Khazanah Journal of Religion and Technology
Universitas Islam Negeri Sunan Gunung Djati Bandung, Indonesia
https://journal.uinsgd.ac.id/kjrt
Online ISSN: 2987-6060
Volume 2 Issue 2 (2024)
DOI: https://doi.org/10.15575/kjrt.v2i2.1099

Text-to-Speech Technology Development Using FastSpeech2 Algorithm for the Story of the Prophet

Muhammad Raihan Firdaus
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
mraihanf11@gmail.com

Pancadrya Yashoda Pasha
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
pancadrya25@gmail.com

Muhammad Rihap Firdaus
Department of Informatics
UIN Sunan Gunung Djati Bandung
Bandung, Indonesia
muhammadrihap448@gmail.com

Received: November 19, 2024; Revised: December 19, 2024; Published: January 6, 2025

Abstract—With the SDGs target point 4.6 for 2030, literacy is a very important thing to improve. With today's technological advancements, improving the accessibility of reading in the digital age is becoming increasingly important, especially for individuals with time constraints. Text-to-Speech (TTS) technology allows users to enjoy text content, such as books or journals, in audio format, which can be listened to while doing other activities. This research develops a TTS model based on the FastSpeech2 algorithm, a non-autoregressive deep learning architecture that utilizes Transformers to generate high-quality audio quickly and efficiently. The LJSpeech dataset, which consists of 13,100 audio chunks with a total duration of 24 hours, is used as the training base. The preprocessing process involves text normalization, audio feature extraction, and data synchronization, while evaluation is performed using objective metrics such as Mel Cepstral Distortion (MCD) and Pitch Error to ensure the quality of the results. The results show that FastSpeech2 can provide fast and accurate performance in generating synthesized voices, making it potential to be used in various audio literacy applications. A key application of this TTS technology is in narrating the stories of the Prophets, which are essential in Islamic teachings for imparting moral values, fostering spiritual connection, and offering timeless lessons. The results show that FastSpeech2 is able to produce high-quality audio quickly, making it an effective alternative for improving audio literacy and providing a solution for individuals with limited reading time.

Keywords- Deep Learning, FastSpeech2, Literacy Reading, Prophet Story, Text-to-Speech

I. Introduction

Reading is something that will not escape from human life. Both from the time humans wrote on stones such as inscriptions to the time when writings were written on paper and made into books. Reading has an impact on a person's emotional and intellectual development. Reading proficiently increases one's chances of success and broadens one's horizons intellectually [1].

The stories of the Prophets hold immense significance in Islamic teachings as they provide moral guidance, spiritual inspiration, and practical lessons for daily life. These narratives emphasize virtues such as patience, gratitude, honesty, and resilience, which remain relevant across generations. By understanding the trials and triumphs of the Prophets, individuals can find solutions to personal challenges and gain a deeper connection to Islamic values. Integrating these stories into modern technologies, such as text-to-speech (TTS) systems, offers an innovative way to preserve and disseminate this knowledge, making it accessible to a global audience. Leveraging advanced algorithms like FastSpeech2 to develop TTS technology tailored for the stories of the Prophets ensures that these timeless lessons can be conveyed with clarity and engagement, particularly to younger generations and those with visual impairments. This approach not only promotes Islamic heritage but also bridges the gap between tradition and technology, enriching spiritual learning in the digital era.

Based on Sustainable Development Goals (SDGs) point 4.6, literacy and numeracy rates of all countries in the world must be improved. Meanwhile in Indonesia, based on data from UNESCO, the percentage of reading interest in this country is only 0.001%. This number is very low, which is likened to only 1/1000 people who have a high interest in reading [2], [3]. Actually, among this percentage, there are people who are actually interested in reading but time is an obstacle for them. Usually, reading requires special and free time. While busy work becomes an obstacle because it is difficult to get free time and even if there is, they are too tired to focus on reading.

While it is difficult for many people to find time to read, many people have started to multitask or do several jobs at the same time. For example, we often find people listening to the radio or music while doing other tasks, such as cooking, cleaning, driving and even doing manual labor.

Research has been conducted with the object of employees that by listening to music while working, the results of work are better than those employees who do not listen to music while working [4]. Apart from listening to music, some people's alternative when doing activities is to listen to podcasts.

Back to the discussion about reading, with the many habits of people doing activities while listening to something, there is certainly a chance that people will listen to the contents of books, journals, or other writings while doing other activities or work. With the existence of technology also called Text-to-Speech, we can enjoy or consume writing or reading into audio. In fact, listening to audiobooks can also be considered literacy because in essence it is the content or content of the reading that is important. Audiobooks and physical books certainly have the same content and only differ in format [5].

Based on the previous discussion, a Text-to-Speech (TTS) model will be built to accommodate people who want to keep reading something but do not have the free time to do so. With this model, people will be able to “read” by listening to the TTS model speak through a given text file while doing other activities. An example could be reading a report while driving or relaxing and listening to a PDF file of a journal or book. This model can also help people who prefer to learn through audio rather than through writing or reading directly. The developed TTS model is based on FastSpeech2, which uses the Transformer algorithm in the Deep Learning architecture.

