P-ISSN: 2715-2448 | E-ISSN: 2715-7199
Vol.6 No.2 July 2025
Buana Information Technology and Computer Sciences (BIT and CS)

Development of a Deep Learning-Based Text-To-Speech System for the Malang Walikan Language Using the Pre-Trained SpeechT5 and Hifi-GAN Models

Aina Avrilia Imani[1], Aviv Yuniar Rahman[2], Firman Nurdiyansyah[3]
[1 2 3] Department of Informatic Engineering, Universitas Widya Gama, Malang, Indonesia
E-mail: avriliaaina@gmail.com[1], aviv@widyagama.ac.id[2], firmannurdiyansyah@widyagama.ac.id[3]

Received: 2025/06/07 | Revised: 2025/06/27 | Accepted: 2025/07/27

Abstract

The Walikan language of Malang is a form of local cultural heritage that needs to be preserved in the digital era. This study aims to develop and evaluate a deep learning-based Text-to-Speech (TTS) system capable of generating speech in the Walikan language of Malang using pre-trained SpeechT5 and HiFi-GAN models without fine-tuning. In this system, SpeechT5 is used to convert text into mel-spectrograms, while HiFi-GAN acts as a vocoder to generate audio signals from the mel-spectrograms. The dataset used consists of 1,000 sentences in the Walikan language of Malang. The system evaluation was carried out using objective metrics of Word Error Rate (WER) and Character Error Rate (CER), by comparing the results of synthetic audio transcriptions against two types of reference audio, namely the original voices of female speakers and male speakers, using the Automatic Speech Recognition (ASR) system. The female voice was recorded with controlled articulation, while the male voice used natural intonation in everyday conversation. The results show that synthetic audio has the highest error rate with a WER of 0.9786 and a CER of 0.9024. Meanwhile, female audio has a WER of 0.5471 and a CER of 0.1822, while male audio shows a WER of 0.6311 and a CER of 0.2541. These findings indicate that the TTS model without fine-tuning is not yet capable of producing synthetic voices that can be recognized accurately by the ASR system, especially for regional languages that are not included in the initial training data. Therefore, the fine-tuning process and the preparation of a more representative dataset are important so that the TTS system can support the preservation of the Walikan Malang language more effectively in the digital era.

Keywords: *HiFi-GAN, Malang Walikan Language, SpeechT5, Text-to-Speech, and Word Error Rate.*

I. Introduction

The Walikan language of Malang is a local cultural heritage that deserves preservation, especially in today's digital era, where regional languages are increasingly threatened with extinction. Deep learning-based Text-to-Speech (TTS) technology has developed rapidly and is one solution for language preservation by converting text into natural speech. However, most existing TTS models focus on resource-intensive languages, making them less than optimal for low-resource regional languages like Walikan.

Research on TTS for regional languages is still very limited, especially in cases where there is little or no training data (zero-shot). Walikan also has unique characteristics not found in other languages, making it crucial to test how pre-trained models like SpeechT5 and HiFi-GAN perform under these conditions without fine-tuning.

Several previous studies have used SpeechT5 and HiFi-GAN models for languages with more resources and have shown quite good results [1], [2]. However, the use of pre-trained models directly (zero-shot) in low-resource regional languages, particularly Malang Walikan, is still rarely researched.

Other studies have compared objective evaluation methods such as Word Error Rate (WER) with subjective Mean Opinion Score (MOS) assessments to assess synthetic speech quality [3].

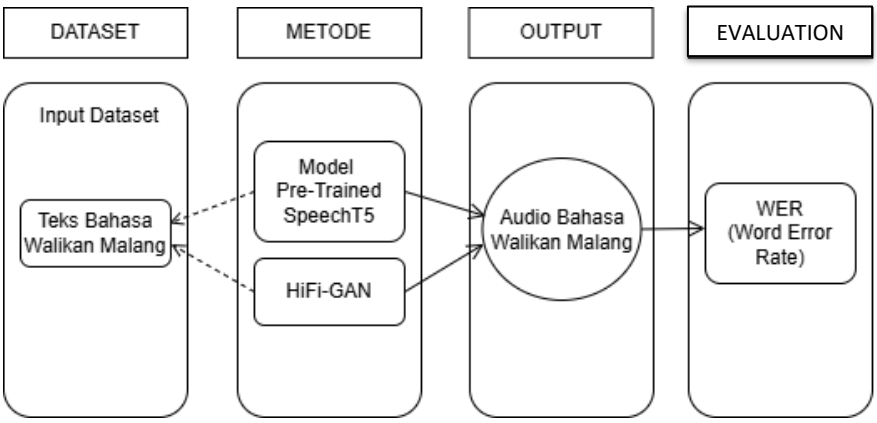
This study aims to develop and evaluate a deep learning-based TTS system using pre-trained models SpeechT5 and HiFi-GAN for Malang Walikan without fine-tuning. The evaluation was conducted using objective metrics WER and Character Error Rate (CER) by comparing the synthetic results to real human speech to assess the model's performance on low-resource languages.

