



Speech-Based Virtual Assistant for Mental Health Support Through Natural Interaction

Dimas Rizqi Kurniawan¹, Donny Avianto²

^{1,2}Undergraduate Program of Informatics, Faculty of Science & Technology, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia

¹rizqikurniawan1863@gmail.com, ²donny@uty.ac.id

Abstract

Mental health is a significant global concern. Indonesia has reported high rates of depression and anxiety, compounded by limited emotional outlets. Although AI virtual assistants are prevalent in e-commerce and education, their application in mental health remains underexplored. Existing solutions are predominantly text-based and transactional, which restricts empathetic and natural interactions. This study develops a voice-based assistant by integrating Automatic Speech Recognition (ASR), a generative AI for empathetic responses, and a Text-to-Speech (TTS) module fine-tuned on an Indonesian dataset to adapt accent and prosody. The system underwent both technical evaluation and human testing to assess its feasibility and user experience. The results showed that the TTS model converged effectively with low loss. Human evaluation indicated 'good' interaction (MS = 3.91, SD = 0.02), 'good' AI responses (MS = 3.83, SD = 0.26), and 'fair' TTS naturalness (MOS = 3.27, SD = 0.05). Most participants found the assistant's responses meaningful, pleasant, and helpful in managing low to moderate anxiety. These results suggest that a voice-based assistant has the potential to support mental health in Indonesia. Future work should enhance speech naturalness and utilize a larger participant pool for evaluation.

Keywords: depression rate; generative ai; speech recognition; tacotron2 waveglow; virtual assistant

How to Cite: [R. Kurniawan and D. Avianto, "Speech-Based Virtual Assistant for Mental Health Support Through Natural Interaction", *J. RESTI (Rekayasa Sist. Teknol. Inf.)*, vol. 9, no. 6, pp. 1493 - 1503, Jan. 2026.
Permalink/DOI: <https://doi.org/10.29207/resti.v9i6.6585>

Received: April 25, 2025

Accepted: December 16, 2025

Available Online: January 11, 2026

*This is an open-access article under the CC BY 4.0 License
Published by Ikatan Ahli Informatika Indonesia*

1. Introduction

Research in 2019 found that millions worldwide experience mental health problems. Specifically, 301 million have anxiety disorders and 280 million have depression [1]. In Indonesia, a 2023 Health Ministry survey sampled each province and reported 630,827 mental health cases among people aged 15 and above [2]. Current factors affecting mental health include traumatic changes, social influences, limited support from friends and family, and life pressures. As a result, many have nowhere to express their problems and keep these issues to themselves [3]. Over time, mental health can worsen, causing symptoms such as irritability, hopelessness, excessive worry, and, at worst, suicidal tendencies [4]. This underscores the urgent need for accessible, confidential, stigma-free emotional support platforms.

One potential solution is using AI virtual assistants' intelligent systems that aid communication and daily tasks. In mental health, these assistants are not just task-

focused robots; they also relieve social emotions [5],[6]. Indonesian research emphasizes other uses of AI assistants. For instance, Rosanti et al. studied Lazada's e-commerce chatbot and found it streamlined transactions [7]. However, such systems do not meet the relational needs of therapeutic dialogues. Similarly, general LLM research on ChatGPT in Indonesian academia [8] shows it helps with idea generation and writing. Though useful, local studies have gone beyond therapeutic uses. Internationally, studies have examined issues like acceptability and professional perspectives. Mental health chatbots are usually seen as acceptable [9], but they are mostly text-based and formal. This points to chances for richer, more natural options, like voice-based tools that may boost empathy and reduce cognitive demands.

Most research on virtual assistants in Indonesia focuses on education and customer service. Even in mental health, studies are limited to text-based chatbots, which restrict natural, dynamic human-AI interactions. To

address this, our work explores voice-based virtual assistants for mental health support. Few Indonesian studies use automatic speech recognition (ASR) for natural, dynamic human-AI interaction. Using speech, not just text, can simulate real conversation and open possibilities for empathetic, accessible mental health support.

This study addresses a research gap by proposing a novel Indonesian voice-based virtual assistant designed to foster more natural, empathetic dialogue for mental health support. The system combines Automatic Speech Recognition (ASR) for free-form input, Generative AI for empathetic responses, and a Text-to-Speech (TTS) module for expressive output. The end-to-end architecture bridges barriers in naturalness, latency, and conversational flow. In addition to showing feasibility

in an underrepresented language, the study evaluates system accuracy, speech quality, and user experience. Results indicate that such an assistant can lower interaction barriers in mental health contexts. The aim is to develop and assess a voice assistant using ASR, a generative model, and a fine-tuned TTS module in Indonesian to enhance naturalness, empathy, and accessibility for mental health support.

2. Methods

This research aims to develop a virtual assistant capable of interacting with humans through relevant responses and accurate pronunciation based on text transcripts. It utilizes the Generative AI Gemini model, Tacotron2, and WaveGlow, as illustrated in Figure 1.