II. Related Works

This research is certainly not the first time research has been conducted on Text-to-Speech (TTS) modeling. There have been many studies out there on this topic. One of them is to synthesize a child's voice from TTS. The challenge in that research was quite big because most of the TTS research results are adult voices and also the speech datasets that are widely available are also for adults. The research resulted in a 5-second long children's voice audio sample [6].

Transformers can also be used in the application of TTS. In research with TTS using Transformer [7], the methods used are techniques such as diagonal constraint, layer normalization, and pre-net bottleneck. The dataset used itself is VCTK which contains 44 hours of speech from 108 speakers and LibriTTS dataset which contains 586 hours of speech from 2456 speakers. The result of the research is a high quality TTS that implements multi-speaker voice because the dataset used is from many types of voice.

More specifically, there is research based on the same model that will be used in our research, namely using FastSpeech2. In that research, the model was further developed so that the TTS synthesized voice could match the emotion of the processed text. The result was the development of FastSpeech2 in the form of an extension called EmoSpeech. With this extension, the natural voice results sound more emotional than the usual FastSpeech2 TTS [8].

In another study, the TTS model was trained with FastSpeech2 and HiFi-GAN with an alignment module. Experiments were conducted with LJSpeech corpus containing 24 hours of recorded speech with a 22.05 kHz sampling rate. The results after being compared with ESPNet2-TTS using the same corpus dataset, this research model gets better results [9].

In addition to several previous studies, there are also more advanced studies that are used as references for this research. In this research [10], pitch prediction is applied which makes the pitch of the TTS result better. It can also be proven by the generation of high-quality mel-spectrograms. It can be concluded that this research effectively mimics the natural modulation of voice, making it a valuable contribution in the field of TTS.

The implementation of FastSpeech2 is not limited to using English, research shows that FastSpeech2 has been used for the development of text-to-speech systems in other languages such as Tibetan [11], [12], Vietnamese [13], [14], [15], Mongolian [16], Indian [17], Turkish [18], Arabian [19], Japanese [20] and Korean [21]. This proves the flexibility and adaptability of the FastSpeech model to the diverse and unique linguistic characteristics of each language, including differences in phonetics, intonation, and sentence structure. This application makes a significant contribution in extending the reach of text-to-speech technology to regions with languages that have their own technical challenges.

III. Research Methods

A. Dataset

This research uses the LJSpeech Dataset. The LJSpeech dataset contains 13,100 short audio snippets of a single speaker reading 7 non-fiction books with a total audio duration of about 24 hours [20]. Each chunk varies in length from 1 to 10 seconds.

This dataset consists of 2 main parts, namely audio files and metadata files. The audio file has a WAV extension type. While the metadata file contains the name of the audio file along with the associated transcript. Both of these are needed by the model to train its ability to process text data into sound. Then, this research also prepared 10 stories of Prophets to implement the model.

B. *Pre-processing*

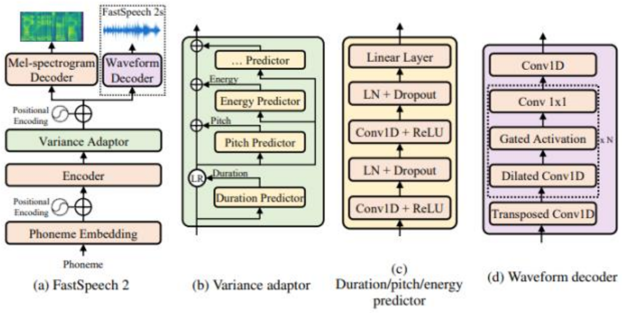
The preprocessing stage is carried out to ensure optimal audio data quality before being used in model training. From the thousands of audio data that have been collected, several main preprocessing steps are carried out as follows:

- Data cleaning, removing irrelevant, low-quality, or duplicate audio to ensure only appropriate, high-quality audio is used.
- Text normalization, converting all text to lowercase; removing irrelevant punctuation; performing short-form expansion, numbers, and symbols into words.
- Audio feature extraction, ensuring audio files are sampled at 22,050 Hz; ensuring volume is consistent; removing silent sections at the beginning and end of the video; converting audio to Mel-spectrogram with parameters (FFT size, Hop length, Window Size, Mel bands) and storing it as a numpy array.
- Synchronization of audio and text, checking to ensure that each pair of text and audio is completely matched so that no data is incorrectly connected.
- Text tokenizer, converts text into numerical representation using phoneme-based tokenizer, which converts text into phonemes to improve the model's understanding of speech.
- Divide the dataset into a Train set, Validation set, and Test set with the proportion of 80%, 10%, and 10% randomly.

C. *Classification Process*

In this study, the capabilities of the Transformer-derived architecture are utilized, namely the FastSpeech2 architecture. FastSpeech2 is an updated version of its predecessor non-autoregressive TTS method, FastSpeech.

