Bahasa Harmony: A Comprehensive Dataset for Bahasa Text-to-Speech Synthesis with Discrete Codec Modeling of EnGen-TTS.

Onkar Kishor Susladkar[1], **Vishesh Tripathi**[2], and **Biddwan Ahmed**[3]

[1,2,3]**Yellow.ai**
[1]onkarsus13gmail.com
[1,2,3]{onkar.susladkar,vishesh.tripathi,biddwan.ahmed}@yellow.ai

Abstract

This research introduces a comprehensive Bahasa text-to-speech (TTS) dataset and a novel TTS model, EnGen-TTS, designed to enhance the quality and versatility of synthetic speech in the Bahasa language. The dataset, spanning ~55.0 hours and 52K audio recordings, integrates diverse textual sources, ensuring linguistic richness. A meticulous recording setup captures the nuances of Bahasa phonetics, employing professional equipment to ensure high-fidelity audio samples. Statistical analysis reveals the dataset's scale and diversity, laying the foundation for model training and evaluation. The proposed EnGen-TTS model performs better than established baselines, achieving a Mean Opinion Score (MOS) of 4.45 ± 0.13. Additionally, our investigation on real-time factor and model size highlights EnGen-TTS as a compelling choice, with efficient performance. This research marks a significant advancement in Bahasa TTS technology, with implications for diverse language applications. Link to Generated Samples: https://bahasa-harmony-comp.vercel.app/

.

1 Introduction

The Bahasa language, spoken by a vibrant and diverse community, serves as a linguistic tapestry that encapsulates the rich cultural heritage of its speakers. In our increasingly digital world, the demand for advanced speech synthesis technologies tailored specifically to Bahasa becomes more pronounced. This necessity arises from the need for synthetic speech that authentically captures the nuances of Bahasa expressions, accommodating the linguistic diversity within the Bahasa-speaking population, including various dialects, registers, and cultural nuances.

Existing text-to-speech (TTS) systems, while making strides in the broader landscape, often fall short in addressing the requirements of Bahasa. This gap underscores the motivation for our research, which introduces a meticulously curated Bahasa TTS dataset and an innovative TTS model, EnGen-TTS. While other models have attempted to address the synthesis challenges for diverse languages, including Bahasa, they may exhibit drawbacks such as limited adaptability, linguistic richness, or efficiency.

Our proposed EnGen-TTS model not only fills these gaps but also showcases superior performance when compared to established baselines. The model achieves a remarkable Mean Opinion Score (MOS) of 4.45 ± 0.13, outperforming existing models even without fine-tuning on additional Bahasa data. Our key strength is, positioning EnGen-TTS as a solution for high-quality, adaptive Text-to-Speech synthesis across various languages.

1.1 Contributions

1. Comprehensive Bahasa Dataset: Our research introduces a meticulously curated Bahasa text-to-speech (TTS) dataset, comprising ~55.0 hours sourced from diverse linguistic contexts. This dataset, enriched with contributions from skilled voice artists and varied textual sources, stands as a valuable resource for the research community and addresses the need for a comprehensive linguistic foundation for Bahasa TTS systems. We will make dataset, trained-model and finetuneing code publicly available. Dataset Link: https://bit.ly/3Vi22x9

2. Efficient Model Architecture: The proposed model architecture leverages a multi-lingual T5 (m-t5) encoder (Xue et al., 2021) for confitioning text latents for decoding audio sequence through neural codec language modeling. This innovative design optimizes the synthesis process, allowing for finetuning more efficiently and reduced computational time

Findings of the Association for Computational Linguistics: EACL 2024, pages 2731–2741

while maintaining high-quality Bahasa speech synthesis.

3. Integration of Neural Codec Language Modeling: The incorporation of a trainable neural codec language modeling module represents a novel contribution. This module captures both textual and audio features, enhancing the model's ability to understand Bahasa linguistic nuances effectively. The integration of trainable weights in this module contributes to the adaptability and expressive power of the TTS system.

4. Versatile Pre-trained Model: Our research presents a pre-trained TTS model, EnGen-TTS, showcasing exceptional performance without additional fine-tuning on Bahasa-specific data. Trained on LJ-Speech (Ito and Johnson, 2017) and VCTK, this model exhibits inherent strengths in adapting to new languages, highlighting its versatility and potential for the development of high-quality Text-to-Speech systems for diverse languages like Bahasa.

2 Related Work

In the domain of multilingual text-to-speech (TTS) datasets and models, several noteworthy contributions have been done for enhanced synthesis capabilities across diverse languages. The IndicSpeech: Text-to-Speech Corpus for Indian Languages (Srivastava et al., 2020a) project recognizes the critical need for TTS systems tailored to the linguistic diversity of India. Presenting a 24-hour corpus for Hindi, Malayalam, and Bengali, the authors not only contribute data but also train state-of-the-art TTS systems for each language.

