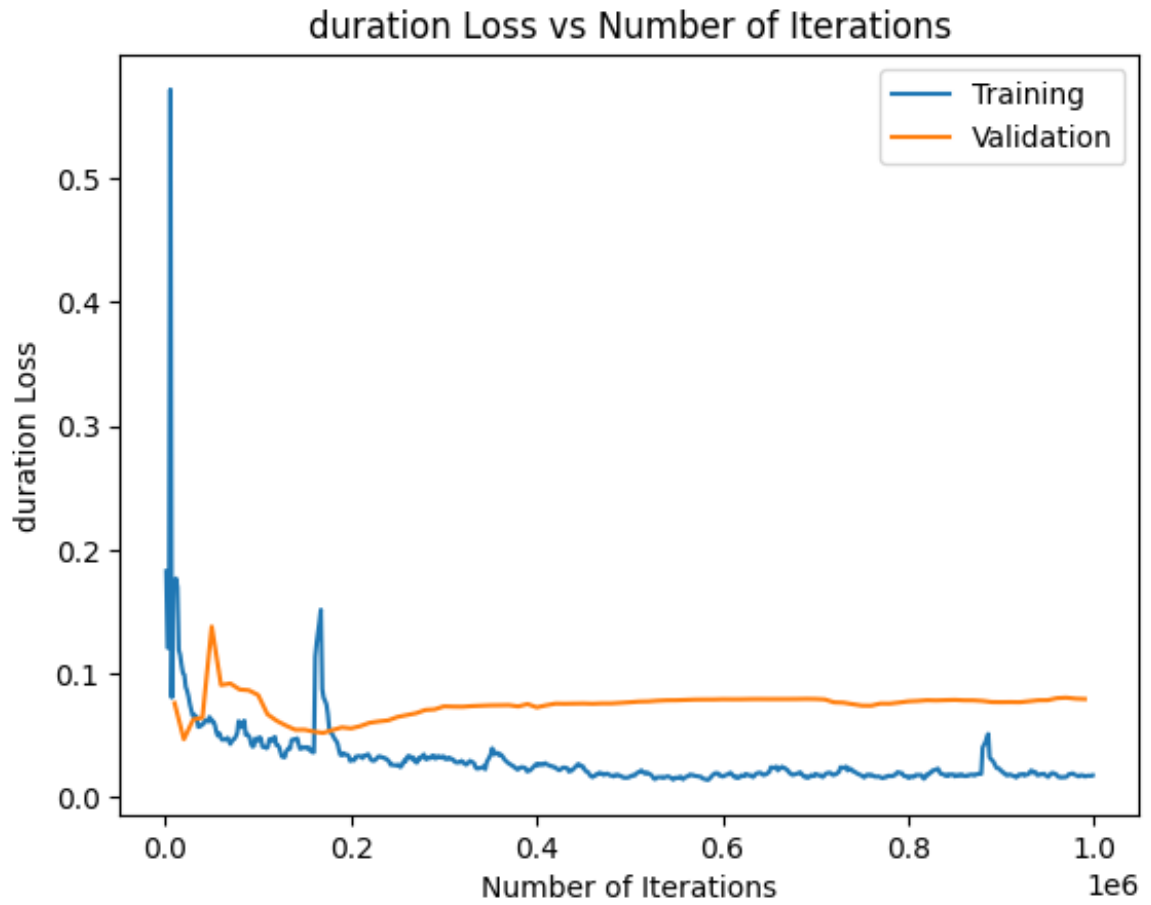


Figure 3.4. Duration loss on training and validation data vs number of iterations.

Just like the pitch, when increasing the number of iterations, the duration loss on training data decreases while increases on validation data.