In the development of deep learning-based Text-to-Speech systems, various architectures have been developed to produce voices that increasingly resemble human speech. One recent model that has demonstrated superior performance is SpeechT5 [4], a transformer-based pre-trained model with an encoder-decoder approach designed for various speech processing tasks, including speech synthesis. This model is capable of understanding text structure and converting it into a mel-spectrogram representation, taking into account intonation and rhythm. To generate the final speech signal, HiFi-GAN [5] was used, a generative adversarial network (GAN)-based vocoder designed to convert mel-spectrograms into high-quality, natural audio. The combination of these two models was found to be effective in generating synthetic speech that approximates human voice quality.

This research provides an important contribution to understanding how large pre-trained models perform on regional languages not yet represented in the training data, and provides a basis for further development through fine-tuning and the collection of a more comprehensive dataset, to support the digital preservation of the Walikan Malang language.

II. Methods

The steps in this research are shown in (Figure 1).



. Figure 1. Steps a research.
(Source: Personal Preparation)

The dataset used in this study consists of 1,000 sentences in the Walikan language of Malang. Each sentence was recorded by two different speakers, one male and one female, resulting in a total of 2,000 audio files. This dataset was used to study the linguistic patterns and acoustic characteristics of Walikan and to evaluate the system's performance on voice variations based on speaker gender.

The collected data was stored in spreadsheet format (Excel) for ease of management and processing. Before use, the data underwent pre-processing to remove irrelevant symbols or emojis and normalize the text. An example of the data used is shown in Table 1.

Table 1. Dataset Language Walikan

Language Walikan	Language Indonesian
etas maya sing cedek gang iku asaib sam rasane	*sate atam yang dekat gang itu biasa mas rasanya*
yang ayahab kuwi gluduk arudam	*yang bahaya itu petir madura*
tambah nade ae nawak kotis iki	*tambah gila aja teman satu ini*
wah umak lihai juga berbahasa arudam	*wah kamu lihai juga berbahasa madura*
rintep pol kakakku	*pintar banget kakakku*
agomes warga licek sukses halokes e halokes sek jum	*semoga warga kecil sukses sekolah nya sekolah dulu jum*
nde wendit akeh sedeb sam	*di wendit banyak monyet mas*
likis e adikku mari dicokot kucing	*kaki nya adikku habis digigit kucing*
bangga nggae soak arema	*bangga pakai kaos arema*

1. Pre-Trained SpeechT5 Model

SpeechT5 is a pre-trained model based on a transformer encoder-decoder architecture used to convert text into a mel-spectrogram, a visual representation of audio frequencies over time. This model was implemented without fine-tuning, using the default configuration from the Hugging Face library. The SpeechT5 model uses a Transformer-based encoder-decoder architecture, with six additional pre-net and post-net modules to process input and output in the form of text and speech. The encoder consists of two parts: the Speech Encoder Pre-net to convert the raw speech signal into speech features, and the Text Encoder Pre-net to convert the text into a textual representation. The outputs of these two parts are combined into a unified representation that is used by the decoder.

