The Evaluation of Performance Related to Noise Robustness of VITS for Speech Synthesis

Jvlie Yang*

Department of Computer Science and Technology, Tongji University, Shanghai, China

*Corresponding author: 2152594@tongji.edu.cn

Abstract. In recent years, the utilization of voice interfaces has gained significant popularity, with speech synthesis technology playing a pivotal role in their functionality. However, speech synthesis technology is susceptible to noise interference in practical applications, which may lead to a decrease in the quality of speech synthesis. In this paper, the noise robustness of the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) model was investigated, which has shown promising results in speech synthesis tasks. This study conducted experiments using six different texts and evaluated the speech synthesis results using three metrics: Mean Opinion Score (MOS), Disfluency Prediction (DIS), and Colorfulness Prediction (COL). The experiments consist of a control group and six experimental groups, which include two types of noise, Additive White Gaussian Noise (AWGN) and real-world noise, at three different signal-to-noise ratios (SNRs). The results demonstrated that both types of noise can significantly reduce the MOS scores of the synthesized speech, with a more severe decrease at lower SNRs. In terms of DIS and COL scores, the VITS model exhibits superior performance with real-world noise compared to AWGN noise, especially at lower SNRs. Moreover, even at an SNR of 3, the VITS model can still generate intelligible speech, which demonstrates its high noise robustness. The findings have important implications for the design of robust speech synthesis models in noisy environments. Future studies may focus on exploring more advanced noise-robust models or investigating the application of these models in practical voice interfaces.

Keywords: Text-To-Speech; Variational AutoEncoder; noise-robust; Mean Opinion Score.

1. Introduction

Speech is one of the most commonly used media for daily human communication, while text serves as the primary carrier for human-machine interaction. Intelligent personal assistants like Microsoft's Cortana and Apple's Siri demonstrate the value of speech technology in human-machine interaction, making it a crucial research area. Text-to-speech (TTS) technology aims to enable computers to convert any input text into speech output. The process of converting text to speech involves two main steps. The first step is text analysis, which aims to convert the input text into symbols or speech representations to build acoustic phoneme models. The second step involves generating natural-sounding speech output from these models. Over the years, the focus of TTS has evolved from intelligibility to naturalness and fluency, with earlier methods such as the Synchronous OverLap Add (PSOLA) algorithm being replaced by more recent approaches based on Hidden Markov Models (HMM) and Deep Learning (DL) [1-3].

As research in the field of speech synthesis progresses, increasing emphasis is being placed on the application of Artificial Intelligence (AI) [4]. There are two categories of AI technology used in this field: traditional machine learning-based methods and deep learning-based methods. This paper focuses on the latter category and presents the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model, a deep learning model based on Variational Autoencoder (VAE) and Generative Adversarial Networks (GANs) [5]. In the past few years, there has been a notable surge in research studies investigating the potential of deep learning in text-to-speech (TTS) technology [6]. For instance, Z. H. Ling et al. [7] proposed a Spectral Envelope Modeling method (SPE-RBM) based on Restricted Boltzmann Machine (RBM). Their research shows that this method can significantly improve the naturalness of speech synthesis based on traditional HMM models. T. Falas et al. [8] used a neural network-based multilayer feed-forward architecture for transforming

Greek text to speech. They successfully determined that a configuration of 60-80 hidden neurons appeared to be the best for training and testing, resulting in the highest classification efficiency. In addition, H. Tachibana et al. proposed a TTS system based on CNN. This system, compared to RNN-based systems [9], does not require a large amount of training time, saving time and economic costs. Nevertheless, J. Y. Lee also creatively proposed a method to use Adversarially trained Variational Recurrent Neural Network (AdVRNN) to design and create speech parameter sequences in [10]. This effectively solves the problem of over-smoothing in the speech synthesis process and increases the dynamic range of the synthesized results.

