End to End Text to Speech Synthesis for Malay Language using Tacotron and Tacotron 2

Azrul Fahmi Abdul Aziz[1], Sabrina Tiun[2], Noraini Ruslan[3]
Center for Artificial Intelligence Technology, National University of Malaysia, Bangi, Malaysia[1, 2]
Environmental Management & Conservation Research Unit, Universiti Tun Hussein Onn Malaysia, Pagoh, Malaysia[3]

***Abstract*—Text-to-speech (TTS) technology is becoming increasingly popular in various fields such as education and business. However, the advancement of TTS technology for Malay language is slower compared to other language especially English language. The rise of artificial intelligence (AI) technology has sparked TTS technology into a new dimension. An end-to-end (E2E) TTS system that generates speech directly from text input is one of the latest AI technologies for TTS and implementing this E2E method into Malay language will help to expand the TTS technology for Malay language. This study involves the development and comparison of two end-to-end TTS models for the Malay language, namely Tacotron and Tacotron 2. Both models were trained using a Malay corpus consisting of text and speech and evaluated the synthesized speech using Mean Opinion Scores (MOS) for naturalness and intelligibility. The results show that Tacotron outperformed Tacotron 2 in terms of naturalness and intelligibility, with both models falling short of human speech quality. Improving TTS technology for Malay can encourage its use in a wider range of contexts.**

Keywords*—*Text to speech; end-to-end TTS; Tacotron; Tacotron 2; Malay language; artificial intelligence; mean opinion score (MOS); naturalness; intelligibility

I. INTRODUCTION

Text-to-speech (TTS) or speech synthesis technology has been under development for decades, with the goal of making a system that can turn text input into natural, expressive, and understandable human speech. Early TTS systems were built using rule-based methods, where the algorithms responsible for transforming text into speech were based on a set of linguistic rules and heuristics. As TTS research moved forward, concatenative synthesis [1][2][3] became a popular method. In this method, small pieces of speech segments were put together to make a complete sequence of speeches. Even concatenative synthesis methods worked better but requires a complex pipeline and requiring significant resources and manpower. Additionally, the synthesized audios often suffer from glitches or instability in prosody and pronunciation, leading to an unnatural sound compared to human speech. [4].

Study by [3] stated that, another traditional and yet proven way to use TTS technology is to use the Statistical Parametric Synthesis (SPSS) model to generate speech. SPSS uses less data than the concatenative model, but the sound is not natural. An example of a SPSS model is the Hidden Markov Model (HMM). With huge developments in artificial intelligence, the Deep Neural Network (DNN) method has been introduced in TTS. DNN is an improved method compared to SPSS but requires huge datasets [5]. DNN is a neural network with multiple hidden layers; learns by mapping text and speech and then predicting the spectral parameters that define the speech signal, such as frequency and spectrum, and then generating speech from the text input.

The end-to-end TTS systems were eventually developed as researchers started to use artificial neural networks to model the complexities of human speech. By learning to generate speech directly from text inputs, these systems did away with the need for explicit rule-based systems and separate processing stages. A groundbreaking advancement in the field of end-to-end TTS came with the introduction of the Tacotron model, which was introduced in 2017 by [6].

Tacotron [6] and Tacotron 2 [7] are two Deep Neural Networks that use an end-to-end pipeline. Both use a sequence-to-sequence model, which eliminates the need for a complex feature extraction or alignment process in TTS compared with traditional TTS. Generally, both models employ an encoder-decoder architecture in which the encoder network processes the text input and generates a compact representation, while the decoder network generates the speech signal from the encoded representation. Both models can generate high-quality speech with natural prosody and intonation with a simple signal processing.

This paper focuses on implementation of end-to-end TTS for Malay language. The Malay language, a member of the Austronesian family, is one of the many languages spoken globally. [8]. The Malay language is widely use in Southeast Asia, particularly in Malaysia, Indonesia, Singapore, Brunei, and some other countries in the region. Approximately, 250 million individuals are estimated to be speakers of this language [9]. The basic writing system in Malay is based on the Rumi script and uses the Latin alphabet. Malay language consists of 26 Latin letters [5] and 25 phonemes [9]. The Malay language has a long history and has evolved over time because of its exposure to various languages and cultures. Today, Malay language plays an important role, especially in Malaysia, and it is used in a variety of contexts, such as education, government, media, and everyday communication. As a result, progress in TTS research using Malay language plays an important role in preserving Malay as a lingua franca language in Southeast Asia.

Numerous studies focusing on Tacotron and Tacotron 2 models for TTS applications have been carried out, predominantly targeting languages like English [6][7], Chinese [10], Spanish, Korean, Japanese, Mongolian [11], and Myanmar [12][13]. Yet, the exploration of these two models

for Malay language TTS applications remains relatively limited. Previous TTS research for the Malay language has not employed an end-to-end methodology.