5. Post-net loss:

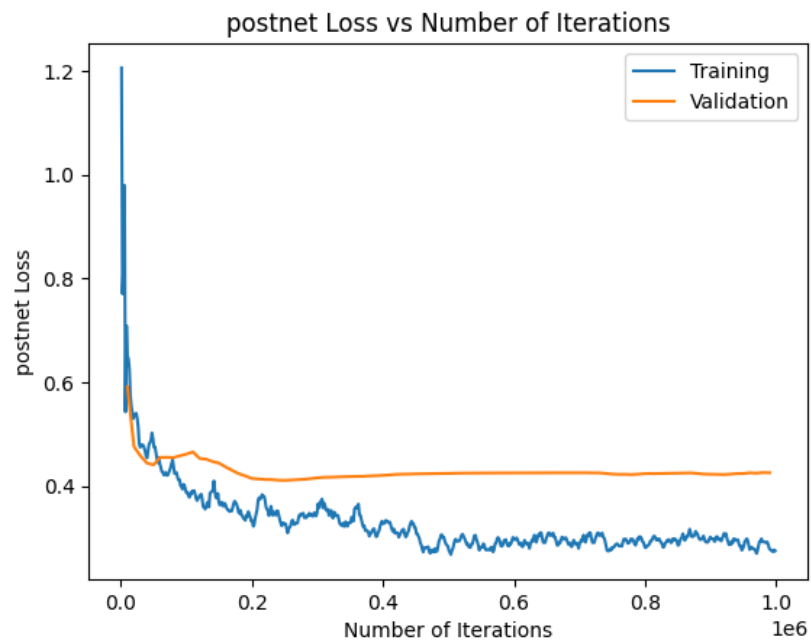


Figure 3.5. Post-net loss on training and validation data vs number of iterations

The post-net loss on validation and training data decreases when we increase the number of iterations.

6. Total loss: is the summation of the above losses

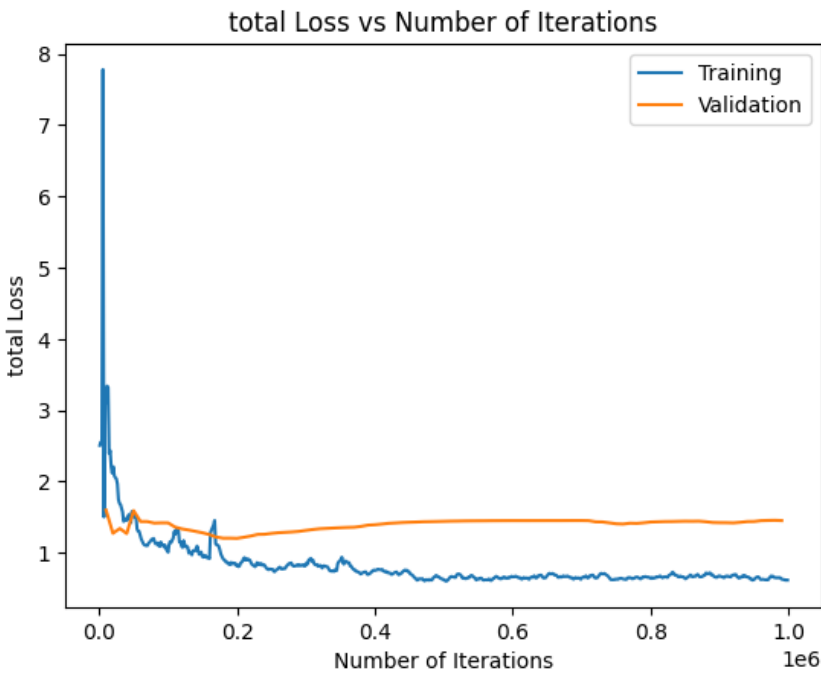


Figure 3.6. Total loss on training and validation data vs number of iterations

The total loss decreases on training data and doesn't change after 500,000 iterations, while on validation data, it decreases a bit then increases again then doesn't change when increasing the number of iterations.

3.2. Integration Testing

To perform integration testing the TTS model, mel-spectrograms for ground truth speech and synthesized speech.

| Text | Ground truth | Synthesized |
|---|--------------|-------------|
| <p>أَحْرَزَ الْقَاصُّ الْمِصْرِيَّ أَشْرَفُ الشَّرِّ بِنِي الْمَرْكَزَ الْأَوَّلَ عَنْ مَجْمُوعَتِهِ الْوَفَاءَ السَّعِيدَةَ لِعَبْدِهِ الْخَلَّاقِ</p> | | |
| <p>أَحْرَزَتِ الْإِمَارَاتِيَّةُ عَلِيَاءُ مُحَمَّدَ سَعِيدَ ذَهَبِيَّةَ سِبَاقِ عَشْرَةِ آلَافٍ مِثْرَ مُسَجَّلَةٍ وَاحِدَةٍ وَثَلَاثِينَ دَقِيقَةً وَوَاحِدَةً وَخَمْسِينَ ثَانِيَةً وَسِتَّ وَثَمَانِينَ جُزْءاً مِنَ الْمِئَةِ</p> | | |
| <p>كَشَفَتْ دِرَاسَةُ أَمْرِكِيَّةٍ أَنَّ الْمَشَاعِرَ الْأَسَاسِيَّةَ السَّتَّ سَعِيدٍ وَحَزِينٍ وَخَائِفٍ وَغَضْبَانٍ وَمُنْذِهَشٍ وَمُسْمِئِزٍ - لَا تُعْطِي كُلَّ التَّعَابِيرِ الَّتِي تُظْهِرُ عَلَى الْوَجْهِ كَمَا كَانَ يُعْتَقَدُ</p> | | |