As a derivative of Transformer, FastSpeech2 utilizes the encoder to convert the phoneme embedding sequence into the phoneme hidden sequence, and then the variant adapter adds diverse information such as duration, pitch, and energy into the hidden sequence, finally, the mel-spectrogram decoder converts the adapted hidden sequence into a mel-spectrogram sequence in parallel. FastSpeech2 uses the self attention layer, i.e. feed-forward transformer block, and 1D convolution as in FastSpeech as the basic structure for the mel-spectrogram encoder and decoder [22].



(a) FastSpeech 2 (b) Variance adaptor (c) Duration/pitch/energy predictor (d) Waveform decoder

Fig. 1. Fastspeech2 architecture

FastSpeech2 trains the model directly using ground-truth targets instead of simplified output. In addition, it also introduces more speech variation information as conditional inputs extracting duration, pitch, and energy from speech waveforms and directly using them as conditional inputs in training and using predicted values in inference. These things make the TTS development task better than before.

D. *Evaluation Process*

Once the model has been trained with FastSpeech 2, the next step is to evaluate the quality of the resulting voice using objective methods. This evaluation involves measurements such as Mel Cepstral Distortion (MCD) to see the similarity of acoustic features, Pitch Error to ensure natural intonation, and Duration Error to check the accuracy of pronunciation duration. With these metrics, the quality of the model results can be assessed consistently without involving subjectivity, thus facilitating the refinement process before widespread use.

IV. RESULT AND DISCUSSION

A. *Model Development Result*

During the course of the research, approximately four main experiments were conducted to create a model that performed reasonably well. The first experiment was conducted by utilizing the pre-trained model from FastSpeech2. The experiment was conducted with an initial training target of 10,000 steps. However, due to a setting error, the training target was missed 90,000 steps and took about 18 hours. In the experiment, the model got a total loss performance value of 36,319.7656. These results were obtained because the model has a mel loss of 0.8031, mel posnet loss of 0.8027, pitch loss of 35,894.8828, energy loss of 423.1924, and duration loss of 0.0861. These results are certainly not very satisfying, especially when inference tests are carried out, the audio results issued are still very robotic and cannot stand for long texts. The output sound will become unclear as the audio progresses.

A second trial was then conducted utilizing the model from the first trial. With many adjustments, the second training started with a target of 10,000 steps. The training went smoothly and only took about 2 hours. Below are the performance values from the second experiment:

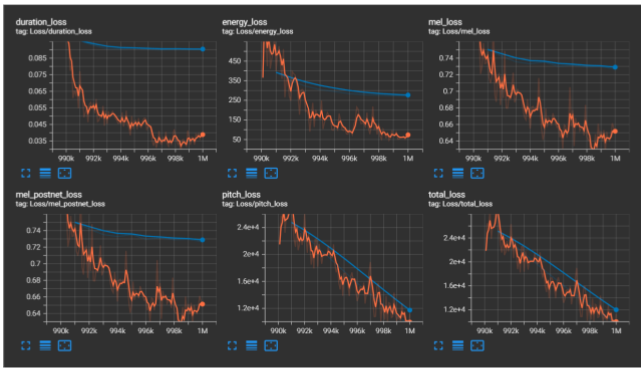


Fig. 2. Loss function from the second experiment

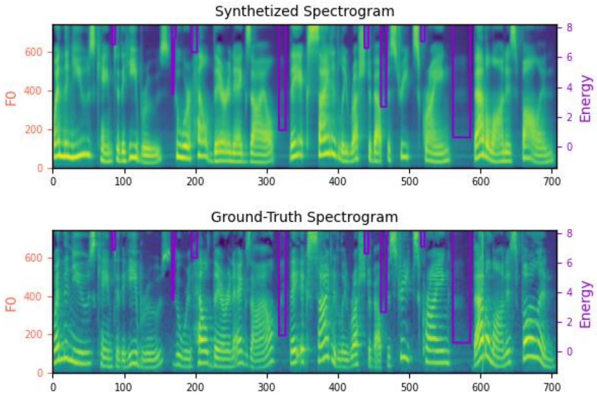


Fig. 3. Spectrogram from the second experiment

There is a significant improvement, especially in the pitch loss value. The results show that the model has a total loss of 11,310.1807, with mel loss 0.6498, mel posnet loss 0.6495, pitch loss 11,219.2598, energy loss 89.5812, and duration loss 0.0401. This is a breath of fresh air for better modeling. In the spirit of further improving the performance of the model, a third experiment was conducted. The various settings that existed in the second experiment were still used.

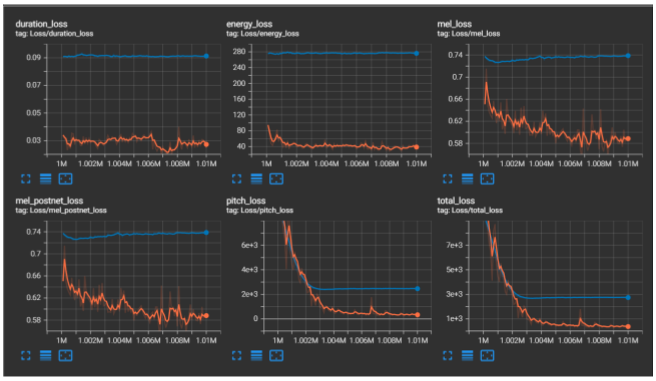


Fig. 4. Loss function from the third experiment

Just like before, there was a sharp decrease in loss in the pitch results. This contributes to the overall loss reduction. The results show that the model has a total loss of 328.3768; mel loss of 0.5893; mel posnet loss of 0.5889; pitch loss of 289.8077; energy loss of 37.3661; and duration loss of 0.0247. As seen in the mel spectrogram image, the pattern of the synthesized audio waveform has started to approach the original. Still, with high curiosity, the fourth experiment was finally conducted. All settings were left the same with the assumption that it would produce a better model like the results of the previous experiments.