This study pioneers the implementation of Tacotron and Tacotron 2 in an end-to-end TTS model for the Malay language, aiming to compare their performance in naturalness and intelligibility, extending beyond traditional confines to harness their potent capabilities for the unique linguistic intricacies of Malay.

This paper consists of six Sections where Section I is the introduction of this paper, Section II will discuss the background of studies, Section III will venture any related works done before, Section IV explained the proposed methodology. Evalaution and discussion is presented in Section V. In Section VIfig

, it was concluded that the Tacotron model surpassed the Tacotron 2 model in performance, though both models were still unable to match the quality of human voice.

II. BACKGROUND

End-to-end TTS models have taken the place of outdated statistical parametric speech synthesis (SPSS) systems based on hidden Markov models (HMMs) and deep neural networks (DNN) because of advancements in deep learning techniques [14]. Previous TTS techniques involved complex pipelines and language specific linguistic features, which were resource-intensive and often resulted in unnatural sounding audio. However, the end-to-end generative TTS models like Tacotron and Tacotron 2, simplified the speech synthesis process by utilizing a single neural network for future generation [4]. This section explains how Tacotron and Tacotron 2 were built.

Both Tacotron and Tacotron 2 models start with text processing, which involves text normalization, tokenization, and character embedding. These processes prepare the input text for speech synthesis by standardizing its format, breaking it down into smaller units, and transforming it into vector representations that capture semantic and syntactic data that helps both models generate more accurate and natural-sounding speech.

A. Tacotron

Tacotron is a text-to-speech (TTS) model that generates speech in an end-to-end manner using a sequence-to-sequence (seq2seq) framework with attention [6] which includes an encoder-decoder that uses a convolutional neural network (CNN) and a recurrent neural network (RNN) to produce linear spectrograms [11]. To convert the linear spectrograms into speech waveforms, Tacotron adopts the Griffin-Lim algorithm [15] for phase estimation, followed by an inverse short-time Fourier transform. This end-to-end TTS model is designed to generate high-quality speech directly from text. Tacotron eliminates the need for phoneme-level alignment, enabling it to efficiently scale with large volumes of acoustic data and accompanying transcripts. With the capacity to be trained entirely from scratch using random initialization, Tacotron represents a significant advancement in the TTS domain. Tacotron has the capability to be trained entirely from scratch, given a set of paired text and audio. The Tacotron's block diagram, shown in Fig. 1, consists of a pre-processing unit, encoder, decoder, and vocoder.

The text is first processed and embedded, and then transformed using a pre-net to reduce overfitting and improve training stability. The encoded characters are then fed into a series of encoder blocks, where a CBHG (Convolutional Banks, Highway Networks, and Gated Recurrent Units) module which contains 1-D convolutional banks, max pooling, a 4-layer highway and a bidirectional GRU is used to convert the pre-net outputs into the final encoder representation. Meanwhile, the attention mechanism computes the context vector using the outputs of the text encoder and the previous decoder state, enabling the model to focus on different parts of the input sequence at each decoding step. In the decoder block, a pre-net, attention RNN, decoder RNN, and CBHG module are used to focus on relevant parts of the input text and align speech signal frames.

Finally, the generated Mel-scale spectrograms from the RNN decoder's previous outputs are input into a CBHG module to correct prediction errors for each frame. From here, linear spectrograms can be predicted. The vocoder block uses Griffin Lim to convert the linear spectrograms into speech waveforms as the output.

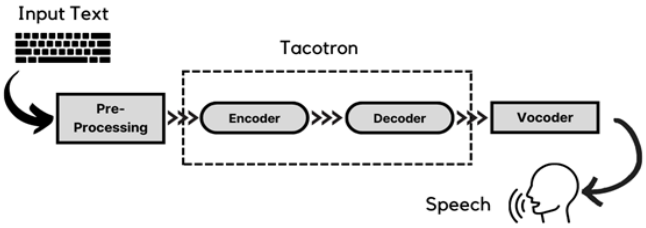

Fig. 1. Tacotron block diagram.

