Adversarial Learning for Duration Prediction in Indonesian Text-to-Speech: Modification to Stochastic and Deterministic Predictors

Yoga Tiara Wiguna*, Bima Prihasto*, Boby Mugi Pratama*, Chia-Hung Yeh†, 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*—Text-to-Speech (TTS) technology has significantly progressed with deep learning, especially through models like Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS). However, improving audio quality particularly in duration diversity remains a challenge, especially for languages like Indonesian due to limited datasets and research. This study compares the performance of VITS using Stochastic Duration Predictor (SDP) and Deterministic Duration Predictor (DDP), while also exploring the impact of adversarial training on duration prediction. Evaluation employed subjective Mean Opinion Score (MOS) and objective Cosine Similarity using Resemblyzer. Two datasets were used: 343 formal audio samples and 1250 mixed (formal and informal) samples. The more diverse dataset achieved better results, with a cosine similarity of 0.91124 and a MOS of 4.54. Findings indicate that SDP produces more natural durations, and adversarial learning enhances audio quality through better duration modeling.**

I. Introduction

Text-to-Speech (TTS) is an artificial intelligence-based technology that generates sound according to text input [1]. This technology mimics the way the human brain works in understanding context, meaning, and language structure. TTS research has so far focused on high-resource languages such as English. In Indonesia, the development of TTS is still minimal due to challenges in its implementation. This has caused Indonesian TTS to not develop optimally, lagging behind countries such as the UK, China, Japan, and Korea. As a result, Indonesian-language TTS applications are still rarely available for public use [2].

Some research on Text-To-Speech has been done by Jungil Kong et al [3], in this study introduced the Hifi-Gan model which serves to produce high-fidelity audio by involving a network of generators to create audio and discriminators in charge of distinguishing the original audio sample with the results of the generator. The next research conducted by Jaehyeon Kim et al [4], this research introduces a parallel end-to-end TTS method with a variational inference approach and an adversarial training process.

Despite the success of previous single-stage models [4], the generated audio still occasionally exhibited unnaturalness due to suboptimal duration prediction, indicating that modeling speech duration remains an area for improvement. Most prior studies have relied solely on subjective evaluation using Mean Opinion Score (MOS) and were primarily conducted using English-language datasets. In contrast, this study focuses on the Indonesian language, which presents unique linguistic and resource-related challenges. It adopts a two-pronged evaluation approach by combining subjective MOS with objective Cosine Similarity metrics using Resemblyzer. Furthermore, earlier works did not incorporate adversarial learning into the duration predictor.

In this study proposes applying adversarial learning to the duration predictor in the VITS model, both Stochastic (SDP) and Deterministic (DDP), as shown in Figure 1. The duration predictor acts as a generator using phoneme and noise input, while real duration data comes from Monotonic Alignment Search (MAS). [5]. The discriminator then compares duration values from the generator and MAS to classify them as REAL or FAKE. This study also compares the performance of stochastic and deterministic duration predictors in Indonesian TTS, evaluates the impact of adversarial learning on duration modeling, and analyzes the dataset to assess model generalization. The dataset and source code are made publicly available. We make the dataset and source publicly available[1][2]. The following contributions are made to our paper:

- We present a new publicly available Indonesian speech dataset recorded by the researcher "Yoga Tiara Wiguna" to support future TTS research.
- We enhance the duration predictor by incorporating adversarial learning to improve duration modeling
- We train the model on two datasets: a small formal transcript set and a larger mixed formal-informal set to examine the impact of dataset size and diversity.
- We evaluate various model configurations by altering the duration predictor type, applying adversarial learning, and comparing performance across both datasets.