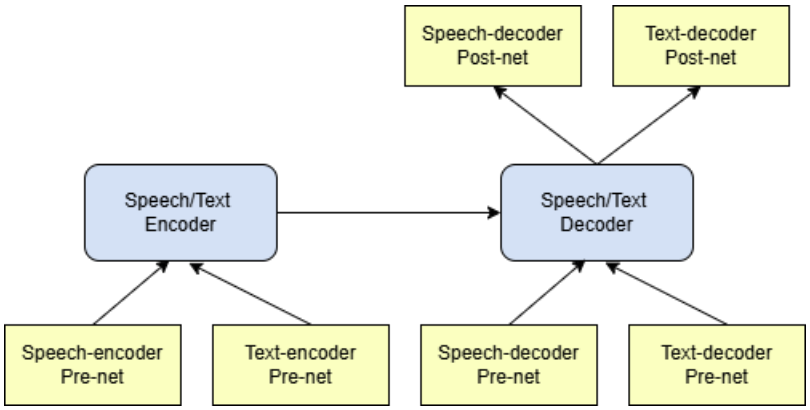


Figure 2. SpeechT5 Architecture

The decoder also has two paths: the Speech Decoder, which generates a mel-spectrogram from text input, and the Text Decoder, which generates text from voice input. The Pre-net and Post-net components in each section enhance the quality of the representation and output. The Speech Encoder pre-net utilizes features from wav2vec 2.0, while the decoder utilizes speaker embeddings (x-vectors) to support multi-speaker synthesis.

The initial pre-training process is multi-task and cross-modal, including bidirectional masking prediction, sequence-to-sequence generation, and text pre-training using an infilling strategy. The encoder output representation is discretized using vector quantization, allowing the model to align the voice and text modalities.

The implementation of the SpeechT5 model as the core of the Text-to-Speech system for the Malang Walikan language is carried out using the following steps:

a. Data Preparation and Pre-trained Model

The pre-trained SpeechT5 model is used in the Walikan language sentence dataset without fine-tuning. The audio output is evaluated using ASR and Word Error Rate (WER).

b. Text Preprocessing

1) Normalization: Standardize the text (lowercase, remove irrelevant symbols/punctuation).
2) Tokenization: Break the text into tokens (words/subwords).
3) Conversion to Numeric ID: Map tokens to unique numbers based on the vocabulary.
4) Embedding: Convert the ID tokens into a numeric representation vector that the model understands.

c. Encoding with a Transformer Encoder

The embedding vector is processed by a Transformer encoder with self-attention to produce a contextual representation of the text (hidden states).

d. Mel-spectrogram Prediction by the Decoder

The Transformer decoder converts the contextual representation into a mel-spectrogram through self-attention and cross-attention with the encoder output. The mel-spectrogram depicts sound energy at various frequencies over time.

e. Mel-spectrogram Conversion to Audio

The decoded mel-spectrogram is converted to audio using HiFi-GAN.

2. HiFi-GAN

HiFi-GAN is used as a vocoder to convert the mel-spectrogram output from SpeechT5 into a high-quality audio signal. This model plays a role in the final stage of the synthesis process, producing natural-sounding voices that approximate human voice quality. The output, in the form of a mel-spectrogram from SpeechT5, serves as input to HiFi-GAN, which then produces the final audio in WAV format.

HiFi-GAN [6] is a Generative Adversarial Network (GAN)-based vocoder designed to convert mel-spectrograms into high-quality raw audio signals. Its architecture consists of a generator and two types of discriminators: the Multi-Scale Discriminator (MSD) and the Multi-Period Discriminator (MPD) [7]. Training is conducted in an adversarial manner, where the generator and discriminator compete to produce increasingly natural sounds.

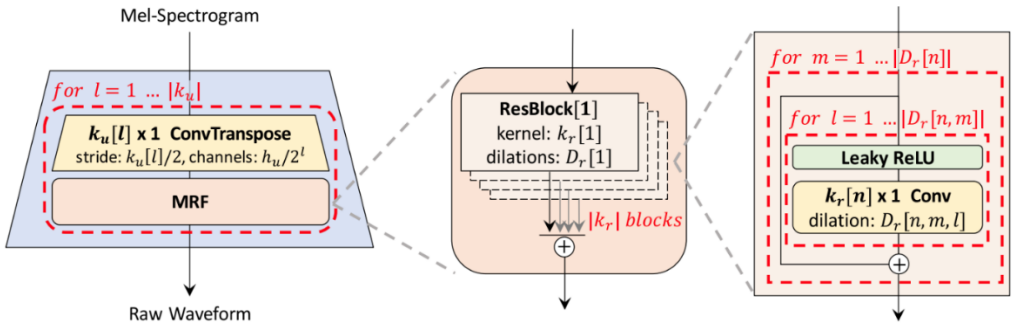


Figure 3. HiFi-GAN architecture [5]

The HiFi-GAN generator is a fully convolutional network with an upsampling process based on transposed convolution [8]. To capture complex temporal patterns, a Multi-Receptive Field Fusion (MRF) module is used, which combines multiple residual blocks with different kernels and dilations.