B. Tacotron 2

Tacotron 2 is a deep neural network that generates speech directly from text, similar to its predecessor Tacotron. Fig. 2 summarizes the Tacotron 2 block diagram. The network utilizes a recurrent sequence-to-sequence model that maps character embeddings to mel-scale spectrograms, which represent the frequency content of audio signals. The encoder processes the character embeddings using convolutional layers to generate encoded features, followed by a bidirectional LSTM. The decoder, which is an autoregressive recurrent neural network, uses a location-sensitive attention network to predict a Mel spectrogram from the encoded input sequence. The decoder also includes pre-net and attention context vectors, which are passed through a stack of unidirectional LSTM layers to predict the target spectrogram frame. The predicted Mel spectrograms are then passed through a post-processing network consisting of CNN layers and a linear projection layer to generate the final Mel spectrograms. Finally, the Mel spectrograms are converted into sound waveforms using either Griffin Lim [13] or Wavenet [7]. Other than Griffin Lim, wavenet is also one of the vocoders that are used to convert mel-spectograms to speech sounds [16]. Wavenet can learn how to match acoustic properties to speech waveforms on a sample-by-sample basis. In conclusion, the Tacotron 2 model is capable of generating speech with high

naturalness and intelligibility and has the potential to be used in various speech synthesis applications [7][13].

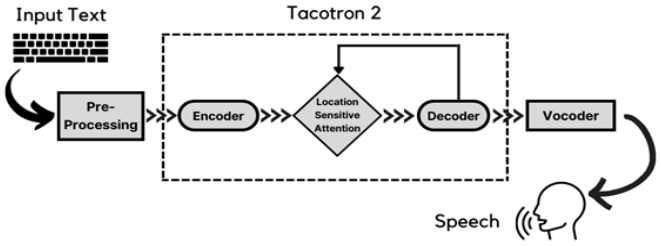

Fig. 2. Tacotron 2 block diagram.

III. Related Work

This section will discuss the development of TTS for the Malay language and previous work done on other languages using Tacotron and Tacotron 2.

A. Development of TTS for Malay Language

Even though TTS technology is rapidly evolving, especially in the English language, Malay language TTS technology however still has untapped potential. Several studies have been done on TTS for the Malay language, including studies that suggest a text to speech system based on Hidden Markov Models (HMM) and context-dependent labels to produce high-quality synthetic speech in Malay language [17]. The use of context-dependent labels is a common technique in TTS that allows for more natural-sounding speech by taking the surrounding linguistic context into account. This study developed a grapheme-to-phoneme database and identified the contextual factors of Malay language to generate the labels. This sequence of labels is then fed into the HMM-based acoustic model, which generates the corresponding speech waveform. The score of intelligibility and naturalness of the synthetic utterances were evaluated to gauge the module's effectiveness.

Another research for Malay language TTS involves using a bilingual voice synthesis system for Standard Malay and Indonesian using a hybrid method of Hidden Markov Model and Deep Neural Network (HMM-DNN) [5]. This study combines the corpus of these two similar languages Malay and Indonesian and introduces speaker codes to examine the bilingual speech synthesis system, comparing it with the monolingual speech synthesis system. The method involves training a hybrid system on a large corpus of speech data using HMMs to model the spectral and duration information, and DNNs to model the acoustic mapping function. The effectiveness of the TTS synthesis was assessed using the MOS score for naturalness of synthesis sound.

Both studies demonstrate that improvements are being made to the TTS system for the Malay language. However, utilizing the most recent TTS technology will help to advance TTS generally in the Malay language. Thus, the introduction of an end-to-end TTS method for the Malay language will enable the development of Malay TTS technology more quickly.

B. Tacotron

One of the famous studies for Tacotron is from [6]. This research focuses on the English language as the input data. The model started with the data collection. The researcher collects a dataset of speech and text pairs using an internal North American English dataset that contains 24.6 hours of speech data. The text data is then processed by normalizing each character. The normalized text and speech data will go to a Tacotron model that contains an encoder and decoder, and the output will be a linear spectrogram. Finally, the linear spectrogram will be synthesized using the Griffin Lim synthesizer to create a speech sound from the designated input text. The quality of the outcome is evaluated based on a MOS score test, in which eight raters evaluate approximately 100 unseen phrases. This study compared the Tacotron MOS to the concatenative and parametric approaches and discovered that the Tacotron method is superior to the parametric method.

Research on Tacotron models is also available in Myanmar. The official language of Myanmar is Burmese, a Sino-Tibetan language that is written and spoken in Myanmar [12]. Like the research from [6], it starts with data collection. For this project, nearly 5000 pairs of text and audio from male speakers and 3000 pairs from female speakers were used. After that, the text is normalized. The Tacotron model will then be used with both normalized text and sound. The linear spectrogram created from the Tacotron model will then be synthesized using the Griffin Lim synthesizer to create an audio sound in Myanmar. The results are compared using the MOS method by comparing the MOS on naturalness and intelligibility of the Tacotron model with the original audio from the recorded speech. Around 20 speech outputs were used for this evaluation, with 5 people as evaluators.