[1]https://github.com/Yogatiara/ITKTTS-IDN.git
[2]https://github.com/Yogatiara/VITS_Indonesia.git

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

II. Method

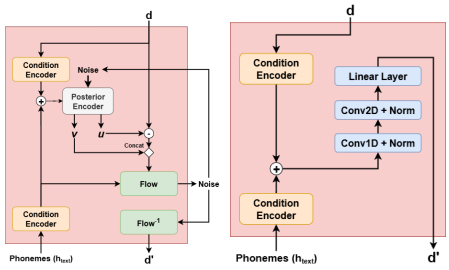
A. Duration Predictor

1) Stochastic Duration Predictor (SDP): The VITS model uses a Stochastic Duration Predictor (SDP) Figure 1a to estimate the duration of audio pronunciation. SDP receives h_{text}, the hidden representation of c_{text}, as input. During training, SDP maximizes the log-likelihood using normalizing-flow. Since duration d is an integer, two random variables u and v are introduced: u for dequantization to convert d into a continuous value for flow processing [6], and v for data augmentation to capture diverse duration distributions [7].

The equation used is based on the Evidence Lower Bound (ELBO), where Equation 1 is used to calculate the log-likelihood probability $\log p_\theta(d \mid c_{\text{text}})$.

$$\log p_\theta(x \mid c_{\text{text}}) \geq \mathbb{E}_{q_\Phi(u,v|d,c_{\text{text}})}\left[\log\left(\frac{p_\theta(d-u, v \mid c_{\text{text}})}{q_\Phi(u, v \mid d, c_{\text{text}})}\right)\right] \tag{1}$$

The training loss for the duration predictor (L_{dur}) is calculated as a negative variational lower bound by adding a minus operator to the variational lower bound of the log-likelihood equation.



(a) (b)

Fig. 1: Comparison of the architectural differences between duration predictors with adversarial learning: (a) Stochastic Duration Predictor (SDP) and (b) Deterministic Duration Predictor (DDP).

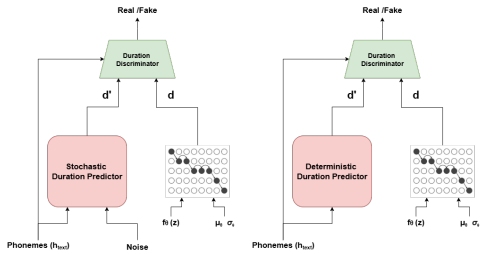
2) Deterministic Duration Predictor (DDP): Duration Predictor is not only Stochastic Duration Predictor which is probabilistic, but there is also a deterministic one, namely Deterministic Duration Predictor (DDP). DDP is different from SDP, based on Figure 1b DDP produces a definite value by training Autoregressive TTS [8] to produce a mel-spectogram with a sequential sequence of input in the form of a phoneme, then the Duration Extractor is tasked with predicting how long the spoken duration is based on the attention matrix of the mel-spectogram generated earlier.

The loss function used in DDP is Mean Squared Error (MSE) to calculate the difference between the actual duration d from the result of text-to-speech alignment and $\hat{d}$, which is the result of DDP prediction.

B. Proposed Modification of VITS Model

Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (VITS) is a text-to-speech model with an end-to-end system, the training process is parallel, which is directly carried out from start to finish in one flow without the involvement of two or more models [4]. In the VITS model, the duration of each part of a word or sentence to be spoken is determined using two types of duration predictors: SDP (Stochastic Duration Predictor) and DDP (Deterministic Duration Predictor).

Duration predictor in this study is implemented with adversarial learning [9]. The duration predictor is trained jointly with a discriminator network. In this setup, the duration predictor acts as a generator: it receives h_{text} and Gaussian noise z as inputs if the Stochastic Duration Predictor (SDP) is used [4], while the Deterministic Duration Predictor (DDP) [5] only takes h_{text} as input. Both predictors produce a predicted duration $\hat{d}$, which is then passed to the discriminator. The discriminator learns to distinguish whether the predicted duration is close to or far from the true duration. The architecture for adding adversarial learning to the duration predictor is illustrated in Figure 2a for SDP and Figure 2b for DDP.



(a) (b)

Fig. 2: a) Stochastic Duration Predictor (SDP) with noise input, while (b) uses a Deterministic Duration Predictor (DDP) without noise. Both use a Duration Discriminator to differentiate real and predicted durations, enhancing duration modeling quality.