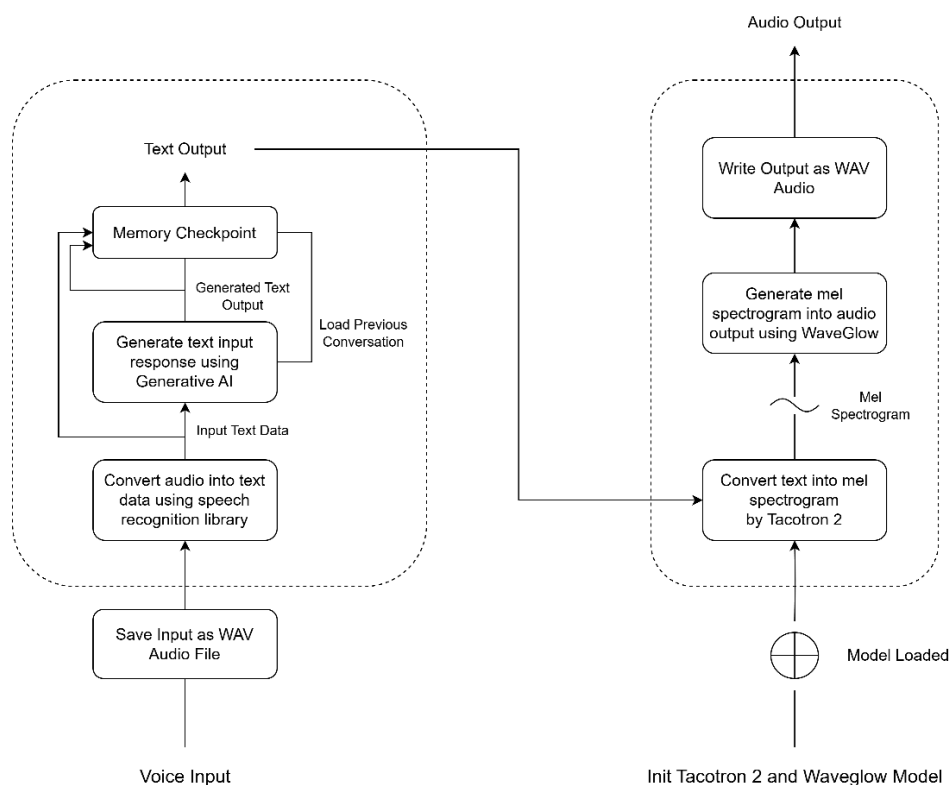


Figure 1. Proposed AI Speech Recognition Workflow

The objective is for the AI system to deliver vocal responses that are accurate and intelligible, as generated by the underlying Generative AI. The process is summarized in Figure 1: the user provides speech input, which is transcribed into text using a speech recognition library. The transcribed text is processed by Generative AI, which generates contextually relevant responses. Subsequently, Tacotron2 converts the AI-generated text into a Mel Spectrogram, and WaveGlow produces the final audio output by synthesizing sound waves based on a pre-trained audio dataset.

2.1 Audio Dataset Collection and Preprocessing

The dataset for this research comprises 300 Indonesian audio files featuring conversational samples sourced from YouTube and a pre-trained Tacotron2 model

utilizing the 13,000-sample LJSpeech dataset. The latter was used for fine-tuning to increase the diversity of model predictions. To enhance the voice clarity, Spleeter was utilized to isolate voice data from background music and noise [10].

On the two datasets, there are differences in waveforms before audio separation and after audio separation. Figure 2 shows that there are waveforms still containing background disturbance, while Figure 3 shows the refined waveforms after audio separation, resulting in a clearer and audible voice. Following the noise reduction, the one-hour dataset was clipped into segments of 1 to 10 seconds per session to suit the model's requirements.



Figure 2 Waveform Before Audio Separation



Figure 3. Waveform After Audio Separation

Subsequently, the sample rate of the audio dataset was adjusted to 22050Hz, 16 bits, mono to match the pre-trained Tacotron2 model. This adjustment is crucial to maintain the signal consistency and prevent accuracy degradation during the model fine-tuning process [11].

Furthermore, audio augmentation was applied to provide dataset variety and improve model performance by compensating for limited datasets. This research utilized pitch shifting to modify sound pressure to be a higher or lower tone and intonation that allows the model to recognize diverse vocal patterns [12]. Additionally, time stretching was implemented to adjust the speed and duration of voice data, providing variations with a longer duration to enhance model robustness [13].

2.2 Metadata Audio Transcription

In this study, the audio transcription was utilized to retrieve information about the spoken words and pronunciation details from the voice dataset, with the results included in the metadata to serve as training labels for the Tacotron2 model [14] as shown in Table 1.

The transcription results are converted into metadata to establish a reference for determining attributes and labels during the fine-tuning process. In the Tacotron2 train/fine-tuning phase, the rules in creating metadata have 2 parameters, which are the audio file location and the corresponding transcription result. This process is crucial for ensuring the model precisely recognizes the text transcription results with the appropriate audio location and correct transcription.

Table 1. Metadata Example for Tacotron2

Metadata Example
/content/TTS-TT2/wavs/audio1-0120.wav They are so amazing as well and they're so entertaining.
/content/TTS-TT2/wavs/audio1-0122.wav I admire everyone in myth council and of course.
/content/TTS-TT2/wavs/audio1-0139.wav hahaha ahah.
/content/TTS-TT2/wavs/audio1-0048.wav Are you scared? Hahaha. Very scared good okay, you should be scared. Cause umm, as a cool and exceptional agent, I literally did my research on you hahahaha.
/content/TTS-TT2/wavs/audio2-17.wav drink a lot drink a lot of warm water
/content/TTS-TT2/wavs/audio1-0059.wav Wow.

2.3 Generating Mel Spectrogram

Feature extraction was performed on the Indonesian voice dataset to convert acoustic features into Mel Spectrogram as the primary step for speech synthesis. These Spectrograms capture essential voice characteristics, rhythm and intonation from the conversational dataset sample [15].

The implementation of this method is illustrated through the sample audio dataset reading "According to my data on this tablet here." as shown in Figures 4 and 5. By extracting the characteristics and sound information of the audio in the dataset, the system stored information such as rhythm, and intonation, which can be used for training the Tacotron2 model.

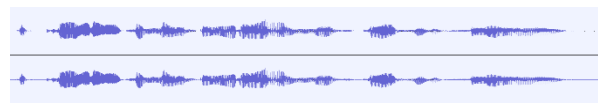


Figure 4. example of a audio wave inside audio1-0052.wav

Mel Spectrogram of audio1-0052.wav

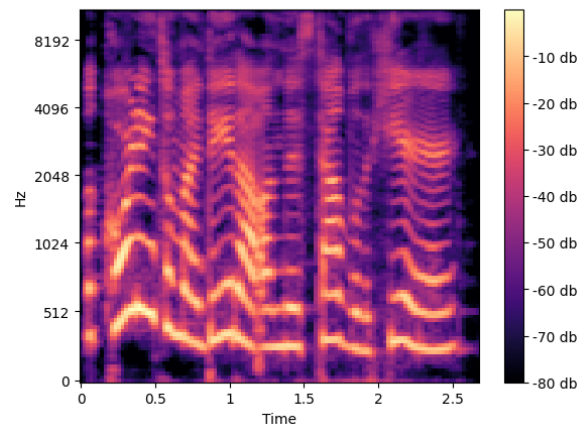


Figure 5. Mel spectrogram from audio1-0052.wav

2.4 Generative AI Model

In this research, the Generative AI model to generate responses for mental health support is the Gemini 1.5 flash 002 model, which is built to manage significantly long context reasoning and natural language processing [16], [17]. This model version offers flexibility in configuring safety thresholds and response parameters, such as Top-P and Top-K. This configuration enables the model to be more expressive but still prevents the

over-filtering of non-hazardous inputs that often occurred in more restricted versions.