C. Tacotron 2

One of the studies on Tacotron 2 done by [7] focuses on the English language. Data from a single female speaker's 24.6 hours of speech were used to begin this study's data gathering. The text will then be processed by normalizing each text and going for character embedding. The pair of processed text and sound is then inserted into the Tacotron 2 model, which consists of an encoder and decoder to create a Mel spectogram. The Mel spectogram is then synthesized using a wavenet synthesizer to create an audio sound from the input text. Similar to Tacotron [6] the subjective evaluation is based on the MOS score from eight raters with 100 unseen phrases. By contrasting the MOS with several TTS techniques, including parametric, Tacotron with Griffin Lim, concatenative, Wavenet, and ground truth audio, which is an authentic human voice, the synthesis audio from the naturalness of the Tacotron 2 model was evaluated. For intelligibility evaluation, this paper runs a MOS evaluation by generating 37 news headlines and comparing the MOS score on the Tacotron 2 model with the Wavenet model. In addition to the subjective assessment, the authors also conducted an ablation investigation to examine the effects of various model components on the performance of the entire system. These investigations aid in their comprehension of the significance of various model components, such as the attention process and WaveNet's conditioning on Mel spectrogram predictions.

The research report by [13] also addresses Tacotron 2. The paper focuses on the Myanmar language as a test subject. From a variety of sources, the researchers created a corpus of Myanmar speech which contains over 5000 pairings of

Myanmar's text and speech, with each audio pair's length ranging from 2 to 12 seconds. They then used a syllable segmenter and text normalizer to separate the Myanmar text into characters. Next, the researcher used a recurrent seq2seq network to map character embeddings to Mel-scale spectrograms. Finally, they used the Griffin-Lim algorithm to produce Myanmar speech output from the input text. The MOS score was used to assess the synthesis's quality. This study examined Tacotron and Tacotron 2 MOS scores for naturalness and intelligibility in the Myanmar language.

IV. PROPOSED METHOD

The journey of end-to-end TTS started with the speech corpus creation. For both models, audio and text from a single Malay speaking male speaker are paired together to make a dataset. The dataset was downloaded through the Speech Malaya website [18]. The datasets consist of text corpus files and wav files for audio. Both text and audio files are paired and numbered accordingly. The text corpus contains over 6445 lines of Malay sentences that were taken from the audio context, and the text corpus was saved in CSV format. The text corpus is a complete sentence ending with a full stop. For the audio file, it was a wav file recorded at a sampling rate of 24 kHz with a total duration of 14.29 hours. The audio data were cut into small sizes with minimum and maximum durations of 1.752 seconds and 21.24 seconds respectively, to make a total of 6445 pairs of text and audio datasets in Malay language.

The text will be fed into the pre-processing channel for text normalization, tokenization, and character embedding. In this paper, the setting for text normalization for Malay words is under the "malay_cleaners" configuration inside the hyperparameter. Table I is an example of text normalization in which an input text is normalized by converting numbers to text, lowercasing the input text, and removing any punctuation marks.

TABLE I. EXAMPLE OF MALAY TEXT NORMALIZATION

No	INPUT TEXT	NORMALIZED TEXT
1	Daripada jumlah tersebut seramai 2,359 iaitu 44.6 % orang ibu tunggal	daripada jumlah tersebut seramai dua ribu tiga ratus lima puluh sembilan iaitu empat puluh empat perpuluhan enam peratus orang ibu tunggal
2	Pada tahun 2010 jumlah pinjaman perumahan yang diluluskan oleh sistem perbankan adalah sebanyak	pada tahun dua ribu satu puluh jumlah pinjaman perumahan yang diluluskan oleh sistem perbankan adalah sebanyak
3	SOALAN 33 Dr Mansor Bin Abd Rahman minta MENTERI PERDAGANGAN ANTARABANGSA DAN INDUSTRI	soalan tiga puluh tiga doktor mansor bin abd rahman minta menteri perdagangan antarabangsa dan industri

From Table I, in the input text number 1 in Malay, "Daripada jumlah tersebut seramai 2,359 iaitu 44.6 % orang ibu tunggal" which means "Out of that number 2,359 44.6% are single mothers" in English, is being normalized into "daripada jumlah tersebut seramai dua ribu tiga ratus lima puluh sembilan iaitu empat puluh empat perpuluhan enam peratus orang ibu tunggal" in Malay, which translate to "out of that number two thousand three hundred and fifty nine forty four point six percent are single mothers" in English.

In Malay input text number 2, "Pada tahun 2010 jumlah pinjaman perumahan yang diluluskan oleh sistem perbankan adalah sebanyak" which translate to, "In 2010 the number of housing loans approved by the banking system was*"* in English. It is then normalized into "pada tahun dua ribu satu puluh jumlah pinjaman perumahan yang diluluskan oleh sistem perbankan adalah sebanyak" in Malay which is translated into English "In two thousand and ten the number of housing loans approved by the banking system was*"*.