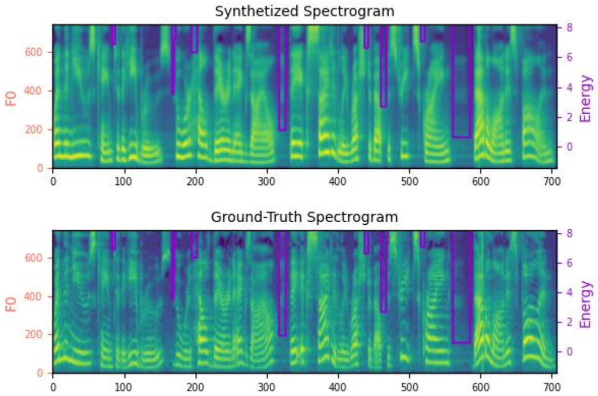


Fig. 5. Spectrogram from the third experiment

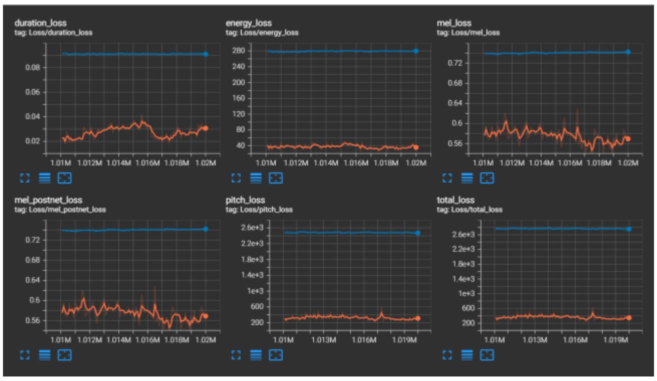


Fig. 6. Loss function from the fourth experiment

In this last experiment, the model showed results with a total loss of 370.4195, consisting of mel loss 0.5667, mel posnet loss 0.5662, pitch loss 338.4276, energy loss 30.8274, and duration loss 0.0317. There is no significant change compared to the results of the previous experiment, so it is concluded that the model performance has reached stability. Therefore, this experiment is the last experiment, and the resulting model will be used as the final model for further implementation.

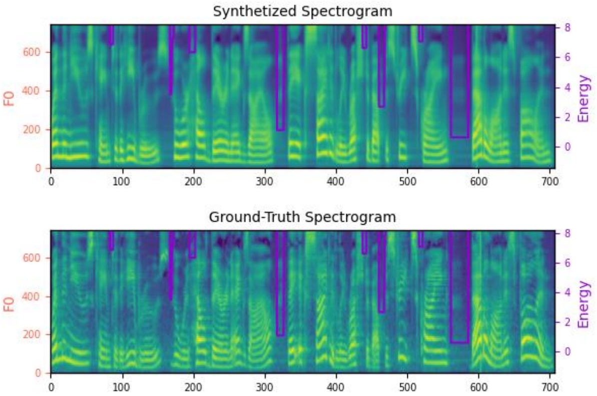
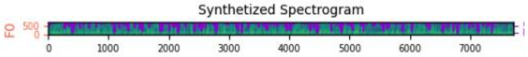
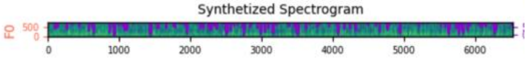
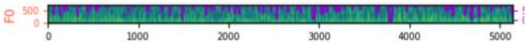


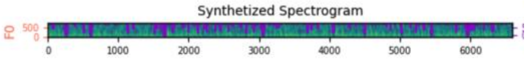
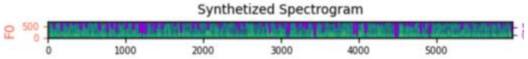
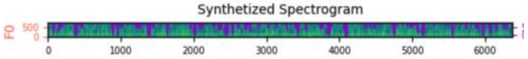
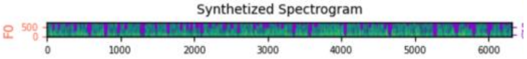
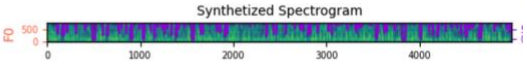
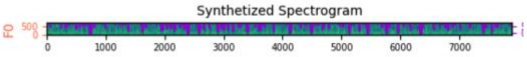
Fig. 7. Spectrogram from the fourth experiment

B. Implementation Result

The implementation of the FastSpeech2 model for the 10 stories of the Prophets involves a systematic process to ensure high-quality audio generation tailored for this specific content. First, the textual content of the stories is curated, ensuring it aligns with authentic Islamic sources. This ensures the synthesized audio captures the solemnity and narrative flow appropriate for the stories of the Prophets. The model generates Mel-spectrograms, which are converted into audio using a vocoder such as HiFi-GAN to produce natural-sounding speech. Post-processing involves quality checks using metrics like Mel Cepstral Distortion (MCD) and subjective evaluation to assess clarity, fluency, and emotional engagement. The resulting audio files provide an accessible, engaging medium for learning the stories of the Prophets, catering to diverse audiences, including those with visual impairments or limited literacy. This approach bridges traditional Islamic education with modern technology, enriching spiritual and educational experiences. Table 1 shows the results of the spectrogram for each Prophet story.