Undoubtedly, VITS, a Conditional Variational Autoencoder with Adversarial Learning developed by KAIST, has emerged as one of the leading performers in the field of Text-to-Speech (TTS). The model has a fast convergence speed and high naturalness and achieves a Mean Opinion Score (MOS) comparable to ground truth [6]. Despite the remarkable naturalness of VITS, the research community tends to overlook its noise resistance, a crucial aspect of TTS. However, noise resistance is an important aspect of TTS because in the real-world scenarios, TTS systems are frequently subjected to varying degrees of noise, resulting in inaccurate, unnatural, and even unintelligible speech synthesis outcomes. In practical applications, the quality of noise resistance directly affects the reliability and usability of TTS systems and is therefore of great significance for improving user experience and reducing the need for manual intervention.

This paper will study the noise resistance of the VITS model and its performance under different conditions by adding noise of different degrees/types to the dataset. The data used in this study will be collected from daily life, which is also the main source of data for TTS. The evaluation of the speech synthesis effect in this study will be based on the score of NISQA, which is a Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction [11]. This will be a relatively objective evaluation criterion. Hopefully, this will provide a perspective for the subsequent optimization of this model, and even the future optimization direction of TTS.

2. Method

2.1. Dataset description and preprocessing

The dataset used in this study consists of audio recordings of a Chinese female speaker reading an article in Chinese. The original audio was lengthy and subsequently divided into 1000 shorter audio clips ranging from 5 to 30 seconds in duration using an audio segmentation tool. As the speaker was reading an article, the speech contained in the dataset was largely devoid of any extraneous background noise, and the speaker maintained a moderate pace of delivery, thus rendering it suitable for use in training a speech model. The audio files in the dataset were in a specific format, which was also used in the original VITS paper. Specifically, the audio files were single-channel, sampled at a rate of 22050 Hz, and quantized using Pulse Code Modulation (PCM) with 16-bit resolution [9].

To investigate the noise resistance of the VITS speech synthesis model, noise was added to this high-quality, noise-free dataset. The study focused on two key aspects of noise resistance: the impact of various noise levels and types. In terms of the first aspect, three levels of noise were added to the audio clips, corresponding to signal-to-noise ratios of 3, 10, and 30 dB. Gaussian white noise was used for this purpose, as it is a common type of noise found in many real-world environments. For the second aspect, two types of noise were used: Gaussian white noise and irregular noise collected from real-life situations, such as street noise. This was done to test the model's ability to handle different types of noise.