From input text number 3 Malay, "SOALAN 33 Dr Mansor Bin Abd Rahman minta MENTERI PERDAGANGAN ANTARABANGSA DAN INDUSTRI" which translated into English "QUESTION 33 Dr Mansor Bin Abd Rahman asked the MINISTER OF INTERNATIONAL TRADE AND INDUSTRY" was normalized into "soalan tiga puluh tiga doktor mansor bin abd rahman minta menteri perdagangan antarabangsa dan industry" in Malay and the English translation is "question thirty three doctor mansor bin abd rahman asked the minister of international trade and industry". All the input text from Table I is being normalized by converting numbers to text, lowercasing the input text, and removing any punctuation marks.

A. Tacotron

After the text is normalized and tokenized, 256-dimensional character embeddings are applied to Malay texts. Subsequently, the embedded text is fed into an encoder pre-net. The encoder CBHG module processes the encoder pre-net output to produce the final encoder representation that the attention module will utilize. As previously mentioned, the decoder comprises the decoder pre-net, attention RNN, decoder RNN, and CBHG (post-net CBHG). Finally, the CBHG synthesizes the linear spectrogram, and the Griffin Lim synthesizer with a power of 1.5 is used to convert it into a sound wave. Table II presents all the parameters.

B. Tacotron 2

In comparison to the Tacotron model, the Tacotron 2 model uses character embeddings of size 512, which are fed into a stack of three convolutional layers. The final convolutional layer's output is used as input for a single bidirectional LSTM layer to generate the encoded features for the model. The Tacotron 2 model utilizes the location-sensitive attention network, which computes location features using 32 1-D convolution filters of length 31. The decoder block includes a pre-net consisting of two fully connected layers with 256 hidden ReLU units that is followed by a stack of two unidirectional LSTM layers, each with 1024 units, and a linear projection to predict the target spectrogram frame. To enhance overall reconstruction of the output, the predicted mel spectrogram is passed through a 5-layer convolutional post-net, which predicts a residual. During inference, the model uses stop token prediction, where the concatenation of the decoder LSTM output and the attention context is projected to a scalar and passed through a sigmoid activation to predict the probability that the output sequence has completed. To synthesize the waveforms, the Griffin Lim synthesizer is used to generate sound waveforms. Unlike the Tacotron model, the

Tacotron 2 model does not employ CBHG stacks and GRU recurrent layers. All parameters can be found in Table III.

TABLE II. MALAY TACOTRON HYPER-PARAMETER

HYPER-PARAMETER NAME	HYPER-PARAMETER VALUE
Audio Parameter	number_mels= 80; number_freq=1025, sample_rate=20000, frame_length_ms=50, frame_shift_ms=12.5, pre-emphasis=0.97
Character Embedding	256-D
Encoder Parameter	Encoder Pre-net : FC-256-ReLU → Dropout(0.5) →FC-128-ReLU → Dropout(0.5) Encoder CBHG : Conv1D bank: K=16; conv-k-128-ReLU; Max pooling: stride=1, width=2; Conv1D projections=conv-3-128-ReLU→ conv-3-128-Linear; Highway net=4 layers of FC-128-ReLU; Bidirectional GRU=128 cells
Attention RNN	1-layer GRU (256 cells)
Decoder Parameter	Decoder Pre-net : FC-256-ReLU → Dropout(0.5)→FC-128-ReLU → Dropout(0.5) Decoder RNN : 2-layer residual GRU (256 cells)
Post-net CBHG	Conv1D bank: K=8, conv-k-128-ReLU; Max pooling: stride=1, width=2; Conv1D projections: conv-3-256-ReLU →conv-3-80-Linear; Highway net: 4 layers of FC-128-ReLU; Bidirectional GRU: 128 cells

TABLE III. MALAY TACOTRON 2 HYPER-PARAMETER

HYPER-PARAMETER NAME	HYPER-PARAMETER VALUE
Audio Parameter	sampling_rate=24000; n_mel_channels=80; ffilter_length=1024; hop_length=200; windows_length=800
Character Embedding	512-D
Encoder parameter	encoder_embedding_dim=512; encoder_kernel_size=5; encoder_n_convolutions=3
Attention Parameter	attention_dim=128; attention_rnn_dim=1024
Location Layer Parameter	attention_location_n_filters=32; attention_location_kernel_size=31
Decoder parameter	n_frames_per_step=1; decoder_rnn_dim=1024; prenet_dim=256; max_decoder_steps=1000; gate_threshold=0.5; p_attention_dropout=0.1; p_decoder_dropout=0.1

C. Training

Training of the data is conducted based on the two models, utilizing the Malay text corpus. Both models use the same data, as explained in previous sections. There are some differences in the training parameters of both models. Table IV summarizes the training parameter summaries. Tacotron 2 requires a smaller batch size than Tacotron because it employs a more complicated representation. A large batch size will cause the processing time to take longer and sometimes cause the program to crash. From Table IV also, Tacotron 2 requires a lengthy run time with nearly the same total steps. Steps for checkpoint intervals are steps where a predicted spectrogram is generated as an output after certain steps. Fig. 3 and Fig. 4 display the alignment plot results for both models at specific checkpoint intervals.

