A Hybrid Attention Mechanism to Improve Tacotron 2 Performance for Indonesian Text-to-Speech Synthesis

Angela Catherina*, Bima Prihasto*, Boby Mugi Pratama*, Li-Wei Kang†, and Jia-Ching Wang‡
* Department of Informatics, Institut Teknologi Kalimantan, Indonesia
† Department of Electrical Engineering, National Taiwan Normal University, Taiwan
‡ Department of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: bima@lecturer.itk.ac.id

***Abstract*—This paper presents a speech synthesis system for the Indonesian language using the Tacotron 2 architecture and HiFi-GAN vocoder. While Tacotron 2 has demonstrated strong performance in high-resource languages, adapting it for Indonesian remains underexplored. To address this, we propose a hybrid attention mechanism that combines location-sensitive and content-based attention to improve alignment accuracy and convergence during training. The system is trained on a custom Indonesian audiobook dataset prepared in LJSpeech format. We compare several model variants, including versions with and without phoneme conversion, and evaluate them using both subjective and objective metrics. Speech quality is assessed through Mean Opinion Score (MOS) ratings, while model performance is evaluated via attention weight visualization. Results show that the proposed hybrid attention mechanism accelerates alignment learning and produces speech with improved naturalness and intelligibility, confirming its effectiveness for low-resource TTS in Indonesian.**

I. Introduction

Text-to-speech (TTS) systems have evolved significantly, from concatenative methods [1], [2] to modern deep neural networks [3]–[5]. TTS plays a crucial role in applications such as virtual assistants, assistive tools for the visually impaired, and educational platforms [6], [7].

While high-quality speech synthesis has been achieved in high-resource languages, TTS research for low-resource languages like Indonesian remains limited [8]. Indonesian's consistent vowel sounds and simpler pronunciation patterns [9] present challenges for models trained on English, highlighting the need for TTS advancements in this language.

Tacotron 2 [4] is a widely adopted architecture that uses an encoder-decoder structure with an attention mechanism to generate mel-spectrograms, which are converted to waveforms by a vocoder. The model's attention mechanism is based on location-sensitive attention [10], which improves alignment by incorporating past attention weights. However, this attention mechanism may struggle in low-resource languages due to data scarcity, prompting the need for further improvements.

This paper introduces a modification to Tacotron 2's attention mechanism by integrating a content-based attention [11] module alongside the location-sensitive one. The two energies are computed independently and combined using learnable weights. This hybrid attention mechanism improves alignment precision while maintaining temporal stability.

The architecture is implemented using SpeechBrain's open-source recipe[1] and trained on a custom Indonesian speech dataset derived from open-license audiobooks. We evaluate the models through subjective MOS tests and objective alignment weight visualizations.

The contributions of this paper are as follows:

1) We introduce a new Indonesian speech dataset, IndoTTS-Book[2].
2) We propose a hybrid attention mechanism that combines location-sensitive and content-based attention using learnable weights.
3) We define four Tacotron 2-based model variants to isolate the effect of phoneme-level input and attention mechanism design.
4) We design three targeted MOS evaluation scenarios to assess speech quality under different model and training conditions.
5) We evaluate the models using subjective MOS and objective alignment weight visualization.

The rest of the paper is structured as follows: Section II describes the baseline Tacotron 2 architecture and the proposed hybrid attention mechanism. Section III outlines the experimental setup, including dataset preparation, model configurations, variation design, and evaluation metrics. Section IV presents and analyzes the results from alignment visualization and MOS evaluations. Finally, Section V summarizes the findings and highlights potential directions for future research.

II. Proposed Method

A. Tacotron 2

Tacotron 2 is an end-to-end TTS architecture that converts text into mel-spectrograms, which are then transformed into waveforms by a vocoder such as HiFi-GAN [12]. The encoder, composed of convolutional layers and bidirectional LSTM,

[1]https://github.com/speechbrain/speechbrain/tree/develop/recipes/LJSpeech/TTS/tacotron2

[2]https://github.com/RandomStrayCat/IndoTTS-Book

979-8-3315-7206-8/25/$31.00

processes the input text into high-level embeddings. The decoder, using an autoregressive LSTM network, generates mel-spectrogram frames while attending to relevant encoder outputs via an attention mechanism.

The attention mechanism in Tacotron 2 is location-sensitive attention [10], an extension of content-based attention [11]. It promotes monotonic alignment by incorporating past attention weights, aiding in the smooth traversal of the input sequence and preventing repeated or skipped tokens.

While Tacotron 2 performs well for high-resource languages like English, its application to low-resource languages like Indonesian faces challenges due to alignment instability. This motivates the need for enhancements in the attention mechanism.