Table 1 Prophet Story Experiment

Name of Prophet	Story and Spectrogram
Zulkifli	Zulkifli or the original name Basyar was a son of Prophet Job a.s. The complete story of Prophet Zulkifli from birth to death is estimated to have lived in 1500 or 1425 BCE. The real name of Prophet Zulkifli a.s. is Basyar the son of Prophet Ayyub a.s. bin Amush bin Tawakh bin Rum bin Al-Aish bin Ishaq a.s. bin Ibrahim a.s.. At the time of his prophethood, he was sent in Damascus and its surroundings. Prophet Zulkifli a.s. was sent to the Amorites in Damascus. He had two sons and died at the age of 95 in Damascus, Syria. There were disobedient people rebelling against the kingdom of Prophet Zulkifli a.s.. Then Zulkifli called upon his people to fight the rebellious people. However, his people were afraid of death so no one wanted to go to war. "O Zulkifli, we will only fight if Allah does not kill us," pleaded Zulkifli's people. Then Prophet Zulkifli a.s. prayed and Allah answered his prayer. Then Zulkifli's troops managed to defeat the disobedient people. From then on, Zulkifli's kingdom and his people lived in peace. But since the people lived long, the population of the Land of Sham became very dense. In fact, every time there were many newborn babies, while the old people did not die. Finally, the people of Prophet Zulkifli a.s. recognized their mistake. They then asked Prophet Zulkifli a.s. to pray to Allah to become an ordinary human being who could die at any time. Synthetized Spectrogram 
Yunus	Prophet Yunus was one of the prophets and messengers sent by Allah SWT to invite the Ninawa people in the Mosul area, Iraq, to believe in Allah and abandon idolatry. However, the Ninawa people rejected the Prophet Jonah's mission and remained adamant in disbelief and misguidance. So Allah revealed to Prophet Yunus that He would punish the Ninavas after three days if they did not repent. Prophet Yunus felt angry and disappointed with the attitude of his people who did not want to listen to his call. He left them without permission from Allah SWT. He boarded a ship that would sail to another place. However, in the middle of the journey, the ship experienced a severe storm that threatened its safety. The passengers of the ship decided to draw lots on who should be thrown into the sea so that the ship could be light and safe. The lot fell to Prophet Yunus three times. Prophet Yunus was willing to jump into the sea as a sacrifice. Allah SWT did not let Prophet Yunus drown in the sea. God told a whale to swallow him without harming him. Jonah entered the belly of the whale and stayed there for several days. In the belly of the whale, Prophet Yunus felt dark, cramped, and cold. He regretted his actions that left his people without the permission of Allah SWT. He prayed earnestly and sincerely to Allah SWT. Synthetized Spectrogram 
Syuaib	Allah SWT sent Prophet Shuaib who was good at preaching and had a strong stance, so he was able to speak on behalf of truth and justice. He also began his mission to call the people of Madyan to monotheism, worship Allah and abandon idol worship. He also said to always be trustworthy in society and not to reduce the measure or scale when trading. However, when Prophet Shuaib preached, many from the Madyan tribe criticized and harassed him. They also threatened to kill Prophet Shuaib and a number of people who had followed him and believed in Allah SWT. "O Shuaib! We do not understand your words. You are a weak person. If you and those people are not part of this tribe, then we will kill or expel you from Madyan." Prophet Shuaib reminded them that the wrath of Allah SWT exists, "O my people, do as you wish. I will also do as I believe. You will know later who among us will be punished and humiliated." One of them said, "You are a liar. If you are true, then let your Lord punish us." Prophet Shuaib replied, "You will find out who is actually lying!" Synthetized Spectrogram
Sholeh	Miracles of Prophet Saleh (peace be upon him) Prophet Saleh realized that the opposition of his people who demanded proof from him in the form of miracles was aimed at eliminating his influence and eroding his authority in the eyes of his people especially his followers if he failed to meet their opposition and demands. Prophet Saleh countered their opposition by demanding a promise from them if he succeeded in bringing the miracle they asked for that they would abandon their religion and worship and would follow Prophet Saleh and believe in him. And in accordance with the request and instructions of the leaders of the Tsamud, Prophet Saleh prayed to God to give him a miracle to prove the truth of his message and at the same time break the resistance and opposition of his people who were still stubborn. He begged from God with His power to create a female camel which he took out of the belly of a large rock on the side of a hill that they pointed to. Then with the permission of Allah, the Almighty, the Creator, the designated rock was split open and a she-camel came out of its stomach. Synthetized Spectrogram 
Muhammad	Muhammad was the prophet and founder of Islam. Most of his early life was spent as a merchant. At age 40, he began to have revelations from Allah that became the basis for the Koran and the foundation of Islam. By 630 he had unified most of Arabia under a single religion. As of 2015, there are over 1.8 billion Muslims in the world who profess, "There is no God but Allah,