TABLE IV. TRAINING PARAMETER DIFFERENCES

TRAINING PARAMETER	TACOTRON	TACOTRON 2
Batch Size	32	28
Total Step Run (Steps)	95850	80000
Total Hours Run (Hours)	59.42	136.33
Steps for checkpoints intervals (Steps)	150	2000

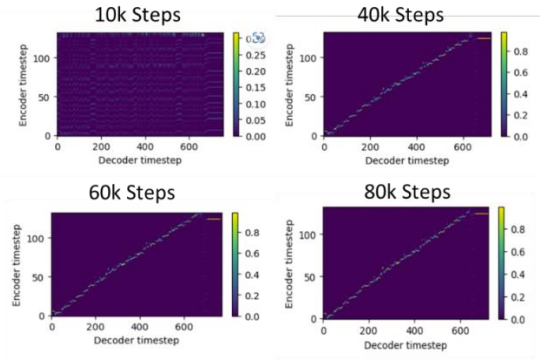

Fig. 3. Alignment for certain steps for Tacotron.

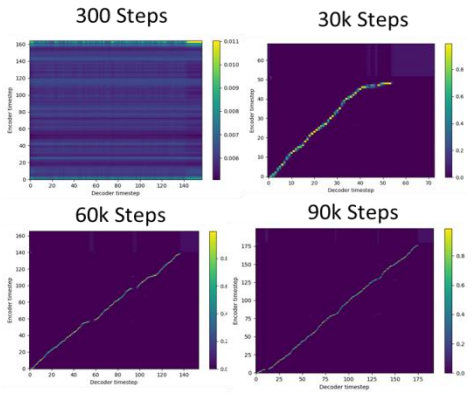

Fig. 4. Alignment for certain steps for Tacotron 2.

An alignment graph shows the visual representation of the learned attention mechanism between the input text and the output and also shows how the model aligns each input character with the corresponding output acoustic features. The graph should have a clear diagonal pattern to indicate that the model is aligning input characters with output features correctly. If it doesn't, it may be having difficulty or not being trained properly. The fig. 3 and 4 also demonstrate that the diagonal pattern on the graph does not appear until after a certain number of training steps.

V. Evaluation and Discussion

A. Measurement Evaluation

To evaluate the effectiveness of the proposed method, a selection of approximately 20 Malay sentences were randomly sampled from an online news portal, as well as several sentences from Malaysian Hansard parliament speeches. These Malay sentences will then be synthesized using the Tacotron and Tacotron 2 models to get the audio speech. By using these 20 Malay sentences, an original human sound was recorded. For both synthesized models, this human voice serves as another point of comparison. The input word is taken from outside of the training data, which is why the original human voice is being used. Around 5 native speakers will then evaluate the naturalness and intelligibility of human sound, synthesized speech from Tacotron, and synthesized speech from Tacotron 2.

This experiment was based on subjective evaluations. According to [19], a MOS served as a gauge of the effectiveness for TTS, following guidance from the International Telecommunication Union (ITU-T P.85, 1994). To evaluate overall sound quality, a 5-point scale was used, with 5 being the highest quality and 1 being the lowest. Around 5 native speakers will then evaluate the naturalness and intelligibility of human sound, the synthesized speech from Tacotron, and the synthesized speech from Tacotron 2 follow the 5-point scale. To calculate the overall individual performance, the mean calculation is used to get the final answer. A simple mean calculation is shown in equation 1, where R is individual rating and N is the total number of speech output.

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

The evaluation for this experiment is based on naturalness and intelligibility. Naturalness and intelligibility are important qualities expected from a TTS system. Naturalness is how the model produces a speech that is human alike in terms of casual, emotional, and spontaneous styles of speaking. Currently, speech recordings used in TTS training typically follow formal reading styles, as pauses, repeats, changes in speed, varying emotions, and errors are not permitted [3]. However, in casual or conversational talk, humans seldom speak in a standard reading style. Intelligibility is how the model can produce speech that can be understood by everyone. Noise is one of the factors that affects the intelligibility of the model [20]. In a real-life situation, a listener is often unable to understand what another person is saying if the environment is noisy. This can lead to some information not being conveyed.

B. Result

A subjective evaluation test was conducted to compare the method including with an original human voice in Malay speech synthesis. Fig. 5 and Fig. 6 show the results of naturalness and intelligibility for the human voice, Tacotron, and Tacotron 2.