The generative process requires user transcription input for contextual reasoning within the Transformer-based architecture. The model then generates an interactive response that is forwarded to the Tacotron2 module for speech synthesis. By utilizing this lightweight with customizable settings, the system maintains low latency and a supportive dialogue flow, which is key for natural Human-AI interaction in mental health support scenarios.

2.5 Speech Synthesis Model

A speech synthesis model was built by integrating an acoustic sequence-to-sequence model with a neural vocoder to produce natural artificial voices [18], [19]. Tacotron2 was implemented as the acoustic model to convert text inputs into a Mel Spectrogram [20]. This model utilizes a recurrent sequence-to-sequence architecture that maps the character embeddings to mel spectrograms, which effectively capture the contextual frequency information in the audio signal [21].

The process involves 3 convolutional layers for feature extraction, followed by Bidirectional LSTM and Location Sensitive Attention layers to manage the context memory. Thereafter, the Pre-Net and Convolution Post-Net Layers will filter and clean the resulting features to clean the Mel spectrogram, reducing noise and enhancing overall data quality.

For the waveform synthesis, WaveGlow was applied to predict and transform Mel Spectrogram features into high-quality audio output [22], [23]. Unlike the autoregressive WaveNet model, which predicts audio data sequentially and often results in suboptimal performance, WaveGlow utilizes a “normalizing-flow” which can predict all audio samples in parallel [24].

2.6 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is implemented to facilitate natural interaction by converting voice user queries into the text transcription for processing by Generative AI model [25]. This process highlights the essential role of ASR in facilitating interactions through AI platforms as a digital assistant [26].

In this research, Whisper Large-v3-Turbo is utilized as the primary ASR model due to stable transcription and translation tasks [27]. This model was selected for its capability of multilingual transcription, particularly in Indonesian speech, by leveraging its training on diverse non-English datasets [28]. The implementation of this state-of-the-art system provides low Word Error Rate (WER), which is vital for maintaining the natural flow of dialogue and understanding in mental health support context.

2.7 User Evaluation Study

Evaluating the perceptual quality of speech synthesis is a vital component of TTS research, for which Mean Opinion Score (MOS) serves as the recognized standard

assessment methodology [29], [30]. This subjective methodology is prevalent in the discipline, with formative studies frequently employing specialized participant cohorts to evaluate performance according to user preferences in certain circumstances [31].

In accordance with this approach, our research aimed to assess a virtual assistant within innovative systems of mental health support. We recruited 27 participants, the sample size ($n=27$) aligns with and surpasses standard methods for user evaluations in the TTS domain. For example, a study in this domain [31] effectively reached conclusions based on 21 participants. Our participants, aged 17-25 with backgrounds as Indonesian students and workers, were chosen due to their probable experiences with academic or work-related anxiety, making their feedback on the system's pertinence more significant.

Participants interacted with the prototype application of the virtual assistant using headphones to ensure the audio quality and then completed eight questionnaires across three categories. The first category, TTS Naturalness, assesses the perceived quality of the synthesized voice, including its naturalness, clarity, and intonation. The second category, AI Responses, evaluates the quality of the AI's replies in terms of their relevance, coherence, and perceived empathy. The final category, Interaction Experience, Measures the overall user experience, including the ease of conversation flow and the system's perceived helpfulness for mental support. These instruments utilized a 5-point Likert scale (1=very poor to 5=very good). For the TTS Naturalness, we computed the standard Mean Opinion Score (MOS) as the average rating, as is common practice in speech synthesis evaluation. For the AI Responses and User Experience categories, we computed the Mean Score for each category to measure the user perception of conversational quality, as is common practice in user experience research in Human Computer Interaction and conversational AI [31], [32], [33].

2.8 Ethical and Safety Considerations

The application of AI Virtual Assistants in mental health shows significant opportunities and several ethical challenges. To address these complexities, this research methodology is developed in line by prevailing legal landscape in Indonesia. Specifically, the utilization of AI Virtual Assistants is conducted under the ethical framework provided by Indonesian Law No. 17 of 2023 concerning Health, which defines Health Technology as any form of tools, products, and methods aimed at assisting in the establishment of diagnosis, prevention, and handling of human health problems. This definition gives the legal basis for integrating AI into mental healthcare. Concurrently, all the aspects of data handling, from data collection to analysis, strictly follow the guidelines of Indonesian Law No. 27 of 2022 concerning Personal Data Protection, safeguarding the privacy and autonomy of individuals involved in the study.

Given the intrinsic fragility of individuals undergoing mental health crises and the sensitive nature of their information, stringent ethical and safety measures have been diligently established throughout this project. Before participation, all participants provide informed permissions, comprehending the study's objectives, potential risks and advantages, data utilization, and their ability to withdraw at any time, in compliance with Indonesian health and data protection laws.

To safeguard participant privacy, all collected data is promptly anonymized or pseudonymized and stored on secure with limited access, in compliance with Law No. 27 of 2022. A critical component of our safety protocol involves the handling of potentially inappropriate or dangerous queries, the AI system is designed with a detection mechanism to identify language indicating self-harm, sexually explicit, harassment and hate speech to others. It is highlighted that while the AI is designed to support individuals experiencing low to moderate stress, cases indicating severe mental health distress or crisis are immediately advised to seek professional or psychological assistance. It's vital to emphasize that AI has inherent limitations in crisis intervention, and human support remains critical in such situations. Furthermore, this research is committed to integrating collaboration with experienced mental health professionals, to ensuring appropriateness and accuracy of AI-generated responses and contributing to the design of robust crisis protocols.

3. Results and Discussions

This section shows the results and discussion of the proposed system. It begins with the fine-tuning configuration of the Tacotron2 model, followed by evaluation of the trained models and the outcomes of human evaluation analysis. This section also outlines system integration, highlights the potential impact and applications of the system, and reflects on limitations and opportunities for further development.

3.1 Tacotron2 Trained Models Configuration

In hyperparameter settings on the fine-tuned Tacotron2 model, researchers provide a configuration of hyperparameter settings including learning rate, batch size, and number of epochs, as well as setting the type of GPU used during training models.