Name of Prophet	Story and Spectrogram
	and Muhammad is his prophet". In his early teens, Muhammad worked in a camel caravan, following in the footsteps of many people his age, born of meager wealth. Working for his uncle, he gained experience in commercial trade traveling to Syria and eventually from the Mediterranean Sea to the Indian Ocean. In time, Muhammad earned a reputation as honest and sincere, acquiring the nickname "al-Amin" meaning faithful or trustworthy. In his early 20s, Muhammad began working for a wealthy merchant woman named Khadijah, 15 years his senior. She soon became attracted to this young, accomplished man and proposed marriage. He accepted and over the years the happy union brought several children. Not all lived to adulthood, but one, Fatima, would marry Muhammad's cousin, Ali ibn Abi Talib, whom Shi'ite Muslims regard as Muhammad's successor. 
Ishaq	Short Story of Prophet Ishaq Son of Prophet Ibrahim Prophet Ishaq we know as the ninth prophet after prophet Ismail and the same son of Prophet Ibrahim AS. The story of Prophet Ishaq begins after Allah SWT granted Ismail to Prophet Ibrahim A.S. Then, Prophet Ibrahim continued to pray to Allah SWT to be granted a child from his wife named Sarah, a wife who was always faithful with him in upholding the sentence of Allah. Allah then answered Ibrahim's prayer and sent several angels in human form to convey the good news to him that a child would be born from his wife named Sarah. They also informed him of their other purpose, which was to go to the people of Luth to inflict punishment on them. Prophet Ibrahim A.S was someone who always honored guests and was also a generous person. Then, not long after, Prophet Ibrahim came with a fat calf that had been roasted and served it to them, but they neither ate nor drank the banquet that he had served. With such conditions, Prophet Ibrahim was afraid then the angels calmed him and immediately they told Prophet Ibrahim about themselves and informed their intentions and objectives to convey good news to him with the birth of a pious child. 
Isa	In Islam, Prophet Isa is one of the Prophets who has the title Ulul Azmi. This is because Prophet Isa had extraordinary patience in going through various trials in his life. How not, the birth of the Prophet Isa from Maryam had become a subject of gossip in his environment, because the Prophet Isa was born without a father. Prophet Isa was born by Maryam without a father with Allah's permission, making many people accuse Maryam of adultery and then giving birth to Prophet Isa. With this event, Allah then gave Prophet Isa the first miracle, namely being able to speak as a baby as a help given by Allah for Prophet Isa and Maryam. In addition to these miracles, Prophet Isa also obtained other miracles that God granted to Prophet Isa. When Prophet Isa was 30 years old, he was given a revelation by God in the form of the book of the gospel as a complement to the book of the Torah that had existed before, and a prophecy about the descent of the Koran to Prophet Muhammad. In addition, Prophet Isa was also blessed with various miracles as a form of help from God. Some of the miracles that God gave to Prophet Isa were bringing down food from the sky, healing lepers, even blindness, and many other miracles of Prophet Isa. 
Idris	Prophet Idris was the sixth descendant of Prophet Adam who was born in Egypt. Prophet Idris' real name was Khanukh, while his main teacher was Shis. Prophet Idris is a prophet who has extraordinary intelligence, he is the first human who managed to tame the horse and make it as a mount. Prophet Idris tamed the horse to help him in various activities, especially lifting goods, with intelligence and what was done by Prophet Idris many people were amazed at Prophet Idris. With the discovery and also the intelligence of Prophet Idris made his people more prosperous in life because it was helped by various inventions, one of which was a way to tame horses. Quoted from the book Hikmah Kisah Nabi dan Rasul by Ridwan Abdullah Sani and Muhammad Kadri, (2021: 32) for the story of Prophet Idris itself there is actually not much information in the Quran. One of the verses that explains the story of Prophet Idris is Surah Maryam verses 56 and 57. "And tell (Muhammad) the story of Idris in the Book (Al-Qur'an). Verily he was a lover of truth and a prophet, and We have raised him to a high dignity." (Q.S Maryam 56-57). 
Harun	Prophet Aaron (Harun) a.s. was one of the 25 prophets who had faithful behavior. Prophet Aaron a.s. was the older brother of Prophet Musa a.s. During his lifetime, Prophet Aaron a.s. faithfully accompanied Prophet Musa a.s. when preaching before King Fir'aun. Prophet Aaron a.s. was known to always accompany Prophet Moses a.s. in every da'wah and his efforts to liberate and guide the Children of Israel to a new land. Prophet Aaron a.s. was chosen to be a prophet as well as a companion of Prophet Moses a.s. in every da'wah, because he had a sharp tongue. His words that came out of his mouth were firm and difficult to be broken by his interlocutors. Prophet Aaron a.s. lived around 1531-1408 B.C. He was appointed prophet around 1450 B.C.. Like Prophet Moses, Prophet Aaron was sent to preach to Pharaoh in Sina, Egypt. 
Ayyub	The genealogy of Prophet Ayyub (peace be upon him) was the grandson of Prophet Ishaq bin Ibrahim (peace be upon him). He was a Prophet who had an extraordinary level of patience (the highest) in facing the trials of life from Allah Subhanahu wa ta'ala. Prophet Ayyub Alaihissalam was a wealthy man, his wealth was abundant and his livestock was very much. Yes, life was prosperous and prosperous, Prophet Job's life was filled with pleasure, but he remained diligent in worship. He also liked to do good and liked to share with anyone, everyone praised the kindness, sincerity, and sincerity of Prophet Job in doing good, even the angels also praised him. This, made the devil feel jealous and spiteful, he did not like there were humans who were so pious, the devil also intended to make Prophet Ayyub become misguided. The devil continued to try to tempt Prophet Job's faith to go astray and disbelieve and not be grateful to Allah SWT. But apparently the devil failed, the devil did not give up, he and his helpers then began to invade the faith of Prophet Job, first they killed all the livestock, then they damaged Prophet Job's garden, and also burned all his wealth. But Prophet Job and his wife's children, remained diligent in worship and never complained, they all accepted fate with sincerity. Iblis and his servants then came to Prophet Job's sons and daughters at home, they shook the pillars of the house, so that it collapsed and all of Prophet Ayyub's children died. 