III. Experiment

A. Dataset

This research uses audio datasets recorded using the voice of a 22 year old male writer. The author recorded using Indonesian with a mixed Javanese and Samarinda accent as many as 1250 audios which were divided into 80% training data and 20% validation data. Dataset collection was carried out in two stages, the first collected 343 datasets and then continued until 1250 datasets were collected. Initially, 343 datasets were collected, characterized by formal, uniformly spoken, and unexpressive sentences. A subsequent collection of 1250 datasets included a variety of formal and informal sentences. dataset details can be seen in the Tabel I.

Recording was done by reading transcripts taken from the TITML-IDN dataset [10], ASR-SIndoDusc [11], ASR-IndoCSC [12], and audiobooks from a book entitled The Art of War in Indonesian by accessing the YouTube channel

TABLE I: Dataset Specifications

Statistic	Value
Number of Datasets	1250 samples (80% training and 20% validation)
Total Duration of All Audio	02 hours 04 minutes
Sample Rate	22,005 Hz
Audio Format	WAV, mono, 16-bit PCM
Language	Indonesian
Speaker Age	22 Years Old
Accent	Javanese and Samarinda Accent
Number of Speaker	1 Speaker (Yoga Tiara Wiguna)

bacasuara on a video entitled "SUN TZU: The Art of War". The video was converted into a transcript using Unmixr[3]. Audio was recorded using a ZOOM H1N microphone placed approximately 2.5 cm from the speaker's mouth in a low-noise room. The recordings were processed using Audacity for quality management. The transcript structure used follows the LJ Speech dataset[4]

B. Preprocessing

1) Audio Silence Trimming: Cutting out the quiet part uses a tool called librosa. The configuration used is the `top_db` parameter in decibels (dB) which is the threshold for detecting silence in audio. First, librosa tries to find the highest energy signal with Root Mean Square (RMS) and then the highest signal is obtained, then the difference between the `top_db` value and the highest energy signal is sought. The difference value is used to be the maximum limit of the signal that will be cut [18]. The `top_db` used in this research is 20 decibels.

2) Transcript Cleaning: This research uses phonemizer [13] to clean the transcript. Phonemizer is a library in the python language that functions to convert transcripts from text that is usually written with letters of the alphabet into phoneme form. Phonemizer has several backends, one of which is esPeak, esPeak provides more than 100 languages including Indonesian by converting raw transcripts into phonetic form using the International Phonetic Alphabet (IPA) [4].

3) Audio Spectrum Transformation: Before entering the training process, the audio is converted into a linear spectrogram via Short-time Fourier Transform (STFT) [14]. In this process, the audio is segmented into small frames using a window size of 1024 samples, and each frame is shifted by 256 samples based on the hop size. Each frame is then transformed into the frequency domain using a Fast Fourier Transform (FFT) with a size of 1024. This results in a two-dimensional matrix where rows represent frequency components and columns represent time progression.

C. Training Setting

Model training on research using a variety of Indonesian datasets as much as 343 and 1250 data. In 343 datasets, the spoken sentences tend to be formal with the characteristics of word pronunciation and speech styles that tend to be uniform, sentence structures using standard language rules, and minimal intonation variations which are not too emotional or expressive. Furthermore, in the collection of 1250 datasets, the sentences spoken are quite varied where there are formal and informal sentences. The pronunciation of informal sentences does not always follow the standard rules, the use of intonation is quite varied, and the style of speech is natural like everyday conversation. This training utilized several configurations as outlined in the Table II.

TABLE II: VITS Model Training Configuration [4].

Parameter	Value
eval interval	500
epoch	10000
Learning rate	2×10^{-4}
betas	$[\beta_1 = 0.8, \beta_2 = 0.99]$
eps	0.000000001
Batch size	32
Lr decay	0.999875
Segment_size	8192
c_mel	45
c_kl	1.0

The device used for training has the specifications of an Intel Core i9 32 Core processor, NVIDIA Geforce RTX 3090 24 GB VRAM graphic card, and 64 GB RAM. Based on Table III, training is conducted using eight scenarios using the VITS-A - VITS-B scenario. These eight scenarios are distinguished by the number of datasets, the use of duration predictor types with and without the addition of adversarial learning to the duration predictor.