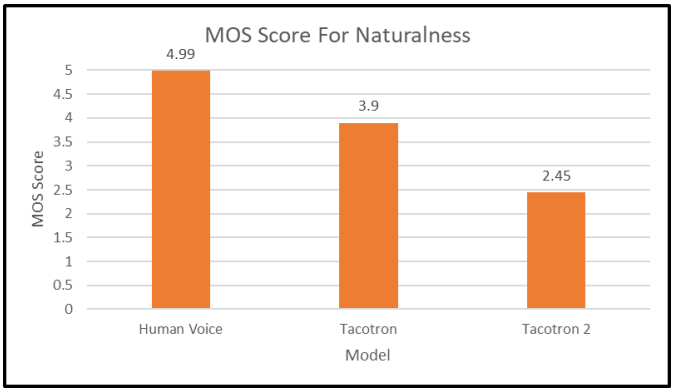

Fig. 5. Comparison of MOS score for naturalness.

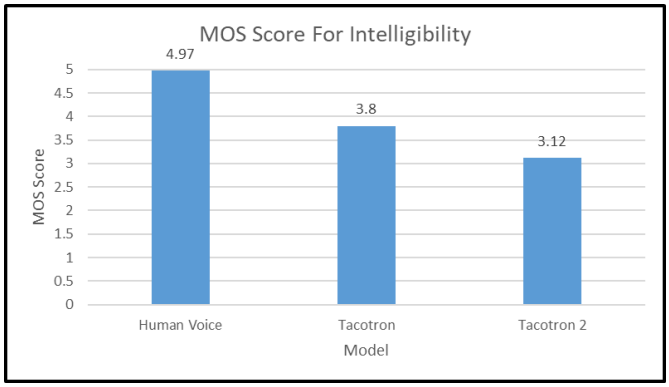

Fig. 6. Comparison of MOS score for intelligibility.

The result shows that in terms of naturalness, the Tacotron 2 model is lower compared to the Tacotron, but for intelligibility, the Tacotron 2 model has a slight improvement, though it is still considered lower compared to the Tacotron.

C. Result and Discussion

From the result however shows that the outcome of this trial contradicts other experiments conducted in [7] for English and [13] for Myanmar language, which suggest that Tacotron 2 performance exceeded the Tacotron. But in this experiment, shows that Tacotron outperformed the Tacotron 2 model. Fig. 7, shows the comparison on the alignment graph between Tacotron and Tacotron 2. The alignment graph shows that Tacotron model diagonal pattern is marginally clearer and have a very straight diagonal pattern compared to Tacotron 2 model, and Tacotron 2 model also generate a longer decoder timesteps. From the previous explanation, a clear diagonal graph indicates that the model is aligning input characters with output features correctly.

This might be for a few reasons, one of which is that the model might not have learned the correct alignment between input text and output speech yet. Most likely, the model needs more training epochs or the hyperparameters need to be changed. Another reason might be issues with the attention mechanism. The model might be struggling to learn the appropriate alignment between the input and output sequences. Overall, the attention mechanism parameter needs to be adjusted to improve the quality of the Tacotron 2 model.

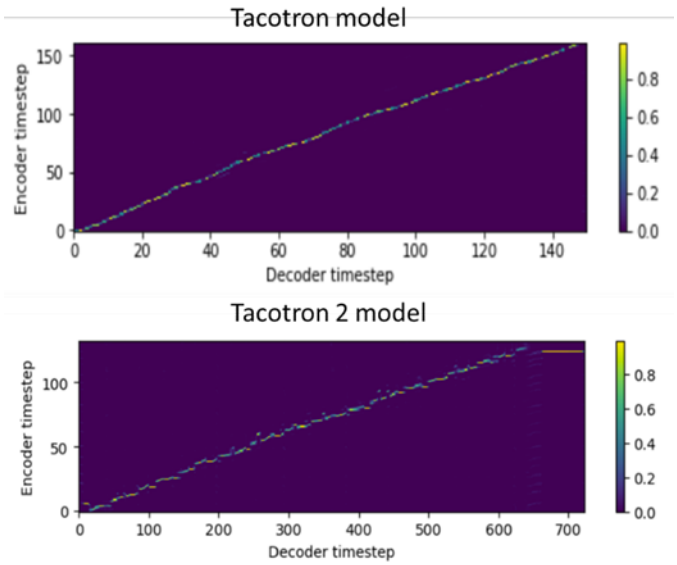

Fig. 7. Comparison alignment graph between Tacotron and Tacotron 2.