In the Tacotron2 model configuration, the Data Type, GPU, and Hyperparameters affect the performance and loss rate results of the model. The result of the output will also affect how good the rhythm and intonation produced by the Tacotron2 model is when the inferences are executed.

3.2 Tacotron2 Model Performance Comparisons

The results of the Tacotron2 model training configuration in Table 2 show that both training loss and validation loss have different variations. Hyperparameters and GPU type also affect the training time and the extent to which the model can read the entire context of the audio dataset.

Table 2 GPU and Hyperparameter Configuration Trained Tacotron2 Models

No	Model Aliases	Augmentation Data	GPU Type	Num of Epochs	Batch Size	Learning Rate
1	v.1.0	Yes	Nvidia A100	60	32	0.0003
2	v1.1	No	Nvidia T4	80	32	0.0003
3	v1.2	No	Nvidia T4	100	32	0.0001
4	v1.3	No	Nvidia T4	100	32	0.0003
5	v1.4	No	Nvidia T4	110	32	0.0004
6	v1.4 rev1	No	Nvidia T4	250	32	0.0003
7	v1.4 rev2	No	Nvidia T4	250	32	0.0001
8	v1.5	No	Nvidia T4	150	32	0.0002
9	v1.6	No	Nvidia T4	100	16	0.0003
10	v1.7	No	Nvidia T4	120	16	0.0001
11	v1.8	No	Nvidia L4	100	64	0.0001
12	v1.9	No	Nvidia T4	250	64	0.0003

Table 3 Performance Comparisons of Tacotron2 Trained Models

No	Model Aliases	Average Estimated Time per Epochs	Estimated Total Training Time	Training Loss	Validation Loss
1	v.1.0	1,2 Minutes	1 hour 24 minutes	0.2621	0.2378
2	v1.1	40 Seconds	54 minutes	0.2743	0.2038
3	v1.2	40 Seconds	1 hour 6 minutes	0.3246	0.2786
4	v1.3	40 Seconds	1 hour 6 minutes	0.2286	0.2132
5	v1.4	40 Seconds	1 hour 12 minutes	0.1984	0.184
6	v1.4 rev1	40 Seconds	2 hours 48 minutes	0.1398	0.1379
7	v1.4 rev2	40 Seconds	2 hours 48 minutes	0.2063	0.2149
8	v1.5	40 Seconds	1 hour 42 minutes	0.2345	0.1978
9	v1.6	1 minutes	1 hour 42 minutes	0.1615	0.1902
10	v1.7	1 minutes	2 Hours	0.2667	0.2314
11	v1.8	15 Seconds	24 Minutes	0.3857	0.2796
12	v1.9	15 Seconds	1 Hours	0.2368	0.1811

The comparison in Table 3 shows that the Tacotron2 v1.4 rev1 model is the type of model with the best hyperparameters because the model has a relatively

small level of training loss and validation loss. Moreover, the difference in training loss and variable loss values shows a small difference, so that both train

and validation can be considered convergent, and the model has a good generalization as illustrated in Figures 6 and 7.



Figure 6 Training Loss Graph of the Best Model

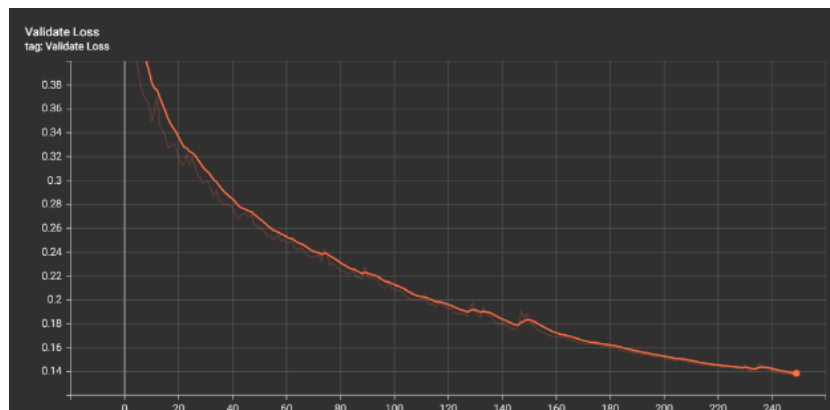


Figure 7 Validation Loss of the Best Model

In the training process, the training dataset tries to fit and synchronize with the validations data model. The model tries to generalize both data and tries to learn the data well. It can be seen that at the 230th epoch until the last, the model starts to learn the training data well and

applies the knowledge gained to the validation data well.

In Figure 8, it shows that the part of the text at the beginning to the middle of the sentence contributes a lot in providing a very clear output.

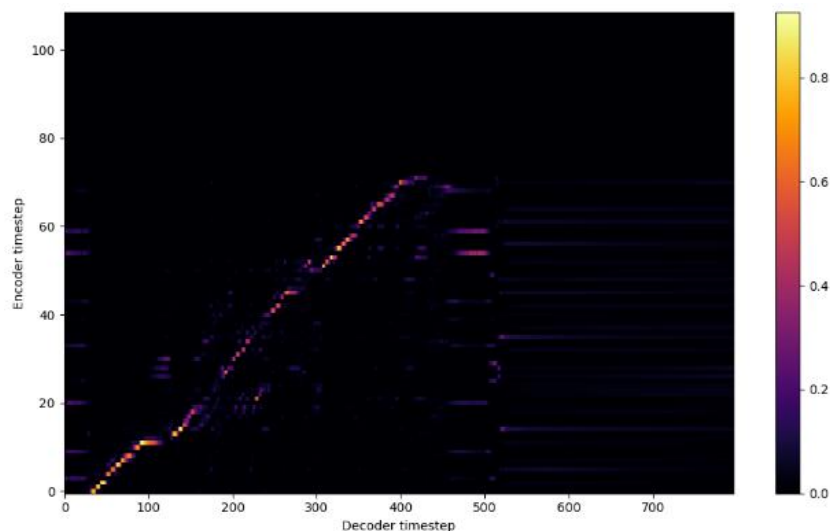


Figure 8 Attention Map Visualization After Train the Tacotron2 Model

The model gives a short voice output at the beginning of the sentence and gives a long voice output at the start of the mid-sentence, while in the last sentence, the model shows a small contribution in the last sentence.