B. Proposed Hybrid Attention Mechanism

We propose a hybrid attention mechanism that computes content-based and location-sensitive attention energies independently, using learnable scalar weights to combine them. The content-based energy is calculated by comparing the decoder hidden state q_t with the encoder outputs h_i using a nonlinear projection, as shown in (1), while the location-sensitive energy incorporates past attention weights processed through a convolutional layer, as shown in (4).

$$e_i^{(c)} = \tanh(W_q q_t + W_h h_i + b) \tag{1}$$

$$\alpha_t^{(c)} = \text{softmax}(e^{(c)}), \quad c_t^{(c)} = \sum_i \alpha_i^{(c)} h_i \tag{2}$$

$$F = \text{Conv1D}(\alpha_{t-1}) \tag{3}$$

$$e_i^{(l)} = \tanh(W_q q_t + W_h h_i + W_f F_i + b) \tag{4}$$

$$\alpha_t^{(l)} = \text{softmax}(e^{(l)}), \quad c_t^{(l)} = \sum_i \alpha_i^{(l)} h_i \tag{5}$$

The hybrid alignment energy is then computed as a weighted sum, as shown in (6):

$$e_t = \alpha \cdot e^{(l)} + \beta \cdot e^{(c)} \tag{6}$$

This approach allows the model to balance content relevance and location history, leading to improved alignment precision while retaining the benefits of temporal consistency. The proposed mechanism is implemented with minimal changes to the original decoder architecture. A block diagram of the full synthesizer with the hybrid attention mechanism is shown in Fig. 1.

III. Experiment Setup

This section describes the experimental configurations used to evaluate the proposed hybrid attention mechanism, covering the dataset, model architecture, model variants, and evaluation metrics.

TABLE I: Configuration of Tacotron 2 Model Variants

Model ID	Hybrid Attention	Phonemizer (IPA)
T-1	✗	✗
T-2	✗	✓
T-3	✓	✗
T-4	✓	✓

A. Dataset

The dataset used to train the models is a custom Indonesian speech corpus derived from the publicly available audiobook "Sejarah Dunia yang Disembunyikan" by Jonathan Black, narrated by Rahman Trahira. This dataset, sourced under an open license[3], consists of 2,087 utterances with durations ranging from 0.40 to 10.10 seconds (mean: 4.28 s), totaling 8,936 seconds. Preprocessing included downsampling from 44.1 kHz to 22.05 kHz, mono conversion, and dereverberation using FFmpeg. Segmentation was performed based on silence detection, and transcription was done using Whisper to ensure accurate timestamping. Metadata was generated following the LJSpeech standard, and IPA-based phoneme conversion was performed using Phonemizer[13] to support Indonesian phoneme processing in Tacotron 2.

B. Model Configuration

All models were implemented using SpeechBrain's Tacotron 2 recipe[4] and trained from scratch on the custom Indonesian dataset. The sequence-to-sequence architecture consists of an encoder (convolution layers + bidirectional LSTM) and a decoder (autoregressive LSTM with attention). Training was done for 700 epochs using the Adam optimizer with an initial learning rate of 10^{-3}, batch size 32, and L2 regularization. The loss function combined mean squared error (MSE) for mel-spectrograms and binary cross-entropy for stop token prediction. HiFi-GAN [12] was used as the vocoder, trained from scratch on the custom Indonesian dataset to better fit the acoustic domain of the synthesized mel-spectrograms. Experiments were conducted on an NVIDIA RTX 3090 GPU with CUDA acceleration.

C. Model Variations

This study evaluates four Tacotron 2-based model variants to isolate the effects of attention mechanism and phoneme-level input:

1) Modifications to the attention mechanism (default vs. hybrid attention),
2) The use of a phonemizer[13] for converting text into phonemes (IPA).

Table I summarizes these configurations.

D. Evaluation Metrics

To assess model performance, we used both subjective and objective evaluation methods:

[3] s.itk.ac.id/audiobooksource

[4] https://github.com/speechbrain/speechbrain/tree/develop/recipes/LJSpeech/TTS/tacotron2