In a similar vein, the paper titled Towards Building Text-to-Speech Systems (Kumar et al., 2023) for the Next Billion Users explores the landscape of deep learning-based TTS systems, specifically focusing on the challenges and opportunities within the context of Indian languages. Some models like SLBERT (Susladkar et al., 2023) a speech and language processing framework, which uses multimodal attention mechanism to get the better transition between the speech and language features. Recognizing the computational expense associated with investigating the multitude of Indian languages, lower resource availability, and untested advances in neural TTS, the authors evaluate various aspects such as acoustic models, vocoders, loss functions, training schedules, and speaker/language diversity. The results indicate that monolingual models with FastPitch (Łańcucki, 2021) and HiFi-GAN V1 (Kong et al., 2020a), trained jointly on male and female speakers, exhibit significant improvements across 13 languages, as measured by mean opinion scores (Viswanathan and Viswanathan, 2005).

Entity	Stats
Hours	~55 Hrs
Mean Audio Length	4.06 Sec
Total Words	458K
Vocab Size	23K
Sentences	68.9K
Mean Word Freq.	9.4
Total Recordings	52K

Table 1: Descriptive statistics of our Bahasa corpus. We see that the corpus consists of a diverse vocabulary and is at a scale well-suited for state-of-the-art neural TTS models.

Considering the landscape of existing TTS models, it's crucial to acknowledge advancements beyond the scope of the aforementioned papers. State-of-the-art models such as Tacotron (Wang et al., 2017), WaveNet (van den Oord et al., 2016), and more recently, FastPitch and HiFi-GAN, have demonstrated significant progress in the realm of TTS. These models leverage deep learning architectures to generate natural and expressive speech, contributing to the evolving landscape of TTS technologies.

3 Dataset

In creating our Bahasa text-to-speech (TTS) dataset, we curated a linguistically diverse textual foundation. This dataset is integral to our innovative TTS model, EnGen-TTS. Drawing from sources like Wikipedia and incorporating content from chat-GPT translation, our approach involved a strategic gathering of text samples. This fusion of varied linguistic contexts lays the groundwork for a robust Bahasa TTS dataset, capturing the language's nuanced breadth of expression.

3.1 Text collection

The textual foundation for our Bahasa text-to-speech (TTS) dataset was meticulously curated from diverse sources, enriching the dataset with varied linguistic contexts. We gathered text samples from prominent repositories such as Wikipedia, ensuring a broad representation of topics and lan-

guage styles. Additionally, we incorporated content generated through chat-GPT translation, further diversifying the dataset with conversational and translated expressions. This eclectic mix of sources contributes to a comprehensive and linguistically diverse textual corpus, laying the groundwork for a robust Bahasa TTS dataset.

3.2 Speaker selection

To imbue the TTS dataset with authentic and expressive voices, we engaged two skilled voice artists—one male and one female. These artists were selected based on their proficiency in Bahasa and their ability to convey the nuances of the language with clarity and naturalness. Both of the speakers are from southern Indonesia. The inclusion of both male and female voices ensures a balanced representation, catering to the diverse preferences of users interacting with the TTS system. The careful selection of voice artists contributes to the overall quality and authenticity of the recorded audio samples.

3.3 Recording Setup

Ensuring optimal recording quality is paramount for the success of any TTS dataset. Our recording setup was designed to capture the richness of Bahasa phonetics and nuances. A controlled acoustic environment was maintained to minimize external interference, and high-quality recording equipment was employed to capture the nuances of the voice artists' performances accurately. The setup included professional microphones, soundproofing measures, and studio-grade audio interfaces, creating an environment conducive to the production of high-fidelity Bahasa TTS audio samples. All the data we have recorded is at a sample rate of 48 kHz.

3.4 Corpus Statistics

We have a report of few statistics of our Bahasa Corpus in Table 1. Upon collecting the text data and organizing it into coherent sentences, the resultant Bahasa TTS corpus exhibits notable statistics reflecting the dataset's scale and diversity. The corpus comprises a total of 55 hours of recorded voice across 52K recordings. This extensive dataset is a testament to the effort invested in capturing a comprehensive range of linguistic expressions, ensuring the TTS system's adaptability to various applications and user preferences. These corpus statistics lay the foundation for subsequent model training and evaluation, fostering advancements in Bahasa TTS technology.