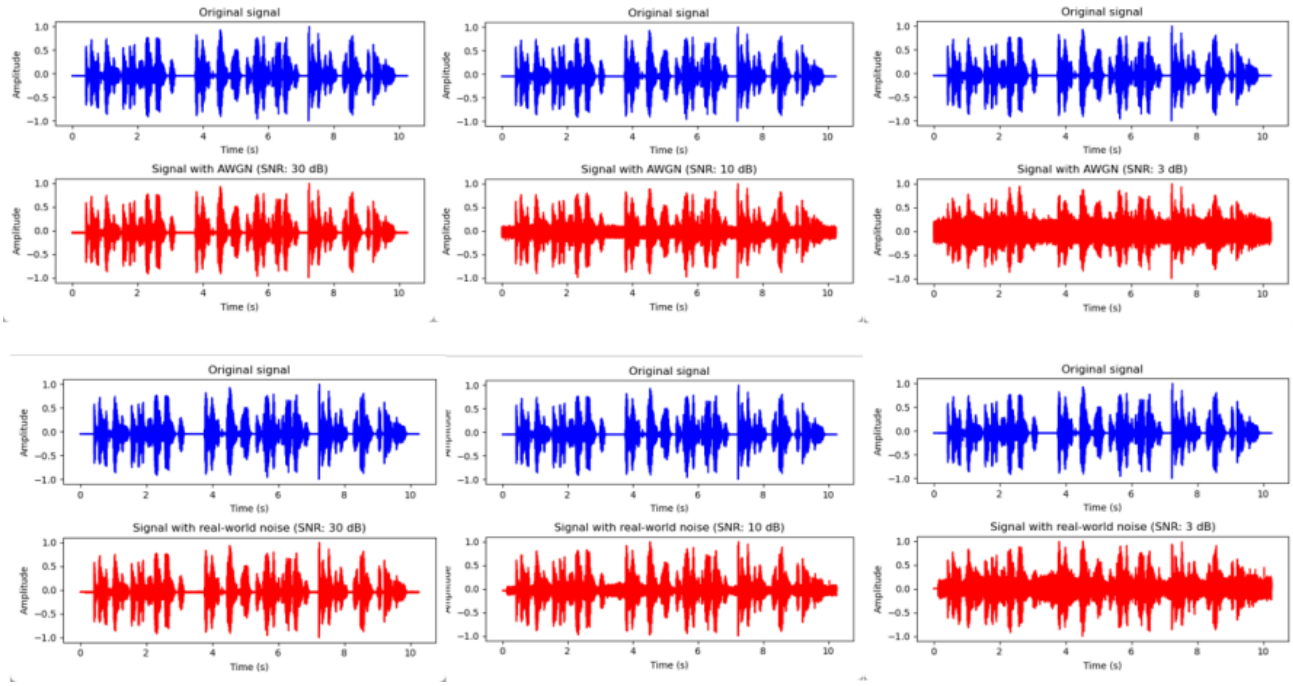

Fig. 1 Original audio and processed audio example from 6 experimental groups.

In total, there were six experimental groups and one control group in the study, corresponding to the six combinations of noise type and level, and the noise-free original audio. The audio data was preprocessed before being used for training the speech synthesis model, but no further details of the preprocessing steps are necessary for this study. Fig. 1 provides the original audio and processed audio example from 6 experimental groups.

2.2. Neural network model

The speech synthesis model studied in this paper is the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) model, which consists of a posterior encoder, a prior encoder, a decoder, a discriminator, and a random event predictor. A concise diagram outlining the entire architecture of the model is presented in Fig. 2.

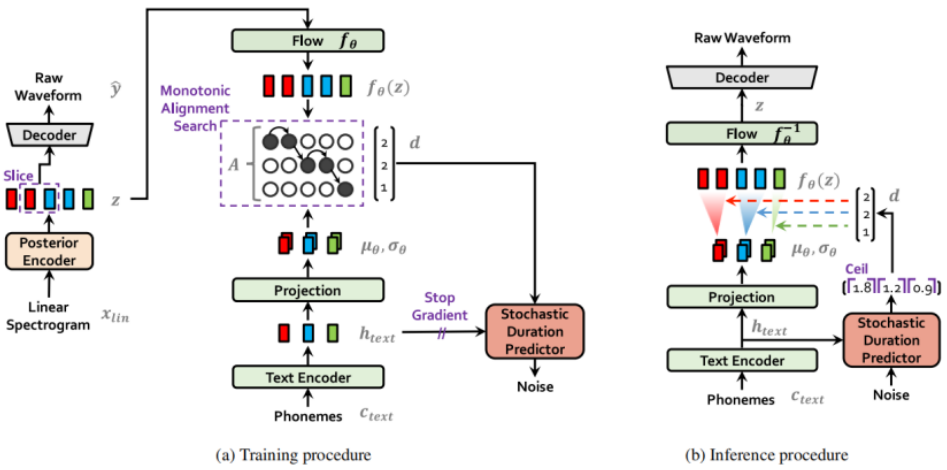

Fig. 2 The architecture of VITS [5].