MPD focuses on periodic patterns in audio signals with sub-discriminators that process the input based on specific time periods (e.g., 2, 3, 5, 7, 11). Meanwhile, MSD captures patterns at multiple temporal scales, processing audio at different resolutions (raw, average pooled ×2, and ×4) to identify global characteristics of the signal.

3. System Output

The system generates synthetic audio in Malang Walikan, reflecting the phonetic and linguistic characteristics of the input text. The resulting audio can be listened to as a spoken representation of the sentences in the dataset.

4. Evaluation

The evaluation was conducted using the Word Error Rate (WER) and Character Error Rate (CER) metrics, which measure the error rate between the automatic transcription of synthetic audio and the original text. The lower the WER and CER values, the higher the system's accuracy and naturalness in producing speech that matches the text.

Evaluation of Text-to-Speech (TTS) results is performed by measuring the accuracy of the synthesized audio transcription against the reference text using two main metrics: the Word Error Rate (WER) and the Character Error Rate (CER).

a. Word Error Rate

The Word Error Rate (WER) is the primary metric for evaluating the accuracy of a Text-to-Speech (TTS) system by comparing the audio transcription output from the model with the original text. The WER is calculated based on the number of word substitutions (S), deletions (D), and insertions (I) divided by the total number of words (N) in the original text, expressed as a percentage:

$$\text{WER} = \frac{S+D+I}{N}\ x\ 100 \qquad (1)$$

If WER = 0%, the system produces perfect output that is identical to the original text. The higher the WER value, the greater the errors in the speech synthesis results [9], [10].

b. Character Error Rate

Character Error Rate (CER) is an evaluation metric similar to WER, but calculated at the character level. CER is more sensitive to small errors, especially in short sentences or languages with non-standard structures, such as Walikan. The CER formula is similar to WER, namely:

$$\text{CER} = \frac{S+D+I}{N}\ x\ 100 \qquad (2)$$

Where S, D, and I are the number of character substitutions, deletions, and insertions, and N is the total number of characters in the original text. CER complements WER by providing a more detailed error analysis at the phonetic level [11]

III. Results and Discussions

1. Word Error Rate

Synthetic audio quality was evaluated by calculating the Word Error Rate (WER) for nine representative test sentences with varying error levels. Three types of audios were analyzed: the SpeechT5 + HiFi-GAN synthesized audio, a female voice, and a male voice, to determine how well the ASR system recognized each type of voice.

Table 2. Word Error Rate.

Test Sentence	WER SpeechT5 + HiFi-GAN	WER Voice Female	WER Voice Male
info lokasi dong nawak	0.5	0.25	0.25
sesuai dengan gambar di bis halokes	0.5	0.17	0.33
wah umak lihai juga berbahasa arudam	0.67	0.5	0.83
rame ilakes area kayutangan di malam minggu	0.71	0.29	0.71
sekali nade tetep nade	0.75	0.75	0.75
agomes lancar rejeki hari ini	0.8	0.6	0.4
nakam lah mbah	1.0	0.33	1.0
umak ngalam	1.0	0.5	1.0
jenenge ae ongis nade yo nade temenan	1.0	0.57	0.71

Table 2 shows examples of WER values for the three audio types. The results indicate that all synthetic audio had WER values above 0.30, indicating a high level of pronunciation errors. Four sentences achieved a maximum WER of 1.0, meaning that all words were not recognized by ASR. The sentences with the lowest WER were "info lokasi dong nawak" and "sesuai dengan gambar di bis halokes," each with a value of 0.5.

In contrast, human voices performed better and more consistently, with some sentences achieving WERs as low as 0.25. This indicates that natural audio is more capable of conveying phonetic information accurately. The high error rate in synthetic audio is likely due to the striking phonetic differences and the model's limitations in handling the non-standard Walikan language structure and its inversions. Sentences such as "wah umak lihai juga bahasa arudam" reflect linguistic challenges for the model that has not undergone fine-tuning.