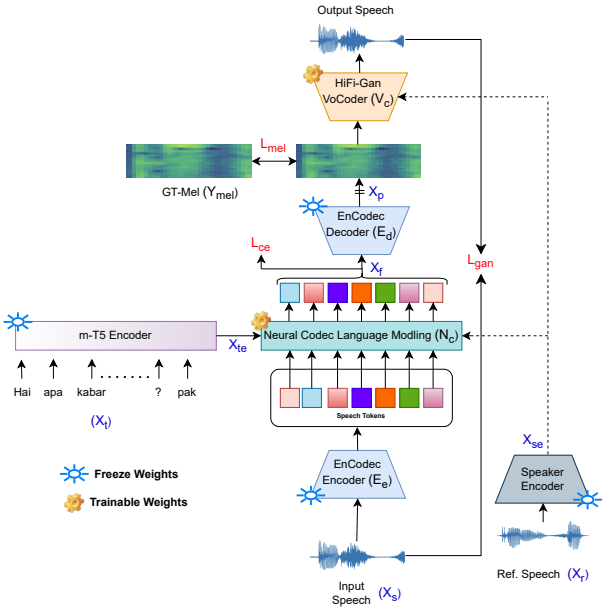

Figure 1: Architectural (EnGen-TTS) Framework for Bahasa Text-to-Speech Synthesis

3.5 Dataset Characterstics

The dataset encompasses 52K recordings, featuring a vocabulary size of 23,000 unique terms. Comprising a total of 55 hours, evenly distributed between male and female speakers, we strategically allocated 5 hours from each speaker for validation and an additional 5 hours from each speaker for testing purposes.So a total of 10 hours of testing & 10 hours of validation. This balanced selection ensures comprehensive coverage and representation in both the validation and test sets, fostering robust evaluation and training of our Bahasa TTS model, EnGen-TTS.

4 Architecture

Our research introduces the EnGen-TTS, a novel Bahasa text-to-speech synthesis system inspired by the state-of-the-art Encodec-based TTS Bark (Schumacher and LaBounty Jr, 2023).The system leverages the architectural framework illustrated in Figure 1:

1. m-T5 Encoder: A frozen multi-li T5 encoder is utilized for conditioning on text latents. This encoder is pre-trained and is kept frozen during training.

2. Audio Encodec: Audio Encodec from meta (Kumar et al., 2024) is pre-trained on an exten-

sive 60K-hour audio dataset and kept frozen during training. This discretizes the audio into tokens, providing a robust audio representation without further training.

3. Neural Codec Language Model: This module generates the audio sequence in an autoregressive manner. It conditions on both the text embeddings from the m-T5 encoder and the speaker embeddings, yielding a sequence that closely follows the linguistic and speaker-specific nuances of the input.

4. Speaker Encoder: A frozen encoder (Wan et al., 2018) trained on the LibriSpeech dataset. This module produces speaker latent vectors that condition the TTS output on the speaker's unique voice characteristics.

5. HiFi-Gan Vocoder: It is for converting the mel spectrogram into natural speech, this vocoder is fine-tuned to adapt to the specific frequency profiles of the Bahasa language. The system is designed to synthesize natural-sounding Bahasa speech by conditioning on both linguistic content and speaker identity.

Audio Codec Setting: We adopt the pre-trained EnCodec model(Kumar et al., 2024) as our tokenizer, a convolutional encoder-decoder model handling 22050 Hz audio at variable bitrates. The encoder generates embeddings at 75 Hz for 22050 kHz input, reducing the sampling rate by 320 times. These embeddings use residual vector quantization (RVQ) with 4 hierarchical quantizers of 1024 entries each, corresponding to a 3K bitrate for audio reconstruction at 22050 Hz.bFor our purposes, we use only the first entity of the 750 × 4 discrete representation matrix, as it contains all phonetic and content information, resulting in a 750 × 1 matrix. Higher bitrates, such as 12K, require more quantizers (e.g., 16) and offer better reconstruction quality. The EnCodec decoder then reconstructs the waveform at 22050 Hz from the discrete codes.

4.1 Method

Let, X_s be the input audio, X_t is the text corresponding to X_s, and, X_r be the reference audio of the same speaker. The methodology commences with byte pair encoding (BPE) to convert X_t into input IDs. These are fed into the frozen m-T5 encoder to derive text embeddings X_{te}. Concurrently, X_s is discritized by the frozen Audio-Encodec encoder E_e, producing discrete audio tokens in between (0 to 1023). The Speaker Encoder processes X_r to generate a speaker latent vector for each frame. These vectors (X_{se}) are then used to condition the model on the speaker's voice, providing a direct sequence that correlates with the length of the X_r sample. Then, Neural Codec Language Modeling N_c predicts the audio sequence conditioned on X_{te} and X_{se}. The Sequential cross-entropy loss (L_{ce}) between the generated audio sequence and the ground truth audio sequence is computed here to ensure the fidelity of the audio tokens X_f. The predicted audio sequence is then passed through the Encodec Decoder E_d to produce an intermediate audio representation. We calculate the loss (L_{mel}) between the predicted mel-spectrogram X_p and the GT-Mel (y_{mel}) to ensure the model accurately captures the handcrafted audio features.