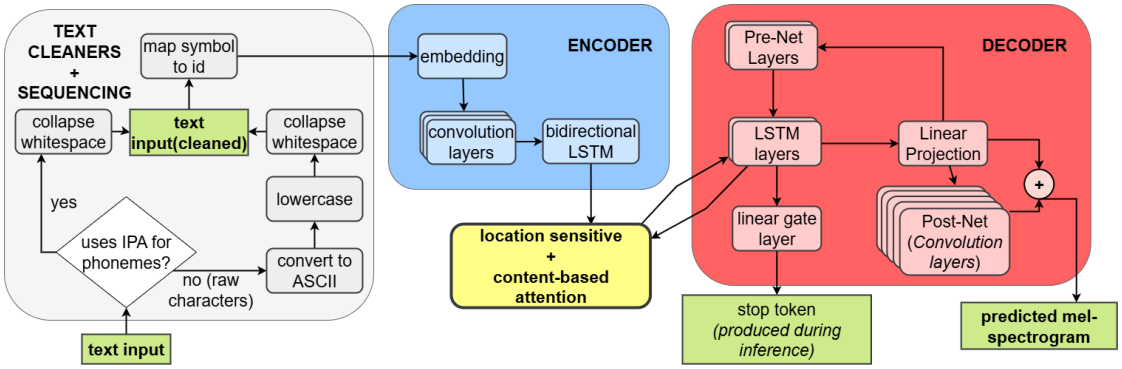


Fig. 1: Modified Tacotron 2 architecture with hybrid attention mechanism.

1) Subjective Evaluation (MOS): Speech quality was evaluated using a Mean Opinion Score (MOS) test [14], where participants rated speech based on *Naturalness* and *Intelligibility* using a 5-point scale with 0.5 intervals (1 = poor, 5 = excellent). Three evaluation scenarios were designed to compare model variants at different training stages and input conditions.

2) Objective Evaluation (Attention Weight Visualization): Model alignment was assessed through attention weight heatmaps [15] generated at selected training epochs. These visualizations track how well the decoder aligns input text with output frames, providing insights into model convergence and stability.

Together, these metrics allow a comprehensive evaluation of both the perceptual quality and model behavior across the Tacotron 2 variants.

IV. Results

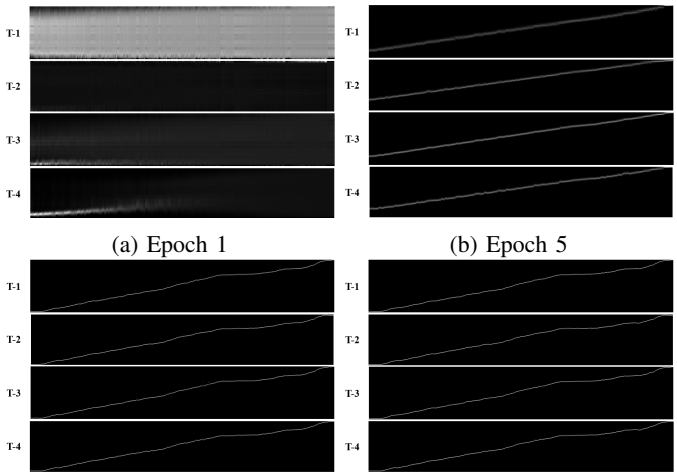
A. Alignment Weight Analysis

We evaluate model alignment by visualizing attention weights at various training milestones: Epoch 1, 5, 200, and 700 (refer to Fig. 2). The heatmaps display attention weights between the character sequence (y-axis) and mel-spectrogram frames (x-axis). Models with hybrid attention (T-3 and T-4) form partial alignments earlier (Epoch 1) and develop clearer diagonal structures by Epoch 5. By Epoch 700, all models produce strong diagonal alignments, with T-3 and T-4 demonstrating faster and more stable alignment learning.

B. Mean Opinion Score (MOS) Evaluation

To evaluate the perceptual quality of the synthesized speech, we conducted a Mean Opinion Score (MOS) test involving human listeners. Three evaluation scenarios were designed: S-1 (final model checkpoint at Epoch 700), S-2 (mid-training at Epoch 350), and S-3 (mixed text input at Epoch 700).

In Scenario S-1, as shown in Table II, T-3 (hybrid attention, no phonemizer) achieved the highest MOS of 4.10, outperforming other models in both naturalness (3.88) and intelligibility (4.32). T-2 consistently scored the lowest, highlighting the ineffectiveness of using a phonemizer alone.



(a) Epoch 1 (b) Epoch 5

(c) Epoch 200 (d) Epoch 700

Fig. 2: Alignment weight heatmaps for all model variants across selected epochs. Each heatmap contains rows for T-1, T-2, T-3, and T-4, respectively.

In Scenario S-2, Table III shows T-3 again achieving the highest overall MOS of 3.82, followed by T-1, T-4, and T-2, with T-2 and T-4 scoring lower than T-1 and especially T-3 due to degradation and instability in the synthesized speech.