In addition, the graph shows that the mid-to-late sentence transition is very influential in providing intonation in the voice output and the model shows a diagonal line, indicating that the model is starting to learn well.

However, the graph also shows that there is a lit point between the diagonal lines, indicating that there is still noise that makes the model a little difficult to train the given dataset, so there is a need for cleaning and adding more datasets on an ongoing basis.

3.3 Output Testing

In the output test using the Tacotron2 and waveglow models, the sound response of the model gives varied results based on the spectrogram results and the sound produced as shown in Table 4.

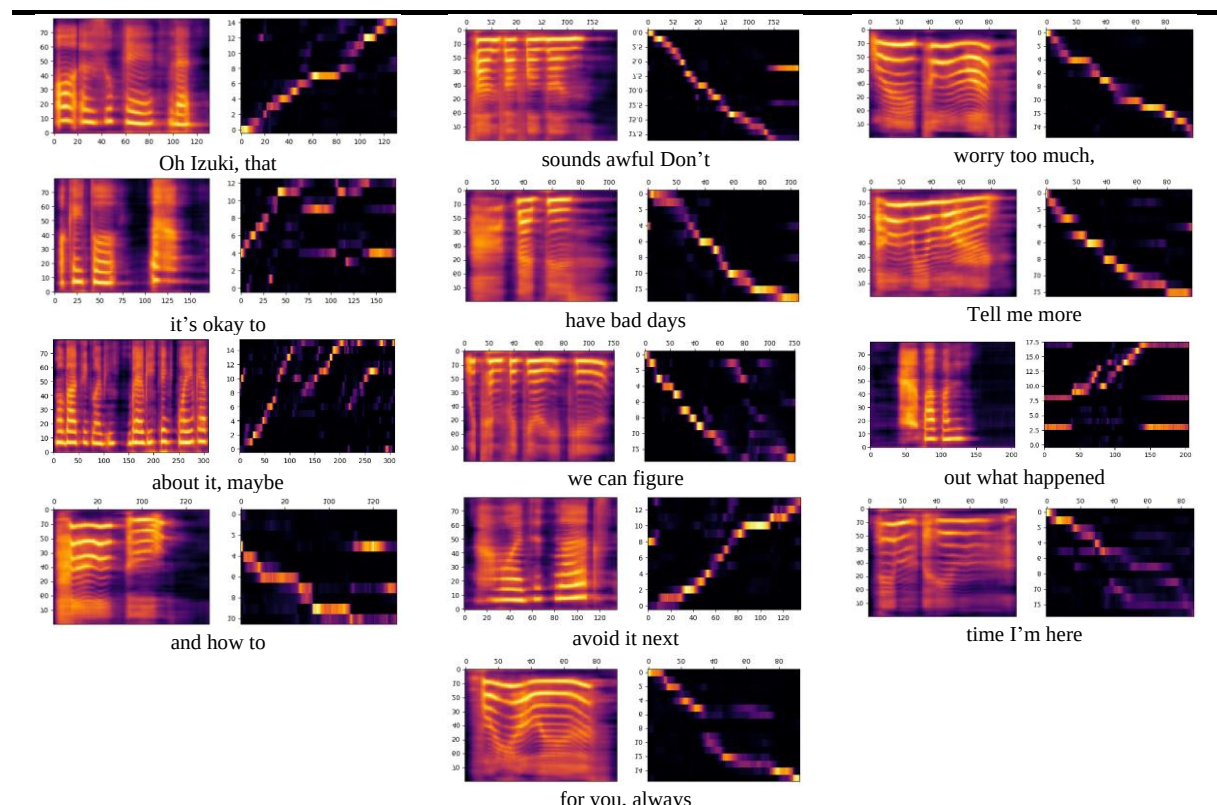
According to Table 5, the results of the generated mel spectrogram and attention map show that almost the majority of the generated voices can capture the context

in the response text generated by Generative AI Gemini properly, and it can also be seen that there has been variation in intonation where it can be seen that the model tries to respond with high and low intonation voices.

Table 4. Output result from voice input with the speech recognition library and Gemini Response

Voice Input	Speech Recognition Text Response	Gemini Response	Tacotron2 + WaveGlow Responses
Input's Link	Do you know what? I want to say something. So, my boss yelled at me because I didn't do my jobwell. I'm really sad right now.	Oh Izuki, that sounds awful Don't worry too much, it's okay to have bad days Tell me more about it, maybe we can figure out what happened and how to avoid it next time I'm here for you, always	Output's Link

Table 5. Mel Spectrogram of the outputs



However, the graph also shows that some of the context is not visible on the diagonal but rather there is a light color around the diagonal graph which indicates that the model is still struggling to handle the noise generated by the previously trained dataset. This limitation may

also be influenced by the characteristics of the dataset used for fine-tuning.

With the limited dataset, the output of the model still suffers from some issues, such as poor clarity, noise and distortion, which make the output less natural and

sometimes not perceptible to human ears. The limited dataset was a tradeoff due to available resources, including availability of training hardware and significant time and effort required for data collection and processing.

Despite this, the approach was considered sufficient for proof of concept and early-stage evaluation of the system. Utilizing the pre-trained model from LJSpeech dataset of 13,100 short audio clips helped create a strong initial foundation. This allowed models to adapt and produce speech with varying intonations, making TTS responses a bit more interactive and dynamic. Our initial qualitative evaluations indicate that the system produced functional speech, which gave us a fair result for this stage.

However, we acknowledge that these datasets are minimal for a model intended for robust, real-world applications. A larger dataset between 1,000 and 3,000 samples would probably be needed to ensure there is more stability and generalizability. Even with this limitation, the model has shown proficiency, and we plan to develop more datasets in the future.

3.4 Human Evaluation Analysis

Before presenting evaluation results from participants, we found a significant contrast revealed in the participant's background. As shown in Figure 9, most participants had prior experience using virtual assistants. However, Figure 10 illustrates that they infrequently utilize voice-based interactions. This gap is noteworthy because our speech-based system for mental health support was evaluated by users who, despite being familiar or experienced in utilizing the technologies, had no existing habit or preferences for voice-based interactions.