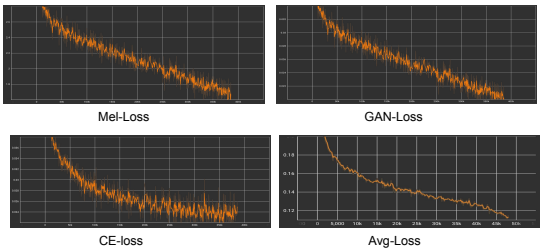

Figure 2: Loss plots

$$L_{ce} = -\log(N_c(E_e(X_s), X_{te}, X_{se})))$$
$$X_f = N_c(E_e(X_s), X_{te}, X_{se}))$$
$$L_{mel} = -\mid y_{mel} - E_d(X_f) \mid$$

At last, To achieve natural-sounding audio, we pass intermediate udio representation to the HiFi-Gan vocoder V_c which is conditioned on the speaker embeddings X_{se} to prevent mode collapse. The loss (L_{gan}) between the predicted speech and input speech is computed to align the output with the input audio distribution. To compute the Gan loss we use the temporal Discriminator module as a D. The loss follows:

$$L_{\text{gan}} = -\log(D(V_c(E_d(X_f), X_{se}))) + |X_s - V_c(E_d(X_f), X_{se})|$$

We adopt a composite loss function, taking a weighted average of L_{ce}, L_{mel}, and L_{gan} to update the model's weights. This combination of losses ensures that the model learns not only the accurate prediction of audio tokens but also the refined generation of melspectrograms and the final audio output. From Figure 2 we can see that all the 3 losses are helping model to learn optimally. The overall loss (L_t) is expressed as follows:

$$L_t = \alpha L_{ce} + \beta L_{mel} + \gamma L_{gan}$$

From our experiment, we determined the optimal values for the coefficients as follows: $\alpha = 1.2$, $\beta = 0.7$, $\gamma = 0.6$.

4.2 Audio Codec language modeling

In this section, we delve into the core component of our model, namely the Audio Codec Language Modeling Module. This pivotal module encompasses a Masked Self-Attention mechanism, Layer Normalization, two Cross-Attention blocks, and FeedForward layers, all integrated with GeLU non-linearity. As depicted in Figure 3, the process begins with discrete audio tokens being right-shifted and passed through the embedding layer to obtain audio embeddings. These embeddings are then directed through the Masked Self-Attention block. This specific block is utilized to decode the audio sequence autoregressively, akin to the GPT model, where the full scope of tokens isn't available during decoding. Utilizing normal self-attention could lead to overfitting on particular sequences and a lack of generalization across different audio sequences. The output from the masked self-attention undergoes layer normalization, augmented with a residual connection from the input audio embeddings. Subsequently, the embeddings are processed through a cross-attention module, conditioned on the text embeddings (X_{te}) derived from the m-T5 encoder. The output from this cross-attention phase is then subject to another layer normalization, followed by a residual connection from the preceding text-conditioned cross-attention block.

In contrast to previous methods (Zhang et al., 2023) that concatenate speaker embeddings directly with audio embeddings, our approach employs an additional cross-attention step for conditioning on speaker embeddings (X_{se}), enhancing the model's generalizability across multiple speakers in TTS applications. This output is finally channeled to a feedforward layer, followed by GeLU non-linearity and layer normalization. The architecture maintains a consistent embedding dimension of 1024 and 16 attention heads across all sub-modules. During inference, we also implement a KV-cache mechanism to enhance efficiency. Our larger model configuration comprises 26 such blocks, each meticulously designed to optimize performance and accuracy in audio processing tasks.

4.3 Inference

During inference, the Neural Codec Language Model (Nc) is initialized with the <SOS> (start of sequence) token as the input. The model then autoregressively generates the entire sequence, conditioned on the text embeddings produced by the pre-trained m-T5 model and the speaker embeddings derived from the specified reference audio. During training, audio codec inputs help the model learn the mapping between audio and text embeddings. However, during inference, the model relies on learned embeddings and the autoregressive mechanism to generate the sequence. Our current training does not use scheduled sampling or a weaning-off period typical of teacher forcing. The model smoothly transitions from training to inference due to robust text and speaker conditioning.

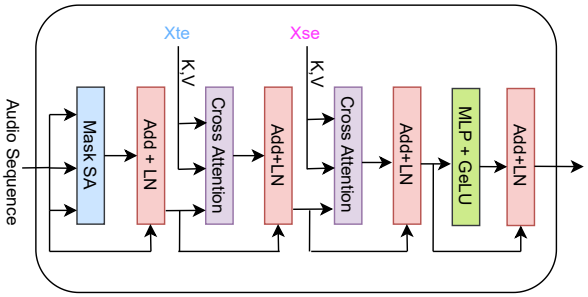

Figure 3: Audio Codec language modeling module

5 Experiments

5.1 Experimental Setup

We assessed our Text-to-Speech (TTS) models quality using Mean Opinion Score (MOS) and Comparative Mean Opinion Score (CMOS). MOS measures average listener preferences, indicating the models' naturalness, pleasantness, and intelligibility. CMOS compares models directly, identifying slight differences in perceived quality. Together, these metrics offer detailed insights into each model's performance and real-world applicability.