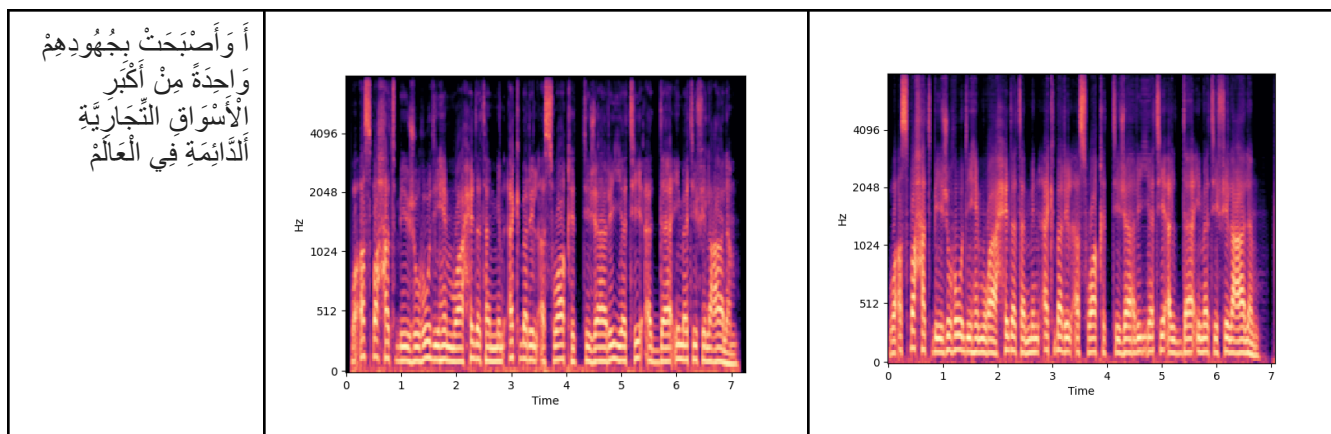
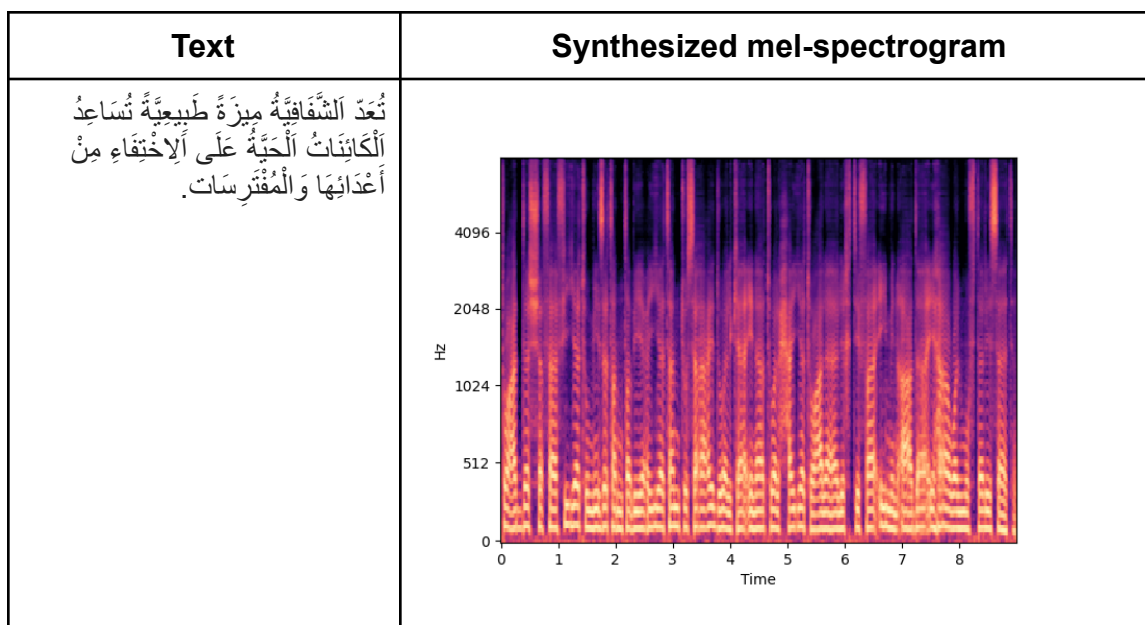