As a follow-up to this evaluation, the scatter plot in Figure 4 was used to visualize the distribution of WER values for each test sentence by voice type category. This visualization provides a more comprehensive picture of the distribution of recognition errors at the sentence level.

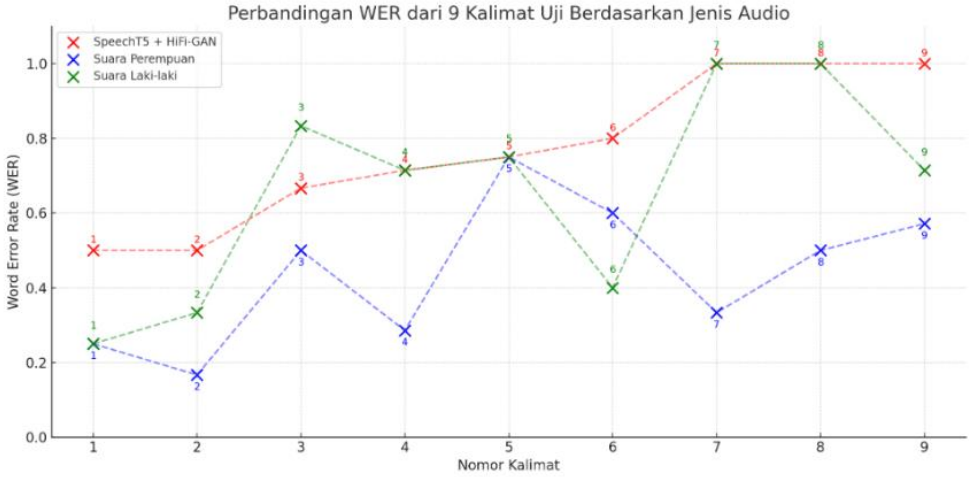


Figure 4. Scatter plot comparing WER values by audio type

Figure 4 shows a comparison of the Word Error Rate (WER) values for nine test sentences based on three audio types: the ones synthesized by the SpeechT5 + HiFi-GAN model (red/orange), the

female voice (blue), and the male voice (green). In general, the synthetic audio produced the highest and most consistent WER values, with all sentences having values above 0.5, even reaching 1.0 in the last three sentences. This indicates that the Text-to-Speech system is not yet capable of producing pronunciations close to human accuracy, especially in the complex context of Walikan. Meanwhile, human audio, particularly female voices, demonstrated better and more stable performance, with WER values ranging from 0.2 to 0.6. Male voices also tended to be more accurate than synthetic audio, but experienced greater fluctuations between sentences. This difference indicates that synthetic audio still faces significant challenges in matching the natural phonetic characteristics of human speakers.