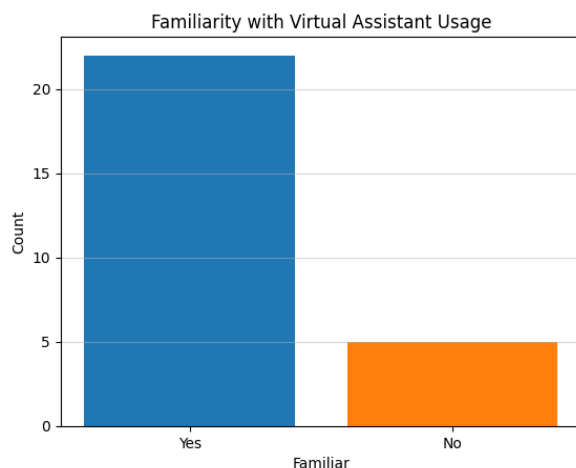


Figure 9 User familiarity with virtual assistant usage

Considering the participants' profile, the Mean Opinion Score (MOS) and Mean Score (MS) Analysis revealed several key insights. The analysis showed that the Interaction Experience category obtained the highest rating (MS = 3.91, SD = 0.02), followed by AI Responses (MS = 3.83, SD = 0.26). Based on our interpretation range in Table 6, both scores fall into

“Good” category. In contrast, the TTS Naturalness category was rated lower (MOS = 3.27, SD = 0.05), corresponding to a “Fair” rating.

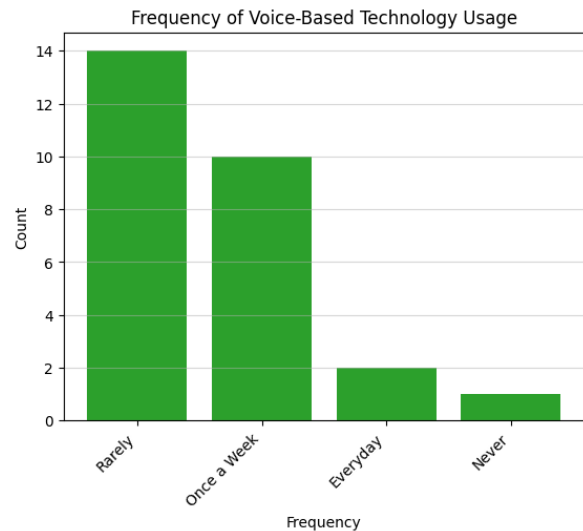


Figure 10 User frequency of using voice-based virtual assistant

A summary of these results is presented in Table 7. Notably, the standard deviations were consistently low across all categories. This suggests a high degree of agreement among participants regarding the strongest and weakest aspects of the system and the fact that the SD values were calculated from average category scores rather than individual responses.

Table 6 Qualitative Interpretation of 5 Point Likert Scale

No	Score Range	Qualitative Interpretation
1	1.00 – 1.49	Very Poor
2	1.50 – 2.49	Poor
3	2.50 – 3.49	Fair
4	3.50 – 4.49	Good
5	4.50 – 5.00	Excellent / Very Good

Table 7 Human Evaluation Analysis Results

No	Category	Metrics	Score (Standard Deviation)
1	TTS Naturalness	Mean Opinion Score (MOS)	3.27 (SD=0.05)
2	AI Responses	Mean Score (Likert)	3.83 (SD=0.26)
3	Interaction Experience	Mean Score (Likert)	3.91 (SD=0.02)

The most prominent findings were in the high score for Interaction Experience Category. This result, which corresponds to a “Good” rating, suggests that participants found the system to be intuitive and easy to follow. Moreover, the low standard deviation indicates a very high degree of consensus, indicating that participants consistently shared this positive perception.

One of the main reasons behind this highest score was the positive perception of the virtual assistant's potential. As illustrated in Figure 11, the majority of participants agreed that this system had the potential to provide mental health support. This is a crucial finding, as it validates the core premise of this research that

participants were open to the idea of utilizing speech-based virtual assistants for mental health support.

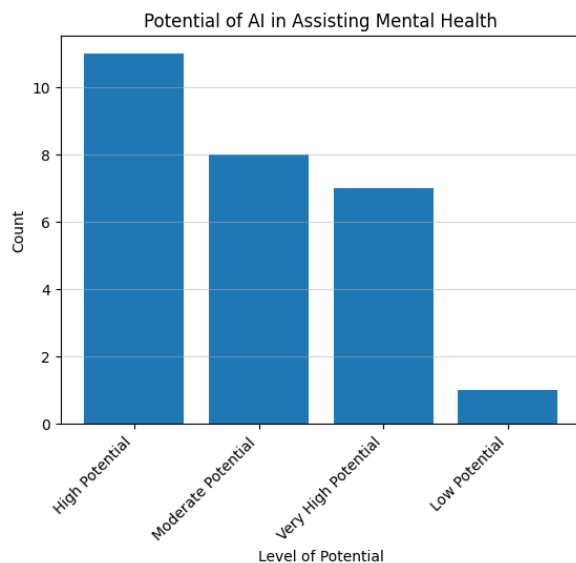


Figure 11 Participants' Perceptions of the System's Potential for Mental Health Support

The AI Responses Category was also rated favorably, achieving a "Good" score. This indicated that the content generated by Gemini was successful, with most participants finding the replies were relevant to their queries and emotionally supportive. The quality of the responses also contributed to a positive overall flow of conversation.

On the other hand, the TTS Naturalness category received the lowest score, corresponding to a "Fair" rating. This result indicates that while the voice was understandable, the natural aspects still aren't at the same level as the others. These findings were consistent with our prior technical analysis regarding the dataset size. Although the fine-tuning dataset was concise, utilizing pre-trained model on the LJSpeech dataset allowed the system to achieve baseline quality. The "Fair" score from human evaluations confirms that this approach was sufficiently provided the functional systems. However, a large dataset would be needed to reach a higher level of naturalness.

In conclusion, although the 'Fair' mark for TTS Naturalness identifies an obvious area for future development, overall system performance was perceived positively because of its great strengths in AI Responses and Interaction Experience. This indicates that the underlying conversational elements are essentially sound. More significantly, this positive perception is strongly reinforced by the result that most participants consider the system to have great potential as a mental health support tool. Overall, these findings indicate that the system proposed here is a valid proof-of-concept that shows much promise as a future platform in delivering accessible assistance for anxiety and other mental health issues. However, the limited number of participants would restrict the generalizability of the result. Future work will include

larger and varied participants to strengthen the robustness of these findings.