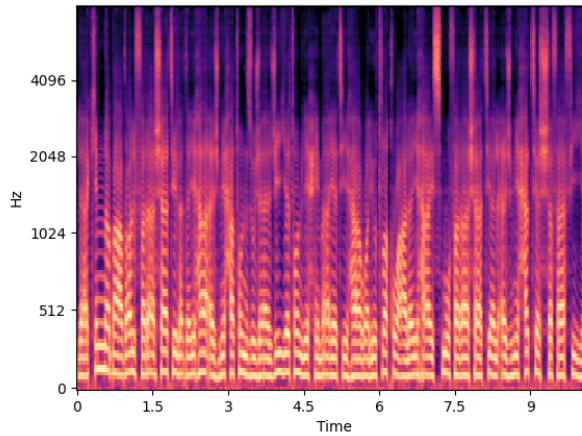
Table 5.4. Comparison between synthesized and ground truth mel-spectrograms on some samples.

As shown in the comparison table, the mel-spectrograms are very similar. This means that the model produces an output waveform that is very similar to the original. Also, the model is able to synthesize speech with high accuracy and understandable words.

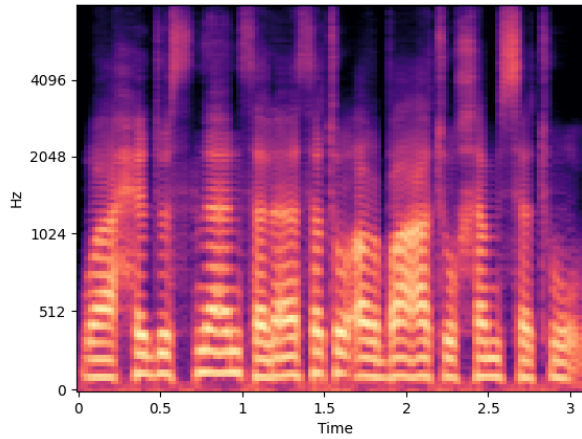
Here are new samples that the model wasn't trained on:



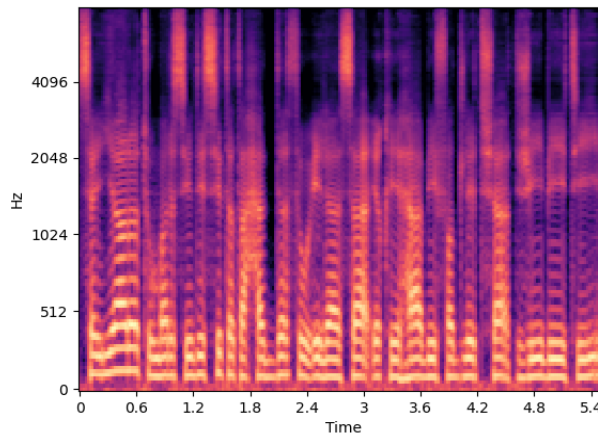
لَكِنَّ بَعْضَ الْفَرَاشَاتِ لَدَيْهَا أَجْنَحَةٌ
رُجَائِيَّةٌ لِحِمَايَتِهَا وَيُطْلَقُ عَلَيْهَا إِسْمُ "
فَرَاشَاتِ الْأَجْنَحَةِ الشَّفَافِيَةِ."



واحد اثنان ثلاثة أربعة خمسة



سَيَّارَةٌ مَرَسِيدِسٍ تَتَحَدَّثُ إِلَى سَائِقِهَا
بِفَضْلِ تَشَاتٍ جِي بِي تِي.