D. Model Evaluation

1) Objective Evaluation: Model evaluation with an objective approach involves statistical measurements. The method used in this research is resemblyzer with an approach using cross-similarity by applying the cosine similarity function [15], [16]. Cosine similarity measures the average similarity between pairs of speech files with identical content, evaluating aspects such as voice character, accent, and articulation. This calculation uses vector representations of each speech pair. The process is performed using the open-source tool Resemblyzer[5]

This evaluation uses 50 and 80 transcript test data to assess the consistency of average values across model outputs, using identical transcripts. Cosine similarity results are rounded to five decimal places to capture small differences between scenarios.

2) Subjective Evaluation: The subjective evaluation used the Mean Opinion Score, Mean Opinion Score (MOS), introduced by the International Telecommunication Union (ITU) in 2006, is a subjective evaluation method involving participants to rate model performance. It uses a 5-point scale (excellent, good, fair, poor, bad). Mos evaluation involved 50 respondents to assess the audio quality. Two aspects were assessed: voice naturalness and pronunciation accuracy. The evaluation included both formal "Prinsip-prinsip ini juga sudah diterapkan dalam politik, bisnis, dan interaksi sehari-har" and

[3]https://unmixr.com/speech-to-text-converter-online/

[4]https://keithito.com/LJ-Speech-Dataset/

[5]https://github.com/resemble-ai/Resemblyzer

TABLE III: Skenario Eksperimen Model VITS

Scenario	Number of Dataset	Type *of Duration Predictor*	Adversarial Learning Applied to Duration Predictor
VITS-A	343	*Stochastic Duration Predictor (SDP)*	✗
VITS-B	343	*Stochastic Duration Predictor (SDP)*	✓
VITS-C	343	*Deterministic Duration Predictor (DDP)*	✗
VITS-D	343	*Deterministic Duration Predictor (DDP)*	✓
VITS-E	1250	*Stochastic Duration Predictor (SDP)*	✗
VITS-F	1250	*Stochastic Duration Predictor (SDP)*	✓
VITS-G	1250	*Deterministic Duration Predictor (DDP)*	✗
VITS-H	1250	*Deterministic Duration Predictor (DDP)*	✓

informal "Lihat situasi kondisi dulu, contohnya pantai mana aja itu?" transcripts. The questionnaire was distributed via Jotform https://www.jotform.com/ with randomized audio order to reduce bias.

IV. Result

1) Objective evaluation: The first evaluation was conducted on a dataset of 343, the model using SDP has a higher average cosine similarity value than using DDP. Further evaluation using 1250 datasets, The average cosine similairty results in this evaluation have improved in terms of using the duration predictor type and adding adversarial learning to the duration predictor.

TABLE IV: Information on Average Cosine Similarity Results

Model	Dataset Size	Testing Data Amount	Cosine Similarity
VITS with DDP	343	50	0.90924
VITS with DDP + *adversarial learning*	343	50	0.91130
VITS	343	50	**0.91839**
VITS + *Adversarial Learning for Duration Predictor*	343	50	0.91158
VITS with DDP	343	80	0.92334
VITS with DDP + *adversarial learning*	343	80	0.92358
VITS	343	80	**0.92555**
VITS + *Adversarial Learning for Duration Predictor*	343	80	0.92208
VITS with DDP	1250	50	0.900021
VITS with DDP + *adversarial learning*	1250	50	0.90048
VITS	1250	50	0.90105
VITS + *Adversarial Learning for Duration Predictor*	1250	50	**0.90385**
VITS with DDP	1250	80	0.91114
VITS with DDP + *adversarial learning*	1250	80	0.91124
VITS	1250	80	0.91161
VITS + *Adversarial Learning for Duration Predictor*	1250	80	**0.91186**