During the training process, the model utilized a loss function that consists of several components. The reconstruction loss, Lrecon, is calculated as the L1-norm difference between the predicted mel-spectrogram, x_mel, and the ground-truth mel-spectrogram, Xmel, using the Variational Inference approach.

$$L_{recon} = \| x_{mel} - \hat{x}_{mel} \| \tag{1}$$

The KL-divergence loss, Lkl, is calculated based on the prior encoder, posterior encoder, and the latent variable, z, as follows:

$$L_{\text{kl}} = \log q_{\phi}(z \mid x_{lin}) - \log p_{\phi}(z \mid c_{text}, A), z \sim \text{q}_{\phi}(z \mid x_{lin}) = N(z; \mu_{\phi}(x_{lin}), \sigma_{\phi}(x_{lin})) \quad (2)$$

The duration prediction loss, Ldur, is obtained from text input, and the adversarial training component includes the generator loss, Ladv(G), and the feature matching loss, Lfm(G).

$$L_{\text{adv}}(G) = \text{E}_{z}[(D(G(z)) - 1)^2] \quad (3)$$

$$L_{\text{fm}}(G) = \text{E}_{(y,z)}[\sum_{l=1}^{T} \frac{1}{N_l} \| D^l(y) - D^l(G(z)) \|_1] \quad (4)$$

The final loss, LVAE, is the sum of all these components, representing the combined training of the VAE and GAN:

$$L_{vae} = L_{recon} + L_{kl} + L_{adv}(G) + L_{fm}(g) \quad (5)$$

2.3. Evaluation

The evaluation of the synthesized speech quality is crucial for assessing the performance of the VITS model under different noise conditions. In this study, the Neural network-based Index for Speech Quality Assessment (NISQA) was utilized to evaluate the synthesized speech quality. NISQA is a deep CNN self-attention model that predicts the multidimensional speech quality of synthesized speech. The inference of NISQA involves four stages: MEL-Spec segmentation, frame-wise modeling using CNN, time-dependent modeling using self-attention, and attention pooling using the pooling model. A comprehensive explanation of NISQA shown in Fig. 3 is available in [11].

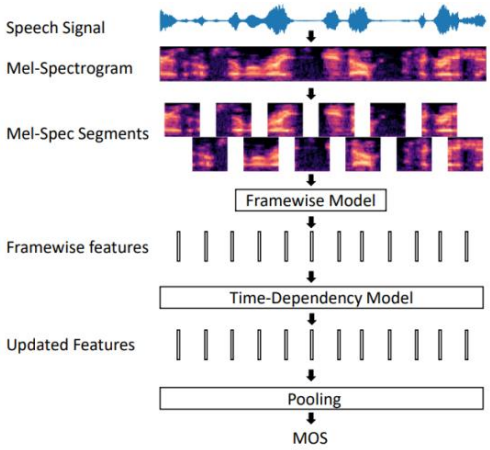

Fig. 3 General speech quality model structure [11].

NISQA is a more objective evaluation standard compared to human rating. It is capable of providing information on various aspects of synthesized speech, including overall quality, noisiness, coloration, discontinuity, loudness, and other aspects of the synthesized speech. In this study, NISQA was utilized as the primary evaluation criterion to compare the quality of synthesized speech across different noise conditions.

2.4. Implementation details

The VITS model was trained on a platform equipped with a NVIDIA 4090 GPU with 24G of memory. The Adam W optimizer was employed for training the network [12], and the configuration file for training included the following hyperparameters: batch size of 32, epochs of 1786, and a learning rate of 2e-4 with a decay rate of 0.999875. The training process utilized mixed-precision training (fp16_run). A segment size of 8192 was used during training, and the initialization learning

rate ratio was set to 1. Additionally, the c_mel and c_kl was set to 45 and 1.0, respectively. The NISQA model can be easily deployed locally and evaluated using a pre-provided model.

3. Results and Discussion

The experiment consisted of a control group and six experimental groups, as discussed earlier. The control group did not include any noise added, while the experimental groups were a combination of three different levels of Signal-to-Noise Ratio (SNR) and two types of noise, resulting in a total of six groups. The seven trained models were used to synthesize six different texts, and the resulting synthesized speech was evaluated using NISQA. The evaluation metrics consisted of Mean Opinion Score (MOS), Degree of Intelligibility Scale (DIS), and Coloration (COL), which measure overall quality, speech intelligibility, and speech naturalness, respectively. Given that the evaluation was conducted on six different texts, the resulting metrics were averaged to generate an overall score.