5.2 Real-Time Factor (RTF)

It is a measure of how quickly a TTS system can generate speech relative to the length of the input text. Specifically, the RTF is calculated as the ratio of the time taken to synthesize speech to the duration of the resulting audio. For example, an RTF of 1.0 means that the TTS system takes one second to generate one second of speech. An RTF of less than 1.0 indicates that the system is faster than real-time, whereas an RTF greater than 1.0 indicates that the system is slower than real-time. Lower RTF values are generally preferred as they indicate a more efficient and faster TTS system.

Baseline	MOS	CMOS	RTF
GradTTS (Popov et al., 2021)	4.01 ± 0.11	-0.0515	0.019
GlowTTS (Kim et al., 2020)	4.16 ± 0.14	-0.0422	0.031
VITs (Kim et al., 2021)	4.19 ± 0.12	-0.0366	0.023
NaturalSpeech (Tan et al., 2024)	4.39 ± 0.19	-0.0201	0.014
XTTS (Casanova et al., 2024)	4.39 ± 0.12	-0.0222	0.013
CLaM-TTS (Kim et al., 2024)	4.41 ± 0.09	-0.0189	0.027
VAENAR-TTS (Lu et al., 2021)	4.36 ± 0.21	-0.0326	0.014
EnGen-TTS-L (Without pretrained)	4.41 ± 0.10	-0.0161	0.016
EnGen-TTS-L	**4.45 ± 0.13**	**-0.0101**	**0.016**

Table 2: Comparison of all models on Bahasa Datasets

5.3 Quantitative Results

As the Bahasa language is written in Latin script, So there is always phonetic misalignment between speech and the input text. To learn those alignments we first trained our models and baselines on LJ-speech(Ito and Johnson, 2017) and VCTK(Yam, 2019) dataset. After pertaining we use these learn weights for fine-tuning on our proposed bahasa dataset. We found that Our EnGen-TTS-L (Large) model outperforms other previous baselines at Bahasa text-to-speech synthesis. We Evaluate our model and baseline on MOS and CMOS metrics from Table 2 Our EnGen-TTS-L model achieves the highest MOS of 4.45 ± 0.13, surpassing all other models in the comparison. This indicates that listeners rated the speech generated by EnGen-TTS-L as more natural and closer to human speech than that of the competing models. Notably, EnGen-TTS outperforms NaturalSpeech and CLaM-TTS, which have MOS scores of 4.39 ± 0.19 and 4.41 ± 0.09, respectively.

Even without pre-training on Lj_speech and VCTK dataset, our Engen-TTS-L model achieves results comparable to NaturalSpeech, CLaM-TTS, and XTTS, shows EnGen-TTS-L is good at adapting to new languages, which is important for creating high-quality speech synthesis for languages like Bahasa.

In terms of CMOS EnGen-TTS also exhibits superior performance. It achieves the lowest (best) score in Metric A with -0.0101, indicating a closer alignment with target speech characteristics than other models.

We also evaluate baselines and our model on the RTF (real-time factor) for generating the speech. From Table 2 we found that EnGen-TTS is comparable fast, producing speech almost in real-time. The EnGen-TTS is not only fast, but it also produces high-quality speech unlike Other methods, GradTTS and GlowTTS, are a bit slower and don't generate speech quality as well as EnGen-TTS. Our findings are important for people who want to produce speech systems and want the real-time inference with high and robust quality of speech.

These results collectively affirm that EnGen-TTS not only advances the state-of-the-art in TTS by delivering the most natural-sounding speech but also maintains high performance across various evaluation metrics. Which highlights the strength of our model's architecture and training methodology.

5.4 Quntiative Results based on Multi-Lingual Dataset

We have trained the model in 7 languages, For Latin languages (Spanish, Portugeas, German, and Dutch) we used the CML tts-dataset (Oliveira et al., 2023) and for indic languages (Hindi, Marathi, and Tamil), we used the indic-speech (Srivastava et al., 2020b) dataset. For Fair comparison, We loaded the models that are pre-trained from LjSpeech and VCTK datasets. We evaluate each model with the MOS score. We with our EnGen-TTS-L, we used two more baselines to showcase our novelty.

The Table 3 presents a comparison of Mean Opinion Scores (MOS) across different languages for three models: VITS, NaturalSpeech, and EnGen-TTS-L. The languages evaluated include Spanish, Portuguese, German, Dutch, Hindi, Marathi, and Tamil, with varying amounts of training data for each language. EnGen-TTS-L consistently outperforms both VITS and NaturalSpeech across all the languages evaluated. For example, in Spanish, EnGen-TTS-L achieves a MOS of 4.28, significantly higher than both VITS (3.39) and NaturalSpeech (3.72). Similarly, in Portuguese, EnGen-TTS-L scores 4.37, surpassing VITS (3.41) and NaturalSpeech (3.78).

The performance gap is especially pronounced for languages like Marathi and Tamil, where EnGen-TTS-L achieves the highest scores of 4.87 and 4.78, respectively. In contrast, NaturalSpeech performs noticeably worse with MOS scores of 4.01 and 3.96 for Marathi and Tamil, respectively, while VITS scores lower at 3.92 for Marathi and 3.88 for Tamil.