3.5 System Integrations

To provide clear understanding of our proposed system, we divide the workflow into four sequential phases. Each phase represents logical stage in the system's overall architecture, ensuring seamless mechanism between the ASR, Generative AI, and TTS Modules.

The system initialization stage begins with the server loads all required models. The models to be loaded include Whisper for Speech-to-Text Transcription, Gemini for Generative AI Model, and Tacotron2 and WaveGlow for Text-to-Speech Model. When the models are loaded, the server will start listening for incoming requests.

The request handling process is initiated when the server captures the conversation ID and voice input from user as blob data, which has been encoded by base64 encoding. After the encoded audio input is decoded, the input audio is stored into temporary storage as .wav file then transcribe to text base using Whisper. The text-based input will be sent to Gemini model to get the generated responses. The generated text output is filtered by regular expressions to remove unnecessary characters, this action required to prevent distorted audio synthesis output from the Tacotron2 and WaveGlow Model. The cleaned text output is synthesized into audio using TTS System then it will be converted to wav extension because this TTS output is still in array form and needed to be converted before being returned to the client in response format.

The text generation process by Gemini begins with the retrieval of conversation history from a memory checkpoint based on conversation ID. The input text from the user is appended to the conversations structure, then Gemini produces an appropriate response. This response will be added to the memory checkpoint again and stored for the next conversation.

In the speech synthesis stage, the cleaned generative text is preprocessed and converted to sequence form. Tacotron2 model generates mel-spectrogram from this sequence, then passed through the WaveGlow vocoder to synthesize the raw audio waveform. Postprocessing is applied to the waveform to generate the final audio output, which is returned to request handler.

3.6 Impacts and Real-World Applications

The proposed system demonstrates potential in real-time automatic speech recognition and interactive response, highlighting its relevance across various digital platforms. In terms of technology, this system introduces a speech recognition approach to virtual assistants by integrating Speech Recognition, Generative AI and TTS as core components, rather than solely dependent on text-based interaction. This system also enhances scalability and offers potential for integration into current infrastructure, elevating the

proposed system as a highlighted contribution to AI-driven virtual assistant research.

For mental health domain, the system offers an accessible tool for individuals dealing with stress and anxiety. The virtual assistant provides availability and a non-judgmental environment and allows users to express themselves without fear of stigma. This form of support has the potential to reduce the severity of anxiety symptoms in daily life and enhance the emotional wellbeing of the users who might otherwise experience a lack of consistent relationship.

Despite the potential, the method also presents several limitations. The allocation of computational resources that involved are relatively high, which may limit implementation in low resources environments. Moreover, more development in speech synthesis technology is necessary to deliver more natural and engaging interactions. It is also important to consider that the system is primarily intended to support individuals with low and moderate anxiety and should not be considered a replacement for professional care in severe cases. Finally, interdisciplinary collaboration with mental health professionals remains necessary in assuring ethical use and to strengthen the system's effectiveness.

4. Conclusions

This study proposed a voice-based virtual assistant integrating Automatic Speech Recognition (ASR), Generative AI model, and Text-to-Speech (TTS) module to explore accessible and stigma-free mental health support through natural speech conversations for individuals experiencing stress and anxiety. The implementation demonstrates that a speech-based interaction framework can enhance naturalness and reduce stigma compared to text-only approaches by allowing spontaneous and empathetic conversational exchanges. Technically, the fine-tuned Tacotron2 and WaveGlow pipeline achieved convergence with low loss in the training and validation metrics, the TTS model produced generally understandable output, and the human evaluations showed the system's usability, with Interaction Experience and AI Responses rated as "Good", showing strong results. However, TTS naturalness received a "Fair", indicating good progress but remaining limitations in prosody and naturalness. These results indicate that combining Automatic Speech Recognition (ASR), a generative conversational model, and Text-to-Speech (TTS) can yield a functional prototype for conversational mental health support in Indonesian and contribute to the field by offering a practical end-to-end implementation and evaluation for a specific underrepresented language. Practically, this system is designed to be an accessible and supportive tool for individuals with low to moderate anxiety, offering immediate empathetic responses, and lowering interaction barriers. However, this system is not a replacement for mental health professionals in severe cases. Limitations include a small and restricted

participant pool and a limited fine-tuning dataset for TTS, followed by computational resource requirements that may prevent its use in low-resource environments. Future work should prioritize larger, more diverse datasets for TTS fine-tuning, improvements to synthesis naturalness and latency, scalability testing across domains, and collaboration with mental health professionals to improve safety, ethical safeguards, and clinical appropriateness.