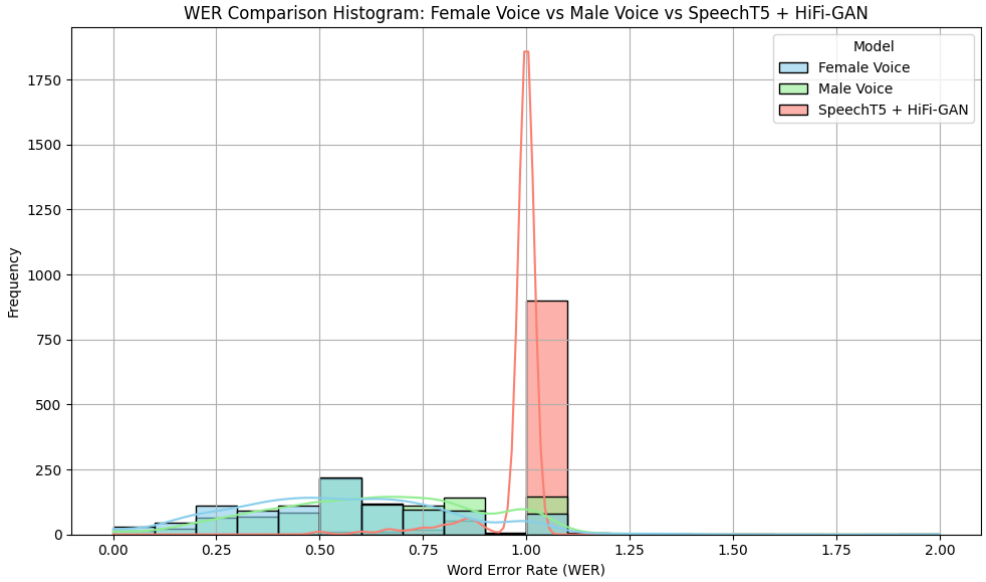


Figure 5. Histogram of WER Comparison

Figure 5 shows a comparison of the distribution of Word Error Rate (WER) values for three types of audios: female, male, and the SpeechT5 + HiFi-GAN synthesized voices. The histogram shows that the WER values of the synthesized voices are mostly close to 1.0, indicating ASR's failure to recognize the sentence content.

In contrast, human voices show a lower and more even distribution of WER values, some even approaching 0.0, indicating near-perfect transcription. This finding confirms that without fine-tuning, the SpeechT5 + HiFi-GAN model is not yet capable of producing pronunciations accurate enough for ASR to recognize.

Table 3. Average Word Error Rate

Average Word Error Rate (WER)	SpeechT5 + HiFi-GAN	Real Voice (Female)	Real (Male)
	0.9786	0.5471	0.6311

Testing 1,000 test sentences showed an average WER of 0.9786 for the synthetic audio from SpeechT5 + HiFi-GAN (Table 3), close to 1.0, indicating that the majority of words were not recognized by ASR. This reflects the low quality of the synthesis, especially in the context of Walikan, which has unusual phonetics.

In contrast, human voices performed better, with an average WER of 0.5471 (female) and 0.6311 (male), a difference of ±0.33–0.43 lower than the synthetic results. This confirms that ASR is better able to recognize natural voices.

In addition to inaccurate pronunciation, the synthetic results also showed anomalies such as random capitalization and phonetic errors due to the model's bias toward English, for example, "sam" being recognized as "some".

2. Character Error Rate

Evaluation using the Character Error Rate (CER) provides a more detailed picture of phonetic errors at the character level, especially for minor errors such as letter substitutions or spelling mistakes. Table 4 summarizes the minimum, maximum, median, and average CER values for the three audio source categories tested.

Table 4. Character Error Rate Evaluation

Voice Category	Minimum CER	Maximum CER	Median CER	Mean CER
SpeechT5 + HiFi-GAN	0.3243	1.0000	1.0000	0.9024
Voice Female	0.0000	1.6154	0.1461	0.1822
Voice Male	0.0000	1.0000	0.2143	0.2541

Character Error Rate analysis (Table 4) shows that the synthetic audio from the SpeechT5 + HiFi-GAN model performed poorly compared to human audio. The average CER was 0.9024, with a median and maximum of 1.00, indicating that many sentences failed to be recognized by ASR. Even the minimum value of 0.3243 still indicated significant errors at the character level.

In contrast, human audio was much more accurate. Female voices had an average CER of 0.1822 and male voices 0.2541, with a minimum value of 0.00 for both, indicating some sentences were perfectly recognized.

These results align with previous WER findings, indicating that errors in synthetic audio occur not only at the word level but also at the character level. This reflects the model's limitations in representing Walikan phonology without fine-tuning.

Based on the evaluation, the SpeechT5 and HiFi-GAN-based TTS systems performed poorly when used on Walikan without fine-tuning. High WER and CER values for synthetic audio indicate that ASR has difficulty recognizing the model's output, in contrast to human speech, which produces better accuracy.

Table 5. Comparison of Regional Language TTS Studies

Author	Dataset	Data Ground Truth	Method	Evaluation		Validate
				WER	CER	MOS
Agustina, C., 2024. [12]	250 Sentence Language Banjar	-	VITS	-	-	3.604
ALHUDA, M.Y., 2025. [13]	500 Sentence Language Palembang	-	VITS	-	-	4.58
Our proposed	1000 Sentence Language Walikan Malang	Voice Artificial	SpeechT5 & HiFi-GAN	0,9786	0,9024	-
		Voice Female	SpeechT5 & HiFi-GAN	0,5471	0,1822	-
		Voice Male	SpeechT5 & HiFi-GAN	0,6311	0,2541	-

Table 5 compares this research with other TTS studies on regional languages. The studies by Agustina (2024) and Alhuda (2025) used VITS with a smaller dataset and only assessed voice quality through MOS (3.604 and 4.58), without objective metrics like WER or CER. Meanwhile, this study used 1,000 Walikan sentences and objectively evaluated them using WER and CER.