2) Subjective evaluation: Based on Figure 3, the calculation of the average value on the assessment aspects of audio naturalness and pronunciation suitability of the eight scenarios is carried out. The figure indicates that increasing the dataset from 343 to 1250 in the MOS evaluation can improve the quality of the model from both aspects of assessment, namely in terms of understanding phonetic information from the text and naturalness in terms of the use of intonation, prosody, speech flow, and pronunciation expression. This also indicates that the model is able to create audio that sounds faithful to the input text, but is still not good at producing natural-sounding audio. Next, a comparison of the MOS values among

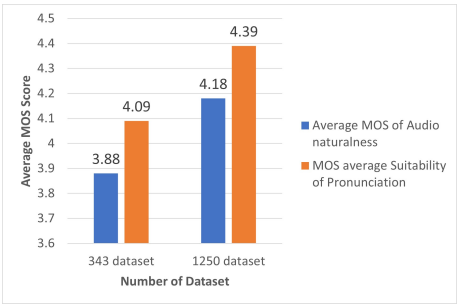
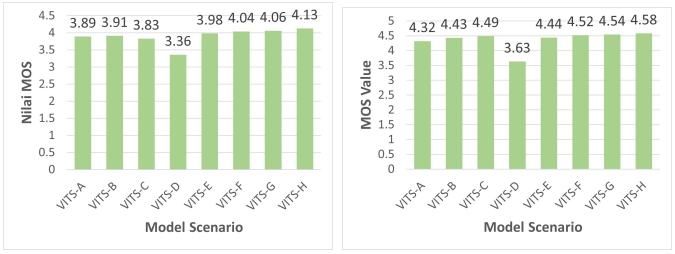


Fig. 3: Comparison Diagram of Average Values of Audio Naturalness and Pronunciation.

the eight scenarios was carried out, where the MOS value used was derived from the calculation of the average value between the two assessment aspects, namely audio naturalness and suitability of pronounciation.



(a) (b)

Fig. 4: Comparison of MOS evaluation results for each model scenario: (a) MOS values using formal transcripts, and (b) MOS values using informal transcripts.

Based on Figure 4a, experiments using 343 formal audio samples showed that the Stochastic Duration Predictor (SDP) achieved the highest MOS score of 4.49, while the Deterministic Duration Predictor (DDP) scored 4.32 and increased slightly to 4.43 with adversarial learning. However, adding adversarial learning to SDP caused a drop in MOS to 3.63, likely due to the small dataset size, which made the training unstable. When the dataset was expanded to 1250 samples with more varied content, performance improved across all

configurations. DDP reached 4.44, increasing to 4.52 with adversarial learning, SDP improved to 4.54, and the highest score of 4.58 was achieved with the combination of SDP and adversarial learning, indicating that sufficient and diverse data can stabilize adversarial training and enhance model quality.

As shown in Figure 4b, experiments using informal transcripts with 343 datasets showed that DDP achieved an MOS of 3.89, which slightly increased to 3.91 with adversarial learning. Meanwhile, SDP resulted in lower scores 3.83 without and 3.36 with adversarial learning. The decline is likely due to the lack of variation in the dataset, which mainly consists of formal speech. Since SDP relies on probabilistic distributions, it requires more diverse data to generate natural prosody and intonation; otherwise, the output may sound unnatural.

TABLE V: Mean Opinion Score (MOS) Evaluation Results

Model	Number of Dataset	Transkrip	MOS
VITS (DDP)	343	Formal	4.32
VITS (DDP)	343	Informal	3.89
VITS (*DDP + adversarial learning*)	343	Formal	4.43
VITS (*DDP + adversarial learning*)	343	Informal	**3.91**
VITS (SDP)	343	Formal	**4.49**
VITS (SDP)	343	Informal	3.83
VITS (*SDP + adversarial learning*)	343	Formal	3.63
VITS (*SDP + adversarial learning*)	343	Informal	3.36
VITS (DDP)	1250	Formal	4.44
VITS (DDP)	1250	Informal	3.98
VITS (*DDP + adversarial learning*)	1250	Formal	4.52
VITS (*DDP + adversarial learning*)	1250	Informal	4.04
VITS (SDP)	1250	Formal	4.54
VITS (SDP)	1250	Informal	4.06
VITS (*SDP + adversarial learning*)	1250	Formal	**4.58**
VITS (*SDP + adversarial learning*)	1250	Informal	**4.13**