C. Discussion

The results indicate that FastSpeech2 demonstrates strong potential in performing Text-to-Speech tasks. Despite the development limitations, the model achieves a relatively good loss value. The synthesized audio exhibits acceptable pronunciation, sounding sufficiently natural and not overly robotic, while effectively handling relatively long texts. However, the model still struggles with a high level of pitch errors, which can be attributed to the short training duration of only 120,000 steps. This is significantly lower compared to pre-trained models typically trained for 900,000 steps. The evaluation of the generated mel spectrograms further highlights areas of improvement, particularly in terms of pitch consistency.

This analysis operates under the assumption that training the model for 120,000 steps would yield satisfactory results for basic Text-to-Speech tasks, even though it falls short of the extensive training duration of pre-trained models.

Additionally, it is assumed that the subjective evaluation of pronunciation quality provides a reliable initial measure, focusing on naturalness and the ability to handle long input texts.

When applied to the narration of 10 selected stories of the Prophets, the model demonstrated performance that aligns well with its general evaluation results. The synthesized audio effectively captured the natural flow and solemnity required for storytelling, making the narratives engaging and accessible to the audience. However, some challenges were observed, such as occasional errors in reading abbreviations and numbers, which slightly disrupted the listening experience. These issues highlight the need for additional fine-tuning to enhance the model's accuracy in handling specific text elements. Despite these limitations, the overall harmony between the model's performance and the delivery of the Prophets' stories underscores its potential for educational and spiritual applications.

Based on the results, it is hypothesized that extending the training duration closer to that of pre-trained models and further tuning hyperparameters would significantly enhance the model's ability to handle pitch-related challenges. Moreover, incorporating a larger and more diverse dataset during training is likely to improve the naturalness of pronunciation and further reduce errors. Future work could explore these directions to refine the model's performance and address its current limitations.

V. Conclusion

This research successfully developed a Text-to-Speech (TTS) model based on FastSpeech2 to improve reading accessibility, especially for time-constrained individuals. By utilizing the LJSpeech dataset and preprocessing techniques such as Mel-Spectrogram Generation and Feature Alignment, the model demonstrated the ability to produce high-quality audio with reasonably good pronunciation and not too robotic. When applied to the narration of 10 stories of the Prophets, the model effectively delivered engaging audio outputs that aligned with the solemnity and significance of the content. Although minor issues were observed, such as errors in reading abbreviations and numbers, the overall performance highlights the model's capability for religious and educational applications.

The process of developing the model through several stages of experimentation showed a significant improvement in performance, although there are still challenges such as high pitch error due to limited training duration. Nevertheless, the resulting model is reliable enough to be used in various audio literacy applications, including reading long texts in audio format.

With the results obtained, the model shows great potential for overcoming reading constraints in modern society, providing an alternative form of audio-based content consumption. In addition, this model can also be used as a foundation for further research, such as the development of emotion features or improving the quality of voice synthesis through more intensive training. In addition, if sufficient datasets are available, it would be great if this model could be applied to other languages.

References

[1] J. Baba and F. Rostam Affendi, “Reading Habit and Students’ Attitudes Towards Reading: A Study of Students in the Faculty of Education UiTM Puncak Alam,” *Asian Journal of University Education*, vol. 16, no. 1, p. 109, Apr. 2020, doi: 10.24191/ajue.v16i1.8988.