Table 1. Results of Control group

Evaluation indicators	Text_1	Text_2	Text_3	Text_4	Text_5	Text_6	Mean
MOS	3.24	3.10	3.18	3.02	2.97	3.33	3.14
DIS	4.26	4.11	4.19	3.98	3.94	4.36	4.14
COL	3.69	3.63	3.69	3.46	3.56	3.82	3.64

Table 2. Results of Group with Additive White Gausian Noise (AWGN)

SNR(dB)	Evaluation indicators	Text_1	Text_2	Text_3	Text_4	Text_5	Text_6	Mean
30	MOS	2.870	2.98	2.97	2.62	2.88	2.79	2.85
	DIS	4.34	4.28	4.27	4.28	4.24	4.36	4.29
	COL	3.76	3.79	3.70	3.75	3.59	3.72	3.72
10	MOS	1.26	1.47	1.18	1.14	1.44	1.40	1.32
	DIS	3.06	3.44	2.87	3.06	3.31	3.65	3.23
	COL	2.25	2.83	2.29	2.76	2.44	3.22	2.63
3	MOS	1.00	1.00	1.045	1.01	1.01	1.02	1.01
	DIS	2.73	2.74	2.71	2.75	2.73	2.68	2.72
	COL	1.26	1.26	1.30	1.29	1.26	1.32	1.28

Table 3. Results of Group with Real world noise

SNR(dB)	Evaluation indicators	Text_1	Text_2	Text_3	Text_4	Text_5	Text_6	Mean
30	MOS	3.16	3.22	2.95	3.05	2.98	2.95	3.05
	DIS	4.25	4.17	4.04	4.19	4.10	4.28	4.17
	COL	3.72	3.77	3.63	3.72	3.66	3.75	3.71
10	MOS	2.04	1.93	1.43	1.75	2.05	1.55	1.79
	DIS	4.31	4.31	4.29	4.26	4.35	4.35	4.31
	COL	3.89	3.90	3.58	3.86	3.81	3.73	3.80
3	MOS	1.42	1.45	1.41	1.11	1.43	1.35	1.36
	DIS	4.14	4.08	4.00	3.80	3.97	3.86	3.98
	COL	3.52	3.43	3.38	2.93	3.44	3.30	3.33

Table 1 shows the results of the control group, which had an average MOS score of 3.14. Table 2 and Table 3 show the results of the AWGN and real-world noise groups, respectively. Both types of noise had a significant impact on the MOS score, with the MOS score decreasing as SNR decreased.

For example, in the AWGN group, the MOS score decreased from 2.85 at SNR 30 dB to 1.01 at SNR 3 dB, which corresponds to a 64.5% decrease. Similarly, in the real-world noise group, the MOS score decreased from 3.05 at SNR 30 dB to 1.36 at SNR 3 dB, which corresponds to a 55.4% decrease.

Furthermore, the results showed that the DIS and COL scores of the real-world noise group were generally better than those of the AWGN group. For example, at SNR 3 dB, the DIS and COL scores of the real-world noise group were 3.98 and 3.33, respectively, while those of the AWGN group were 2.72 and 1.28, respectively. This may be due to the fact that real-world noise is more complex and diverse than AWGN, and the model may have learned to handle this type of noise better during training.

Notably, the MOS score remains relatively stable when SNR=30, indicating that the proposed model is less sensitive to high levels of Gaussian white noise. However, as the SNR level decreases, the impact on the MOS score becomes more pronounced. The DIS and COL metrics also showed a similar trend, indicating a decrease in speech intelligibility and naturalness with decreasing SNR.