Notably, EnGen-TTS-L also performs exceptionally well in Hindi with a MOS of 4.55, significantly outperforming both VITS (3.56) and NaturalSpeech (3.88). In German and Dutch, which have relatively more training data, EnGen-TTS-L continues to lead with MOS scores of 4.17 and

Languages	No of Hours	VITS	NaturalSpeech	EnGen-TTS-L
Spanish	23.34	3.39	3.72	**4.28**
Portuguese	17.82	3.41	3.78	**4.37**
German	18.03	3.27	3.47	**4.17**
Dutch	29.76	3.18	3.31	**4.14**
Hindi	15.11	3.56	3.88	**4.55**
Marathi	09.07	3.92	4.01	**4.87**
Tamil	10.56	3.88	3.96	**4.78**

Table 3: Comparison of MOS scores for different languages using VITS, NaturalSpeech, and EnGen-TTS-L models.

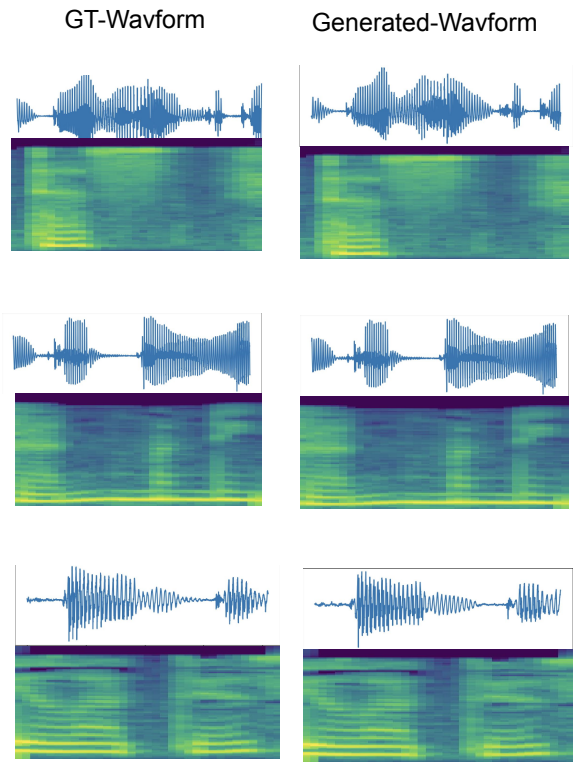

Figure 4: Comparison between Ground truth and Generated Audio

4.14, respectively, demonstrating consistent high-quality speech synthesis across languages, even for those with limited training data.

Overall, the results clearly indicate that EnGen-TTS-L excels in generating more natural-sounding speech across multiple languages, especially when compared to the baseline models, VITS and NaturalSpeech. This demonstrates the robustness of the EnGen-TTS-L model in multilingual TTS tasks, even with varying amounts of training data for each language.

5.4.1 Model Performance

The model generates audio at a 24 kHz quality. When it comes to pronouncing acronyms, the task can be challenging. A helpful strategy is to articulate each letter separately, spacing them out to improve clarity. For numerical data, converting digits into their word equivalents often yields better results. An important observation is that the model might inadvertently replicate the reference speaker's audio in its output, particularly when the input text closely mirrors the reference material. The overall quality of the output is heavily influenced by the caliber of the reference audio. Ideally, the reference should be between 4 to 6 seconds long and exclusively contain clear speech, free from any background noises. It's worth noting that employing a cartoon-like voice in audio references might lead to model failure, as such inputs are significantly different from the data used during the training process. The model's capacity is constrained to 604 audio tokens and 1024 text tokens, where 600 audio tokens equate to approximately 16 seconds of sound. Look in Figure 4 for a comparison between the Ground truth waveform and the EngenTTS Generated waveform. We can see that our model-generated output is very close to the ground truth.

5.4.2 Implementation details

The EngenTTS-L's Audio Codec language modeling module utilizes transformer architecture with 26 blocks, 16 attention heads, a hidden dimension of 1024, a feed-forward layer dimension of 1024. The average length of the waveform in LJspeech and VCTK is 9.8 seconds, for our Bahasa dataset average length of the audio is around 7 seconds. During training, we randomly crop the waveform to a random length between 2 seconds and 6 seconds. Its corresponding phoneme alignments are used as the phoneme prompt. We remove the con-

secutive repetitions in the force-aligned phoneme sequence. While training we keep max sequence length of 500. The models are trained using 3 NVIDIA RTX 3090Ti 24 GPUs with a batch size of 4 with gradient accumulation steps of 24 per GPU for 800k steps. We optimize the models with the AdamW optimizer, warm up the learning rate for the first 32k updates to a peak of 5×10^{-4}, and then linear decay it.

