LoGAN: Attention-based GAN Vocoder using Longformer

Anonymous submission to Interspeech 2025

Supplementary Material

A0. Pre-Processing and Post-Processing (Volume normalisation)

As shown in Algorithm 1 and 2, it refers to the input folder containing generated files, O=original folder containing original files, and N=output folder to store normalized files. x_{gen}, sr_{gen} refers to generated audio data and its sampling rate, x_{orig}, sr_{orig} original audio data and its sampling rate. μ_{orig} mean volume of the original audio. μ_{gen} Mean volume of generated audio, n and x_{norm} are normalization factors and normalized audio signal, respectively.

1. **In-distribution (ID) data scenario:**

Algorithm 1 Volume Normalization Algorithm

Input: I, O, N
Output: Normalized files in N
for $f \in I$ **do**
    $x_{gen}, sr_{gen} \leftarrow \text{Load}(f, I)$
    $x_{orig}, sr_{orig} \leftarrow \text{Load}(f, O)$
    $\mu_{orig} \leftarrow \frac{1}{N}\sum_{i=1}^{N} |x_{orig}(i)|$
    $\mu_{gen} \leftarrow \frac{1}{N}\sum_{i=1}^{N} |x_{gen}(i)|$
    $n \leftarrow \frac{\mu_{orig}}{\mu_{gen}}$
    $x_{norm}(i) \leftarrow x_{gen}(i) \times n, \forall i$
    Save(x_{norm}, N)
end for

2. **Out-of-distribution Data (OOD) data scenario:**

Algorithm 2 OOD Volume Normalization Algorithm

Input: $I, N, threshold$
Output: Normalized files in N
for $f \in I$ **do**
    $x_{gen}, sr_{gen} \leftarrow \text{Load}(f, I)$
    $\mu_{gen} \leftarrow \frac{1}{N}\sum_{i=1}^{N} |x_{gen}(i)|$
    if $\mu_{gen} > threshold$ **then**
        $n \leftarrow \frac{threshold}{\mu_{gen}}$
        $x_{norm}(i) \leftarrow x_{gen}(i) \times n, \forall i$
        Save(x_{norm}, N)
    end if
end for

A1. Hyperparameters and Training Setup

Parameter	Value
fmin	0
fmax	8000 Hz
Sampling rate	22050 Hz
Number of sub-bands	80
Number of FFT	1024
Global Attention Tokens	N (input sequence)
Local Attention Window Size	$\{T_i, T_{i+1}, T_{i+2}, T_{i+3}\}$ from N
Learning Rate	0.001
Batch Size	8
Optimizer	Adam
Upsampling configuration	[8,8,2,2]
Kernel size	[16,16,4,4]
Training iterations	1.5×10^5

A2. Evaluation Matrices

Objective Matrices

1. **Perceptual Evaluation of Speech Quality (PESQ (↑))**
It predicts the perceived quality of speech based on a perceptual model. It is computed as a weighted sum of disturbance metrics:

$$\text{PESQ} = \alpha \cdot D_{\text{sym}} + \beta \cdot D_{\text{asym}} + \gamma \tag{1}$$

where D_{sym} and D_{asym} are symmetric and asymmetric disturbances, and α, β, and γ are empirically determined constants.

2. **Short-Time Objective Intelligibility (STOI (↑))**
It measures the intelligibility of degraded speech with respect to clean speech by computing correlations between short-time envelope segments:

$$\text{STOI} = \frac{1}{N}\sum_{n=1}^{N} \text{corr}\left(\mathbf{x}_n, \hat{\mathbf{x}}_n\right) \tag{2}$$

where $\mathbf{x}_n$ and $\hat{\mathbf{x}}_n$ are short-time temporal envelopes of the clean and degraded speech, and corr$(\cdot)$ denotes the Pearson correlation coefficient.

3. **Modulation Spectra Distance (MSD (↓)):** It calculates the likeness or disparity between two signals through modulation spectra. As given in Eq.(3),

$$\text{MSD} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(s(y)_i^t - s(y)_i^{\hat{t}}\right)^2}. \tag{3}$$

4. **Mel Cepstral Distortion (MCD (↓)):** It is used to quantify

the variation among two sets of Mel cepstral coefficients, i.e.,

$$MCD = \frac{\sqrt{\sum_{t=1}^{N}\left(\sqrt{\sum_{i=1}^{D}(A(t,i)-B(t,i))^2}\right)^2}}{N}. \quad (4)$$

Subjective Matrices

5. **Subjective Mean Opinion Score (SMOS (↑))** It is a subjective metric obtained from human listeners who rate the naturalness, or speech quality on a scale, from 1 to 5:

$$\text{SMOS} = \frac{1}{N}\sum_{i=1}^{N}\text{Rating}_i \quad (5)$$

where Rating_i is the score given by the i-th listener, and N is the number of raters or utterances evaluated.

A3. Recipes of the baseline systems

	Application	**Dataset Used**	**Opensource**
[1]	Speech synthesis (TTS)	Google internal TTS dataset (English and Mandarin, high-quality 48 kHz recordings)	No
	Unconditional speech generation	TIMIT dataset (clean phonetic speech corpus) [2]	Yes
	Music generation	Custom internal piano music corpus)	No
[3]	Vocoder for TTS	LJSpeech [4], VCTK [5], JSUT [6]	Yes
	Fine-tuned audio synthesis	Custom internal corpora	No
	[7] Vocoder TTS	LJSpeech [4], VCTK [5], LibriTTS [8]	Yes
	[Proposed] Vocoder for TTS	LJSpeech [4], VCTK [5]	-

Table 1: *Datasets used for training baseline systems across different vocoder models.*

A4. Additional Statistical Analysis

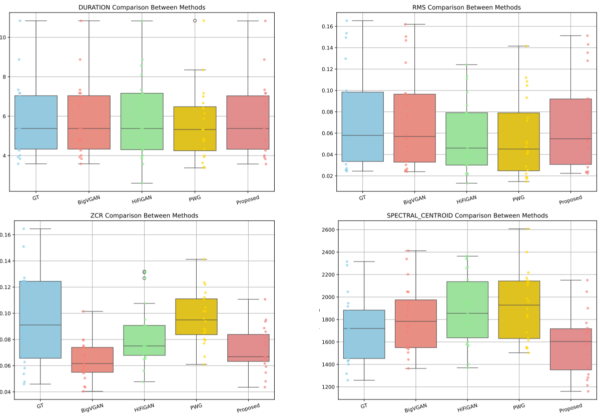

Figure 1: *Distributions and characteristics Analysis of proposed LoGAN with existing baseline systems, such as BigVGAN [9], HiFiGAN [7], and Parallel WaveGAN [1] across four metrics: Duration, RMS, ZCR, and Spectral Centroid, using box plots.*

1. References

[1] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 3918–3926.

[2] J. S. Garofolo, L. F. Lamel, W. M. Fisher, D. S. Pallett, N. L. Dahlgren, V. Zue, and J. G. Fiscus, "Timit acoustic-phonetic continuous speech corpus," *(No Title)*, 1993.

[3] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 6199–6203*, 2020, Barcelona, Spain.

[4] K. Ito and L. Johnson, "The LJ Speech Dataset," https://keithito.com/LJ-Speech-Dataset/ {Last Accessed: August 18^{th}, 2024}.

[5] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019 {Last Accessed: August 18^{th}, 2024}.

[6] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[7] 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 (NIPS), Vol. 33, pp. 17022–17033*, 2020, Virtual-only Conference.

[8] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," *arXiv preprint arXiv:1904.02882*, 2019.

[9] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022 {Last Accessed: August 18^{th}, 2024}.