VI. CONCLUSION

The purpose of this study is to provide an end-to-end TTS technique for the Malay language, hence contributing to the development of Malay language technology. In conclusion, this study shows that Tacotron and Tacotron 2 can both translate Malay text into speech, but the Tacotron model performs better when compared to Tacotron 2. In the future, to improve the Tacotron 2 model, some hyperparameters will need to be adjusted, especially in the attention and decoder parameter. Additionally, there could be a contemplation of the use of WaveNet as a vocoder in Tacotron 2. Given its flexibility, the WaveNet vocoder allows high user adaptability in manipulating synthesized speech to suit various scenarios, thereby enhancing overall performance. To further explore end-to-end TTS methods for the Malay language, other end to end models such as FastSpeech, Transformer TTS, and Parallel WaveGAN can be considered. This approach could provide increased flexibility in the application of end-to-end techniques for Malay language.

ACKNOWLEDGMENT

The authors acknowledge that part of this research is supported by Universiti Kebangsaan Malaysia under GUP grant with grant number: GUP-2020-063.

REFERENCES

[1] J. H. Andrew and W. B. Alan, “Unit selection in a concatenative speech synthesis system using a large speech database,” In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, vol. 1, pp. 373-376. IEEE, 1996.

[2] G. Xavi, T. Siamak, C. Chun-an, B. Markus, G. Alexander, and S. Hanna, “Recent advances in Google realtime HMM driven unit selection synthesizer,” 2016.

[3] X. Tan, Q. Tao, S. Frank, and L. Tie-Yan, “A survey on neural speech synthesis,” arXiv preprint arXiv:2106.15561 2021.

[4] L. Naihan, L. Shujie, L. Yanqing, Z. Sheng, and L. Ming, “Neural Speech Synthesis with Transformer Network,” In Proceedings of the AAAI conference on artificial intelligence, vol. 33, no. 01, pp. 6706-6713. 2019.

[5] F. Chen, J. Yang, and L. Zhao, “A Bilingual Speech Synthesis System of Standard Malay and Indonesian Based on HMMDNN,” in Proceedings of the 2020 International Conference on Asian Language Processing (IALP), pp. 181–186, IEEE, Kuala Lumpur, Malaysia 2020.

[6] Y. Wang et al., “Tacotron Towards end to end speech synthesis,” in Proc. Interspeech, Aug. 2017, pp. 4006–4010 2017.

[7] J. Shen et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4779-4783. IEEE, 2018.

[8] C. Adrian and D. David, “Standard Malay (Brunei),” Journal of the International Phonetic Association 41(2). 259–268 2011.

[9] M.A Asyafie, M. Harun, M.I Syapiai and P.I Khalid, “Identification of Phoneme and Its Distribution of Malay Language Derived From Friday Sermon Transcripts”. In 2014 IEEE Student Conference on Research and Development, pp. 1-6. IEEE, 2014.

[10] Y. Zhang et al., “Learning to speak fluently in a foreign language Multilingual speech synthesis and cross language voice cloning,” arXiv preprint arXiv:1907.04448 2019.

[11] J. Li, H. Zhang, R. Liu, X. Zhang, and F. Bao, “End-to-end mongolian text to speech system,” In 2018 11th international symposium on chinese spoken language processing (ISCSLP), pp. 483-487. IEEE, 2018.

[12] Y. Win, H. Pyae Lwin, and M. Masada, “Myanmar Text to Speech System based on Tacotron End to End Generative Model,” In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 572-577. IEEE, 2020.

[13] Y. Win and T. Masada, “Myanmar text to speech system based on Tacotron 2,” In 2020 International Conference on Information and Communication Technology Convergence (ICTC), pp. 578-583. IEEE, 2020.

[14] T. Hayashi et al., “Espnet2 tts Extending the edge of tts research,” arXiv preprint arXiv:2110.07840 2021.

[15] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” IEEE Trans. Acoust., Speech, Signal Process., vol. 32, no. 2, pp. 236–243, Apr. 1984.

[16] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, “Speaker-dependent WaveNet vocoder,” in Proc. Interspeech, vol. 2017, 2017, pp. 1118–1122 2017.

[17] B. Mustafa Mumtaz, Z. M. Don, and G. Knowles, “Context dependent labels for an HMM based speech synthesis system for Malay HMM based speech synthesis system for Malay,” n Computational Science and Technology: 5th ICCST 2018, Kota Kinabalu, Malaysia, 29-30 August 2018, pp. 205-214. Springer Singapore, 2019.

[18] Z. Husein, "Malaya-Speech: Speech-Toolkit library for Bahasa Malaysia, powered by Deep Learning Tensorflow," GitHub, 2020. https://github.com/huseinzol05/malaya-speech.

[19] M. Viswanathan and M. Viswanathan, “Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale,” Comput. Speech Lang., vol. 19, no. 1, pp. 55–83, 2005.

[20] D. Paul, M. P V Shifas, Y. Pantazis, and Y. Stylianou, “Enhancing speech intelligibility in texttospeech synthesis using speaking style conversion,” 2008.