For the evaluation of inference performance, all models were tested under identical hardware conditions to ensure consistency and comparability. Specifically, we utilized an NVIDIA T4 GPU equipped with 16 GB of VRAM. Inference was performed with a batch size of 16, and each input had an average token length of 20. Under this setup, the inference speed was approximately 200 milliseconds per batch for all models. This uniform testing environment ensured that the performance metrics reported are directly comparable across all evaluated TTS systems.

5.5 Ablation

In our experimentation with different vocoders for generating high-quality audio from latent representations, Univnet emerged as the top performer, achieving a Mean Opinion Score (MOS) of 4.41 ± 0.12. This slight edge over MelGan (4.39 ± 0.11) and Wavgrad (4.36 ± 0.13) suggests its superiority in preserving speech quality and naturalness during waveform reconstruction (see Table 5). While HiFi-GANs (4.35 ± 0.13) exhibited comparable performance, its slightly lower MOS indicates room for further optimization. Overall, these results highlight the importance of vocoder selection in the Text-to-Speech pipeline, with Univnet demonstrating its potential for creating highly faithful and human-sounding synthetic speech.

Additionally, to explore the impact of model size on both perceptual quality and real-time efficiency, we conducted an ablation study as outlined in Table 4. The table presents results for three variants of our EnGen-TTS model, denoted as EnGen-TTS-S, EnGen-TTS-M, and EnGen-TTS-L, with varying parameters. As model size increases from 87M to 570M, we observe a corresponding improvement in Mean Opinion Score (MOS), indicating enhanced speech quality. Specifically, EnGen-TTS-L achieves a MOS of 4.42 ± 0.08, outperforming the smaller variants. However, this comes at the cost of increased Real-Time Factor (RTF), with EnGen-TTS-L demonstrating a slightly longer synthesis time (0.021) compared to EnGen-TTS-S (0.014). This trade-off between model size, perceptual quality, and synthesis speed provides valuable insights into tailoring the EnGen-TTS architecture based on specific application requirements and resource constraints.

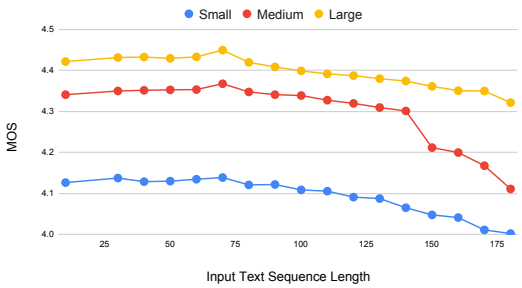

Figure 5: Input text sequence length vs MOS

We also conducted an ablation study focusing on the impact of different loss components. Table 6 summarizes the Mean Opinion Scores (MOS) obtained from three model variants: Lgan, Lmel, and the combination of both (Lgan + Lmel). The results indicate that incorporating both the adversarial loss (Lgan) and the mel-spectrogram loss (Lmel) leads to a MOS of 4.35 ± 0.12, showcasing a marginal improvement over individual losses. This nuanced exploration of loss components provides valuable insights into the synergy between adversarial and mel-spectrogram losses in our training pipeline, contributing to the optimization of our Bahasa TTS model for enhanced speech synthesis quality. Note: For every loss we are always computing L_{ce}, without it model can't be trained.

We conducted an ablation study on three different model sizes—large, medium, and small—focusing on their performance with varying lengths of input text sequences. Our observations indicate that for text sequences ranging from 5 to 75 tokens, there is minimal variation in the MOS metrics. However, as the sequence length exceeds 75 tokens, we noticed a decline in MOS metrics. This decline correlates with deteriorations in pronunciation and timbre quality of generated speech, along with an increase in the Word Error Rate (WER). As depicted in Figure 5, extending the sequence length beyond 100 tokens results in a significant decrease in MOS metrics, likely due to the models' inability to manage longer contexts effectively, leading to catastrophic forgetting which is discuss in the paper (Liu et al., 2024).

Model	Parameters	MOS	Hidden Dim	Attention Heads	No. of Blocks	RTF
EnGen-TTS-S	87M	4.12 ± 0.19	512	4	6	0.014
EnGen-TTS-M	280M	4.35 ± 0.12	768	8	13	0.016
EnGen-TTS-L	570M	4.42 ± 0.08	1024	16	26	0.021

Table 4: Ablation Based on Model Size

Models	MOS
Wavgrad (Chen et al., 2020)	4.36 ± 0.13
MelGan (Kumar et al., 2019)	4.39 ± 0.11
Univnet (Jang et al., 2021)	4.41 ± 0.12
Hifi-Gans (Kong et al., 2020b)	4.35 ± 0.13