However, when the dataset was expanded to 1250 samples, similar to the formal transcript experiments, all configurations showed improved performance. DDP reached an MOS of 3.98, and with adversarial learning, it increased to 4.04. SDP achieved an MOS of 4.06, and the best result 4.13 was obtained when combining SDP with adversarial learning. This suggests that a larger and more varied dataset helps unlock the full potential of both duration predictors, especially SDP.From several experiments conducted on the MOS evaluation also outlined in Table V.

V. Conclusions

This research modifies an Indonesian TTS model by exploring two duration predictors—SDP and DDP—and the impact of adversarial learning. The modified model shows strong audio generation, with cosine similarity scores of 0.90105 (50 samples) and 0.91124 (80 samples). Subjective MOS evaluations reached 4.54 for formal and 4.06 for informal transcripts, indicating natural and accurate speech output. The comparison shows that SDP outperforms DDP in both cosine similarity and MOS, producing more natural and varied audio despite its complex training. Adversarial learning further improves both predictors by enhancing duration naturalness. The best results were achieved with SDP and adversarial learning, reaching a cosine similarity of 0.91186 and a MOS of 4.58 for formal transcripts.

References

[1] M. Bera, "Text to speech synthesis," 2021.

[2] M. R. Hasanabadi, *An overview of text-to-speech systems and media applications*, 2023. arXiv: 2310.14301 [eess.AS]. [Online]. Available: https://arxiv.org/abs/2310.14301.

[3] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *ArXiv*, vol. abs/2010.05646, 2020.

[4] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *ArXiv*, vol. abs/2106.06103, 2021.

[5] J. Kim, S. Kim, J. Kong, and S. Yoon, "Glow-tts: A generative flow for text-to-speech via monotonic alignment search," *ArXiv*, vol. abs/2005.11129, 2020.

[6] J. Ho, X. Chen, A. Srinivas, Y. Duan, and P. Abbeel, "Flow++: Improving flow-based generative models with variational dequantization and architecture design," *ArXiv*, vol. abs/1902.00275, 2019.

[7] J. Chen, C. Lu, B. Chenli, J. Zhu, and T. Tian, "Vflow: More expressive generative flows with variational data augmentation," *ArXiv*, vol. abs/2002.09741, 2020.

[8] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *AAAI Conference on Artificial Intelligence*, 2018.

[9] J. Kong, J. Park, B. Kim, J. Kim, D. Kong, and S. Kim, "Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design," *ArXiv*, vol. abs/2307.16430, 2023.

[10] D. P. Lestari, K. Iwano, and S. Furui, "A large vocabulary continuous speech recognition system for indonesian language," 2006.

[11] MagicHub, *Asr-sindodusc*, https : / / magichub . com / datasets/indonesian-scripted-speech-corpus-daily-use-sentence/, Accessed: 4 April 2025, 2021.

[12] MagicHub, *Asr-indocsc*, https://magichub.com/datasets/indonesian-conversational-speech-corpus/, Accessed: 4 April 2025, 2021.

[13] M. Bernard and H. Titeux, "Phonemizer: Text to phones transcription for multiple languages in python," *J. Open Source Softw.*, vol. 6, p. 3958, 2021.

[14] D. Goyal and B. S. Pabla, "Condition based maintenance of machine tools—a review," *Cirp Journal of Manufacturing Science and Technology*, vol. 10, pp. 24–35, 2015.

[15] W. Gomaa and A. Fahmy, *A survey of text similarity approaches. international journal of computer applications 68 (04 2013)*, 2013.

[16] C.-W. Bang and C. Chun, "Effective zero-shot multi-speaker text-to-speech technique using information perturbation and a speaker encoder," *Sensors*, vol. 23, no. 23, p. 9591, 2023.