[2] Organisation for Economic Co-operation and Development, “PISA 2022 Results: Factsheets Indonesia,” oecd.org. Accessed: Mar. 04, 2024. [Online]. Available: https://www.oecd.org/publication/pisa-2022-results/country-notes/indonesia-c2e1ae0e/

[3] OECD (Organisation for Economic Co-operation and Development), “Program from International Student Assessment (PISA) Result from PISA 2018,” 2018.

[4] S. Serpian, S. F. Alzah, J. Jusnawati, and K. Handayani, “Music at Workplace: Is it trully Improving Employees’ Performance?,” *Jurnal Office*, vol. 8, no. 2, p. 369, Mar. 2023, doi: 10.26858/jo.v8i2.44749.

[5] E. Tattersall Wallin, “Reading by listening: conceptualising audiobook practices in the age of streaming subscription services,” *Journal of Documentation*, vol. 77, no. 2, pp. 432–448, Dec. 2020, doi: 10.1108/JD-06-2020-0098.

[6] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, “A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis,” *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.

[7] M. Chen *et al.*, “MultiSpeech: Multi-Speaker Text to Speech with Transformer,” Jun. 2020.

[8] D. Diatlova and V. Shutov, “EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech,” Jun. 2023.

[9] D. Lim, S. Jung, and E. Kim, “JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech,” Mar. 2022.

[10] A. Lancucki, “Fastpitch: Parallel Text-to-Speech with Pitch Prediction,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Jun. 2021, pp. 6588–6592. doi: 10.1109/ICASSP39728.2021.9413889.

[11] Q. Zhou, X. Xu, and Y. Zhao, “Tibetan Speech Synthesis Based on Pre-Traind Mixture Alignment FastSpeech2,” *Applied Sciences*, vol. 14, no. 15, p. 6834, Aug. 2024, doi: 10.3390/app14156834.

[12] B. Zu, R. Cai, Z. Cai, and Z. Pengmao, “Research on Tibetan Speech Synthesis Based on Fastspeech2,” in *2022 3rd International Conference on Pattern Recognition and Machine Learning (PRML)*, IEEE, Jul. 2022, pp. 241–244. doi: 10.1109/PRML56267.2022.9882187.

[13] Y. Hu, P. Yin, R. Liu, F. Bao, and G. Gao, “MnTTS: An Open-Source Mongolian Text-to-Speech Synthesis Dataset and Accompanied Baseline,” in *2022 International Conference on Asian Language Processing (IALP)*, IEEE, Oct. 2022, pp. 184–189. doi: 10.1109/IALP57159.2022.9961271.

[14] N. Le Minh, A. Q. Do, V. Q. Vu, and H. T. K. Vo, “TTS - VLSP 2021: The NAVI’s Text-To-Speech System for Vietnamese,” *VNU Journal of Science: Computer Science and Communication Engineering*, vol. 38, no. 1, Jun. 2022, doi: 10.25073/2588-1086/vnucsce.347.

[15] Z. Qiao, J. Yang, and Z. Wang, “Multi-Feature Cross-Lingual Transfer Learning Approach for Low-Resource Vietnamese Speech Synthesis,” in *Proceedings of the 2023 3rd International Conference on Artificial Intelligence, Automation and Algorithms*, New York, NY, USA: ACM, Jul. 2023, pp. 175–180. doi: 10.1145/3611450.3611476.

[16] K. Liang, B. Liu, Y. Hu, R. Liu, F. Bao, and G. Gao, “Comparative Study for Multi-Speaker Mongolian TTS with a New Corpus,” *Applied Sciences*, vol. 13, no. 7, p. 4237, Mar. 2023, doi: 10.3390/app13074237.

[17] I. Gupta and H. A. Murthy, “E-TTS: Expressive Text-to-Speech Synthesis for Hindi Using Data Augmentation,” 2023, pp. 243–257. doi: 10.1007/978-3-031-48312-7_20.

[18] T. M. Koçak and M. Büyükzincir, “Building a Turkish Text-to-Speech Engine: Addressing Linguistic and Technical Challenges,” in *2023 24th International Conference on Digital Signal Processing (DSP)*, IEEE, Jun. 2023, pp. 1–4. doi: 10.1109/DSP58604.2023.10167970.

[19] M. K. Ben Mna and A. Ben Letaifa, “Exploring the Impact of Speech AI: A Comparative Analysis of ML Models on Arabic Dataset,” in *2023 IEEE Tenth International Conference on Communications and Networking (ComNet)*, IEEE, Nov. 2023, pp. 1–8. doi: 10.1109/ComNet60156.2023.10366659.

[20] M. Ikeda and K. Markov, “FastSpeech2 Based Japanese Emotional Speech Synthesis,” in *2024 IEEE 12th International Conference on Intelligent Systems (IS)*, IEEE, Aug. 2024, pp. 1–5. doi: 10.1109/IS61756.2024.10705252.

[21] Y. Choi, J. H. Jang, and M. W. Koo, “A Korean menu-ordering sentence text-to-speech system using conformer-based FastSpeech2,” *Journal of the Acoustical Society of Korea*, vol. 41, no. 3, pp. 359–366, 2022, doi: 10.7776/ASK.2022.41.3.359.

[22] Y. Ren *et al.*, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” Jun. 2020.