Table 5: Ablation With Different Vo-Coder

Model	MOS
Lgan	4.30 ± 0.12
Lmel	4.32 ± 0.13
Lgan + Lmel	4.35 ± 0.12

Table 6: Ablation based on Loss

6 Conclusions

In conclusion, our study presents a pivotal advancement in Bahasa text-to-speech (TTS) synthesis, combining a richly curated dataset with a groundbreaking model design. Our comprehensive Bahasa TTS dataset, encompassing over 55 hours of audio and 52K recording, is a robust resource, crafted with inputs from proficient voice artists and varied textual content. The introduced model, EnGen-TTS, excels in performance, surpassing conventional benchmarks with its innovative architecture, which includes a multi-task T5 (m-T5) encoder and a neural codec language modeling module, without necessitating extra fine-tuning for Bahasa. This design not only enhances speech synthesis quality but also ensures computation efficiency, establishing a new standard in TTS technology. Our work not only pushes forward the boundaries of Bahasa TTS but also lays the groundwork for future developments in multilingual text-to-speech systems, promising high-quality and diverse linguistic applications.

7 Limitations

One limitation of our proposed method is its reliance on audio sampled at 22.05 KHz. This sampling rate is necessitated by the use of Meta's pre-trained Audio Encodec, which requires 22.05 kHz audio data. However, this presents a challenge for applications such as automatic voice calling, where telephony standards typically mandate an 8 kHz sampling rate. The required down-sampling from 22.05 KHz to 8 KHz results in a significant reduction in audio quality, manifesting as "muffled speech" due to the drastic decrease in sampling rate. Future work will focus on enabling high-quality audio generation directly at 8 kHz to better align with telephony requirements without compromising speech clarity.

Another limitation of our method lies in the maximum sequence length used during training, which is capped at 500 audio tokens. This constraint is well-suited for generating high-quality speech for shorter sentences or sentences containing up to 70-80 words. However, when the word count exceeds this limit, the generated speech may exhibit unnatural pauses or occasional missing words. This issue is likely due to catastrophic forgetting of longer contexts. Our future research will focus on increasing the context window up to 2048 audio tokens to better handle larger sentences or paragraphs, thereby improving the naturalness and continuity of generated speech.

References

2019. English multi-speaker corpus for cstr voice cloning toolkit. *https://doi.org/10.7488/ds/2645*.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, et al. 2024. Xtts: a massively multilingual zero-shot text-to-speech model. *arXiv preprint arXiv:2406.04904*.

Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. 2020. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. `https://keithito.com/LJ-Speech-Dataset/`.

Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. 2021. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. *arXiv preprint arXiv:2106.07889*.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Jaehyeon Kim, Keon Lee, Seungjun Chung, and Jaewoong Cho. 2024. Clam-tts: Improving neural codec language model for zero-shot text-to-speech. *arXiv preprint arXiv:2404.02781*.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020a. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020b. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033.

Gokul Karthik Kumar, Praveen S V, Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2023. Towards building text-to-speech systems for the next billion users. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kundan Kumar, Rithesh Kumar, Thibault De Boissiere, Lucas Gestin, Wei Zhen Teoh, Jose Sotelo, Alexandre De Brebisson, Yoshua Bengio, and Aaron C Courville. 2019. Melgan: Generative adversarial networks for conditional waveform synthesis. *Advances in neural information processing systems*, 32.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2024. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Hui Lu, Zhiyong Wu, Xixin Wu, Xu Li, Shiyin Kang, Xunying Liu, and Helen Meng. 2021. Vaenar-tts: Variational auto-encoder based non-autoregressive text-to-speech synthesis.

Frederico S Oliveira, Edresson Casanova, Arnaldo Candido Junior, Anderson S Soares, and Arlindo R Galvão Filho. 2023. Cml-tts: A multilingual dataset for speech synthesis in low-resource languages. In *International Conference on Text, Speech, and Dialogue*, pages 188–199. Springer.

Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. 2021. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR.

Devin Schumacher and Francis LaBounty Jr. 2023. Enhancing suno's bark text-to-speech model: Addressing limitations through meta's encodec and pretrained hubert. *Available at SSRN 4443815*.

Nimisha Srivastava, Rudrabha Mukhopadhyay, Prajwal K R, and C V Jawahar. 2020a. IndicSpeech: Text-to-speech corpus for Indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6417–6422, Marseille, France. European Language Resources Association.

Nimisha Srivastava, Rudrabha Mukhopadhyay, KR Prajwal, and CV Jawahar. 2020b. Indicspeech: text-to-speech corpus for indian languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6417–6422.

Onkar Susladkar, Prajwal Gatti, and Santosh Kumar Yadav. 2023. Slbert: A novel pre-training framework for joint speech and language modeling. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Xu Tan, Jiawei Chen, Haohe Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. 2024. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alexander Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *Arxiv*.

Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (mos) scale. *Computer Speech Language*, 19(1):55–83.

Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. 2018. Generalized end-to-end loss for speaker verification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4879–4883. IEEE.

Yuxuan Wang, R.J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc

Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. In *Proc. Interspeech 2017*, pages 4006–4010.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. 2023. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*.

Adrian Łańcucki. 2021. Fastpitch: Parallel text-to-speech with pitch prediction. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6588–6592.