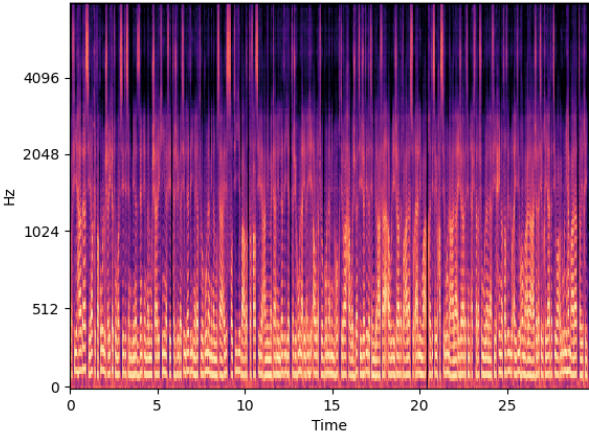
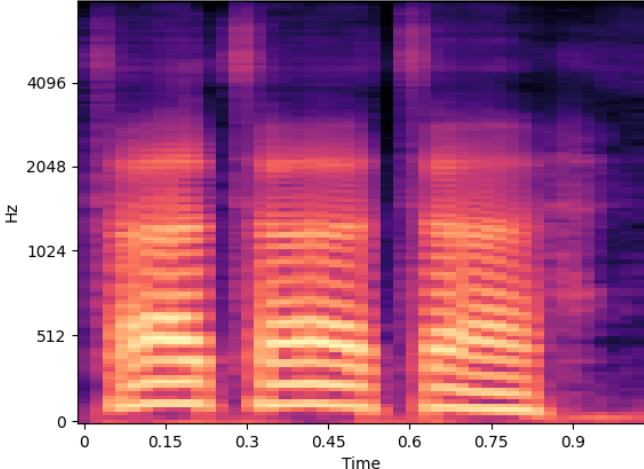
| | |
|---|---|
| <p>وَفِي هَذَا السِّيَاقِ تَقُولُ شَرَكَةُ مَرْسِيدِسْ بِنَزْ إِنَّ الْمَشَارِكِينَ فِي التَّجَرِبَةِ الَّذِينَ يَطْلُبُونَ مِنَ الْمُسَاعِدِ الصَّوْتِيَّ بَعْضَ التَّفَاصِيلِ عَنِ الْمَكَانِ الَّذِي يَتَجَهَّوْنَ إِلَيْهِ أَوْ يَطْلُبُونَ مِنْهُ إِفْتِرَاحَ وَجْهَةٍ جَدِيدَةٍ لِلْعِشَاءِ أَوْ الْإِجَابَةَ عَلَى سُؤَالٍ مُعَقَّدٍ ، سَيَحْصِلُونَ عَلَى إِجَابَةٍ أَكْثَرَ شُمُولًا ، فِي حِينٍ نَظْلُ أَيْدِيهِمْ عَلَى عَجَلَةٍ الْقِيَادَةِ وَغَيْرِهِمْ عَلَى الطَّرِيقِ.</p> |  |
| <p>ت ا ت ا</p> |  |

Table 3.2. The model output on unseen text.

3.3. Comparative Results to Previous Work

In this section we compare FastSpeech2 with previous work by indicating the performance measure used for evaluating these models and their scores. For an objective evaluation, we chose **the mean opinion score**.

Mean Opinion Score (MOS) is the arithmetic mean over all individual values on a predefined scale that a subject assigns to his opinion of the performance of a system quality.

| Model | MOS |
|-----------------|--------------------|
| GT | 4.30 ± 0.07 |
| Tacotron 2 | 3.70 ± 0.08 |
| Transformer TTS | 3.72 ± 0.07 |
| FastSpeech | 3.68 ± 0.09 |
| FastSpeech 2 | 3.83 ± 0.08 |

Table 3.3. The MOS with 95% confidence intervals.

The above table indicates that FastSpeech 2 is the best choice to implemented in this project.

References

- [1] Sengupta N, Sahidullah M, Saha G. Lung sound classification using cepstral-based statistical features. *Comput Biol Med.* 2016 Aug 1;75:118-29. Epub 2016 May 22. PMID: 27286184.
URL:<https://www.sciencedirect.com/science/article/abs/pii/S0010482516301263>
- [2] Sejdić E.; Djurović I.; Jiang J. (2009). "Time-frequency feature representation using energy concentration: An overview of recent advances". *Digital Signal Processing.* 19 (1): 153–183.
URL:<https://www.sciencedirect.com/science/article/abs/pii/S105120040800002X>
- [3] Stevens, Stanley Smith; Volkman; John & Newman, Edwin B. (1937). "A scale for the measurement of the psychological magnitude pitch". *Journal of the Acoustical Society of America.* 8 (3): 185–190.
URL:https://archive.ph/20130414065947/http://asadi.org/jasa/resource/1/jasman/v8/i3/p185_s1
- [4] Halabi, Nawar (2016). *Modern Standard Arabic Phonetics for Speech Synthesis* (PDF) (PhD Thesis). University of Southampton, School of Electronics and Computer Science.
URL:<http://en.arabicspeechcorpus.com/Nawar%20Halabi%20PhD%20Thesis%20Revised.pdf>