References

- [1] J. Liu, W. Ning, N. Zhang, B. Zhu, and Y. Mao, "Estimation of the Global Disease Burden of Depression and Anxiety between 1990 and 2044: An Analysis of the Global Burden of Disease Study 2019," *Healthcare*, vol. 12, p. 1721, Dec. 2024, doi: 10.3390/healthcare12171721.
- [2] Tim Penyusun SKI 2023 Dalam Angka, "SURVEI KESEHATAN INDONESIA 2023 DALAM ANGKA," 2024. Accessed: Dec. 16, 2024. [Online]. Available: <https://www.badankebijakan.kemkes.go.id/ski-2023-dalam-angka/>
- [3] S. Umadiyan and A. D. Kalifia, "Perbedaan Respon Emosional Antara Remaja Perempuan Dan Laki-Laki Terhadap Faktor-Faktor Penyebab Depresi," *Gudang Jurnal Multidisiplin Ilmu*, vol. 2, no. 1, pp. 293–296, 2024.
- [4] K. A. Ainunnida and N. Hartini, "Hubungan antara kesepian dengan ide bunuh diri yang dimoderasi oleh depresi pada remaja dengan orang tua bercerai," *Berajah Journal*, vol. 2, no. 3, pp. 647–660, 2022.
- [5] S. Zhang, Z. Meng, B. Chen, X. Yang, and X. Zhao, "Motivation, social emotion, and the acceptance of artificial intelligence virtual assistants—Trust-based mediating effects," *Front Psychol*, vol. 12, p. 728495, 2021.
- [6] J. de Freitas, A. K. Uguralp, Z. Uguralp, and S. Puntoni, "AI Companions Reduce Loneliness," no. 24–078, Jun. 2024, doi: 10.2139/ssrn.4893097.
- [7] M. D. Rosanti, S. H. Wijoyo, and A. Rachmadi, "Analisis Pengaruh Automated Customer Service Berbasis Artificial Intelligence Pada Aplikasi E-commerce (Studi Kasus Aplikasi Lazada)," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 9, no. 5, 2025.
- [8] T. M. Rahardyan, C. H. Susilo, A. M. N. Iswara, and M. L. Hartono, "ChatGPT: The Future Research Assistant or an Academic Fraud?[A Case Study on a State University Located in Jakarta, Indonesia]," *Asia Pacific Fraud Journal*, vol. 9, no. 2, pp. 275–293, 2024.
- [9] T. Koulouri, R. D. Macredie, and D. Olakitan, "Chatbots to Support Young Adults' Mental Health: An Exploratory Study of Acceptability," *ACM Trans. Interact. Intell. Syst.*, vol. 12, no. 2, Jul. 2022, doi: 10.1145/3485874.
- [10] R. Hennequin, A. Khelif, F. Voituret, and M. Moussallam, "Spleeter: a fast and efficient music source separation tool with pre-trained models," *J Open Source Softw*, vol. 5, no. 50, p. 2154, 2020, doi: 10.21105/joss.02154.
- [11] R. Yoneyama, R. Yamamoto, and K. Tachibana, "Nonparallel High-Quality Audio Super Resolution with Domain Adaptation and Resampling CycleGANs," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10097002.
- [12] V. C. Govoreanu and M. Neghină, "Speech emotion recognition method using time-stretching in the preprocessing phase and artificial neural network classifiers," in *2020 IEEE 16th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2020, pp. 69–74.
- [13] Q. M. M. Zarandah, S. M. Daud, and S. S. Abu-Naser, "Spectrogram flipping: a new technique for audio augmentation," *J Theor Appl Inf Technol*, vol. 101, no. 11, 2023.
- [14] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational End-to-End TTS for Voice Agents," in

- 2021 *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 403–409. doi: 10.1109/SLT48900.2021.9383460.
- [15] A. Karim, “Text to speech using Mel-Spectrogram with deep learning algorithms,” *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 10, pp. 380–386, Dec. 2022, doi: 10.21533/pen.v10i3.3113.
- [16] K. Alexander, “THE ANALYSIS OF THE EFFICIENCY OF GENERATIVE AI ALGORITHMS FOR CREATING A NATURAL DIALOGUE,” *The American Journal of Interdisciplinary Innovations and Research*, vol. 6, no. 6, pp. 26–34, Jun. 2024, doi: 10.37547/tajirr/Volume06Issue06-08.
- [17] G. Team et al., “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context,” 2024. [Online]. Available: <https://arxiv.org/abs/2403.05530>
- [18] R. Liu, B. Sisman, G. Gao, and H. Li, “Expressive TTS Training with Frame and Style Reconstruction Loss,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 29, pp. 1806–1818, 2021, doi: 10.1109/TASLP.2021.3076369.
- [19] M. J. Hwang, R. Yamamoto, E. Song, and J. M. Kim, “Tts-by-tts: Tts-driven data augmentation for fast and high-quality speech synthesis,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6598–6602. doi: 10.1109/ICASSP39728.2021.9414408.
- [20] M. R. Hasanabadi, “An overview of text-to-speech systems and media applications,” *arXiv preprint arXiv:2310.14301*, 2023.
- [21] A. F. A. Aziz, S. Tiun, and N. Ruslan, “End to end text to speech synthesis for malay language using tacotron and tacotron 2,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 6, 2023.
- [22] R. J. Weiss, R. J. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, “Wave-Tacotron: Spectrogram-Free End-To-End Text-To-Speech Synthesis,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 5679–5683. doi: 10.1109/ICASSP39728.2021.9413851.
- [23] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, “Universal neural vocoding with parallel wavenet,” in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 6044–6048. doi: 10.1109/ICASSP39728.2021.9414444.
- [24] G. Sanjay, K. C. Sooraj, and D. Mishra, “Natural Text-to-Speech Synthesis by Conditioning Spectrogram Predictions from Transformer Network on WaveGlow Vocoder,” in *2020 7th International Conference on Soft Computing and Machine Intelligence, ISCMi 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 255–259. doi: 10.1109/ISCMi51676.2020.9311564.
- [25] C. Xu et al., “Recent Advances in Direct Speech-to-text Translation,” 2023. [Online]. Available: <https://arxiv.org/abs/2306.11646>
- [26] H. Kheddar, M. Hemis, and Y. Himeur, “Automatic speech recognition using advanced deep learning approaches: A survey,” *Information Fusion*, vol. 109, p. 102422, 2024, doi: <https://doi.org/10.1016/j.inffus.2024.102422>.
- [27] D. Machacek, R. Dabre, and O. Bojar, “Turning Whisper into Real-Time Transcription System,” 2023. [Online]. Available: <https://arxiv.org/abs/2307.14743>
- [28] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in *Proceedings of Machine Learning Research*, vol. 202. PMLR, Apr. 2023, pp. 28492–28518. [Online]. Available: <https://proceedings.mlr.press/v202/radford23a.html>
- [29] 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.
- [30] C. Chen, Y. Hu, W. Wu, H. Wang, E. S. Chng, and C. Zhang, “Enhancing Zero-shot Text-to-Speech Synthesis with Human Feedback,” 2024. [Online]. Available: <https://arxiv.org/abs/2406.00654>
- [31] S. Wang, E. Székely, and J. Gustafson, “Contextual Interactive Evaluation of TTS Models in Dialogue Systems,” in *25th Interspeech Conference 2024, Kos Island, Greece, Sep 1 2024-Sep 5 2024*, 2024, pp. 2965–2969.
- [32] N. M. Johari and P. N. Nohuddin, “Developing a Good Quality Chatbot,” *J. Inf. Syst. Technol. Manag*, vol. 6, pp. 88–102, 2021.
- [33] I. K. F. Haugeland, A. Følstad, C. Taylor, and C. A. Bjørkli, “Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design,” *Int J Hum Comput Stud*, vol. 161, p. 102788, 2022, doi: <https://doi.org/10.1016/j.ijhcs.2022.102788>.