The results showed that the synthetic audio had a WER of 0.9786 and a CER of 0.9024, indicating significant errors. However, the female and male voices recorded WERs of 0.5471 and 0.6311, and CERs of 0.1822 and 0.2541, indicating better ASR recognition of human voices.

The advantage of this approach lies in the use of more measurable objective metrics and the hybrid SpeechT5 + HiFi-GAN model, which has potential, although it still requires further optimization and training to handle the complexity of regional languages and natural voice variations.

IV. Conclusions

This study developed a preliminary TTS system for Malang Walikan using SpeechT5 and HiFi-GAN without fine-tuning. Although the synthesis results were not optimal, the system demonstrated basic potential for improvement through further training with local data. Evaluation using ASR-based WER and CER showed that the synthetic audio (SpeechT5 + HiFi-GAN) produced a WER of 0.9786 and a CER of 0.9024, indicating a very high error rate. Audio from a female voice had a WER of 0.5471 and a CER of 0.1822, while audio from a male voice recorded a WER of 0.631 and a CER of 0.2541. These differences indicate that voice characteristics significantly impact system performance.

The system was unable to replicate the speaker's voice characteristics, such as intonation and clarity of pronunciation, likely due to the lack of fine-tuning and a lack of adaptation to the unique structure of Walikan. However, the pronunciation of some words was quite accurate, indicating potential for further development.This research is an initial contribution to the development of TTS for undocumented regional languages, and emphasizes the importance of adapting models to local contexts for more accurate and natural results.

References

[1] J. Lehečka, Z. Hanzlíček, J. Matoušek, and D. Tihelka, "Zero-Shot vs. Few-Shot Multi-speaker TTS Using Pre-trained Czech SpeechT5 Model," pp. 46–57, 2024, doi: 10.1007/978-3-031-70566-3_5.

[2] H. Wang, "Understanding Zero-shot Rare Word Recognition Improvements Through LLM Integration," 2025, [Online]. Available: http://arxiv.org/abs/2502.16142

[3] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the MOS pit: A critical analysis of MOS test methodology in TTS evaluation," no. August, pp. 41–47, 2023, doi: 10.21437/ssw.2023-7.

[4] J. Ao et al., "SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing," Oct. 2021, [Online]. Available: http://arxiv.org/abs/2110.07205

[5] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.05646

[6] J. Su, Z. Jin, and A. Finkelstein, "HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks," Jun. 2020, [Online]. Available: http://arxiv.org/abs/2006.05694

[7] Z. Qiu, J. Tang, Y. Zhang, J. Li, and X. Bai, "A Voice Cloning Method Based on the Improved HiFi-GAN Model," Comput Intell Neurosci, vol. 2022, 2022, doi: 10.1155/2022/6707304.

[8] D. Lim, S. Jung, and E. Kim, "JETS: Jointly Training FastSpeech2 and HiFi-GAN for End to End Text to Speech," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2022-Septe, pp. 21–25, 2022, doi: 10.21437/Interspeech.2022-10294.

[9] A. Ali and S. Renals, "Word error rate estimation without asr output: E-WER2," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2020-Octob, pp. 616–620, 2020, doi: 10.21437/Interspeech.2020-2357.

[10] A. Ali and S. Renals, "Word Error Rate Estimation for Speech Recognition: e-WER," Jul. 2018. [Online]. Available: https://github.com/qcri/e-wer

[11] I. Kottayam and J. James, "Advocating Character Error Rate for Multilingual ASR Evaluation," 2023.

[12] C. Agustina, "Implementasi Teknologi Text To Speech Bahasa Banjar Menggunakan Metode Vits," (Doctoral dissertation, Universitas Islam Negeri Sultan Syarif Kasim Riau)., 2024.

[13] M. Y. ALHUDA, "Text To Speech Bahasa Palembang Menggunakan Metode Vits," (Doctoral dissertation, UNIVERSITAS ISLAM NEGERI SULTAN SYARIF KASIM RIAU)., 2025.