Overall, the results of this experiment demonstrate that both Gaussian white noise and real-world noise can cause a significant decrease in speech synthesis quality, with the MOS scores showing a more significant decrease as the SNR decreases. Interestingly, the proposed model was still able to produce intelligible and natural-sounding speech even at an SNR of 3, which is an encouraging result. However, there is still significant room for improvement, as the MOS scores decreased by over 60% for an SNR of 10, and by over 67% for an SNR of 3. Further experimentation with different types of noise and varying SNRs could provide additional insights into the performance of the proposed model. Additionally, investigating alternative methods for handling noise in speech synthesis, such as denoising or noise robust feature extraction, could also be an interesting avenue for future research.

4. Conclusion

In conclusion, this study aimed to evaluate the performance of the VITS model in the presence of varying levels of noise interference. This study has successfully achieved its goal by conducting experiments to evaluate the VITS model's performance under various noise conditions. The findings provide valuable insights into the challenges and limitations of the model in practical applications, particularly in noisy environments. The experiments included a control group without noise and six experimental groups with different noise types and levels. The evaluation criteria consisted of MOS, DIS, and COL scores. The results demonstrated that both Gaussian white noise and real-world noise had a significant negative impact on the MOS scores, with lower signal-to-noise ratios leading to more severe degradation. Overall, the experimental results suggest that the VITS model is sensitive to noise levels and type, which is consistent with previous studies in the field. The superiority of the VITS model over the baseline in dealing with real-world noise was also observed. In the future, further research can be conducted to improve the VITS model's performance in noisy environments, such as exploring new approaches to noise reduction and optimizing the model's architecture. In terms of limitations, this study only investigated the impact of two types of noise on speech synthesis quality, and only one speech synthesis model was used. Further studies can explore the impact of other types of noise and compare the performance of different speech synthesis models under noisy conditions. Additionally, the current study only focused on objective evaluation metrics, and future studies can incorporate subjective evaluation to further understand the impact of noise on the subjective perception of synthesized speech quality.

References

[1] Norbert S et al. Synthesizing a choir in real-time using pitch synchronous overlap add (PSOLA), ICMC, 2000.

[2] Yoshimura T, Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis, in Proc. Sixth European Conference on Speech Communication and Technology, 1999.

[3] Ning Y et al. A Review of Deep Learning Based Speech Synthesis. Appl. Sci. 2019, 9, 4050. https://doi.org/10.3390/app9194050

[4] Kumar Y et al. A deep learning approaches in text-to-speech system: a systematic review and recent research perspective. Multimed Tools Appl, 2022. https://doi.org/10.1007/s11042-022-13943-4

[5] Jaehyeon K and Jungil K and Juhee S. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. arXiv. 2021.

[6] Fahima K et al. Text to Speech Synthesis: A Systematic Review, Deep Learning Based Architecture and Future Research Direction, Journal of Advances in Information Technology, Vol. 13, No. 5, pp. 398-412, October 2022.

[7] Amjady N. Short-term hourly load forecasting using time series modeling with peak load estimation capability. IEEE Transactions on Power Systems, 2001, 16(4): 798-805.

[8] Falas T and Stafylopatis A G, Neural networks in text-to-speech systems for the greek language, In Proc. 10th Mediter-ranean Electrotechnical Conference. Information Technology and Electrotechnology for the Mediterranean Countries, 2000, pp. 574- 577.

[9] Tachibana H, Uenoyama K, and Aihara S, Efficiently trainable textto-speech system based on deep convolutional networks with guided attention, in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, April 2018, pp. 4784- 4788.

[10] Lee J Y et al. Acoustic modeling using adversarially trained variational recurrent neural network for speech synthesis, in Proc. INTERSPEECH, 2018, pp. 917-921.

[11] Gabriel Mi et al. NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets aug Interspeech, 2021. 10.21437/interspeech.2021-299

[12] Loshchilov I and Hutter F Decoupled Weight Decay Regularization. In International Conference on Learning Representations, 2019, URL https://openreview. net/forum?id=Bkg6RiCqY7.