In Scenario S-3, Table IV demonstrates that T-3 maintains the highest MOS of 4.09, with T-1 close behind. T-2 and T-4 again performed poorly, showing the limitations of IPA phonemized inputs in this configuration.

Across all three evaluation scenarios, T-3 consistently outperformed other models in both naturalness and intelligibility,

TABLE II: MOS Results for Scenario S-1 (Epoch 700, Uniform Text)

Model	Naturalness	Intelligibility	Overall MOS
T-1	3.67	4.30	3.98
T-2	2.85	3.21	3.03
T-3	**3.88**	**4.32**	**4.10**
T-4	3.13	3.65	3.39
Ground truth	4.20	4.60	4.40

TABLE III: MOS Results for Scenario S-2 (Epoch 350, Uniform Text)

Model	Naturalness	Intelligibility	Overall MOS
T-1	2.57	2.72	2.64
T-2	2.62	2.47	2.55
T-3	**3.58**	**4.07**	**3.82**
T-4	2.58	2.56	2.57

TABLE IV: MOS Results for Scenario S-3 (Epoch 700, Mixed Text)

Model	Naturalness	Intelligibility	Overall MOS
T-1	3.70	4.29	4.00
T-2	2.73	2.97	2.85
T-3	**3.88**	**4.29**	**4.09**
T-4	3.02	3.12	3.07
Ground truth	4.22	4.63	4.43

confirming the effectiveness of the hybrid attention mechanism. Interestingly, T-1 outperformed T-4 in all scenarios, suggesting that the phonemizer-based IPA transcription format may disrupt Tacotron 2's training, which processes text at the character level. Despite the presence of robust attention mechanisms, models T-2 and T-4 suffered from alignment issues due to mismatches in the IPA representation.

These findings confirm that while hybrid attention improves synthesis quality and convergence speed, it is most effective when paired with compatible input formats. Further investigation into phoneme conversion methods may help to fully leverage the benefits of Tacotron 2's attention mechanism.

V. Conclusions

This paper proposed a hybrid attention mechanism for Tacotron 2 that independently computes content-based and location-sensitive energies, combined via learnable weights. The method was evaluated on **IndoTTS-Book**, a new Indonesian TTS dataset sourced from open-license audiobook recordings. We trained four model variants to assess the effects of attention design and phoneme-level input.

Key findings include: (1) the hybrid attention mechanism improved alignment quality and accelerated alignment training; (2) the T-3 model (hybrid attention without phonemizer) consistently achieved the highest MOS scores in naturalness and intelligibility.

Overall, the proposed mechanism enhances synthesis quality and learning efficiency in Indonesian TTS. Future work may explore phoneme-aware input encodings or extend this approach to other low-resource languages and non-autoregressive architectures.

References

[1] A. Chalamandaris, S. Karabetsos, P. Tsiakoulis, and S. Raptis, "A unit selection text-to-speech synthesis system optimized for use with screen readers," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 3, pp. 1890–1897, 2010.

[2] S. S. Hande, "A review of concatenative text to speech synthesis," *Int. J. Latest Technol. Eng. Manag. Appl. Sci. IJLTEMAS*, vol. 3, no. 9, pp. 12–15, 2014.

[3] Y. Wang, R. Skerry-Ryan, D. Stanton, *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[4] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2018, pp. 4779–4783.

[5] Y. Ren, C. Hu, X. Tan, *et al.*, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[6] P. Taylor, *Text-to-speech synthesis*. Cambridge university press, 2009.

[7] M. R. Hasanabadi, "An overview of text-to-speech systems and media applications," *arXiv preprint arXiv:2310.14301*, 2023.

[8] M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, "A survey on recent approaches for natural language processing in low-resource scenarios," *arXiv preprint arXiv:2010.12309*, 2020.

[9] B. Andi-Pallawa and A. F. A. Alam, "A comparative analysis between english and indonesian phonological systems," *International Journal of English Language Education*, vol. 1, no. 3, pp. 103–129, 2013.

[10] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[13] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *Journal of Open Source Software*, vol. 6, no. 68, p. 3958, 2021. DOI: 10.21105/joss.03958. [Online]. Available: https://doi.org/10.21105/joss.03958.

[14] P. L. Salza, E. Foti, L. Nebbia, and M. Oreglia, "Mos and pair comparison combined methods for quality evaluation of text-to-speech systems," *Acta Acustica united with Acustica*, vol. 82, no. 4, pp. 650–656, 1996.

[15] F. Zalkow, P. Govalkar, M. Müller, E. A. Habets, and C. Dittmar, "Evaluating speech–phoneme alignment and its impact on neural text-to-speech synthesis," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.