BemaGANv2: A Vocoder with Superior Periodicity Capture for Long-Term Audio Generation

Taesoo Park
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
taesoo0707@kw.ac.kr

Mungwi Jeong
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
ansrnl9178@kw.ac.kr

Mingyu Park
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
a6000692@kw.ac.kr

Narae Kim
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
wing02@kw.ac.kr

Junyoung Kim
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
kimjnyoung@kw.ac.kr

Dohyun Park
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
pdh6608@gmail.com

Mujung Kim
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
kmj1026@kw.ac.kr

Jisang Yoo
Department of Electronic Engineering
Kwangwoon University
Seoul, South Korea
jsyoo@kw.ac.kr

Hoyun Lee
Department of Otorhinolaryngology-Head and Neck Surgery
Ewha Womans University College of Medicine
Seoul, South Korea
hoyun1004@gmail.com

Sanghoon Kim
Department of Otorhinolaryngology, Head and Neck Surgery
School of Medicine, Kyung Hee University
Seoul, South Korea
hoon0700@naver.com

*Soonchul Kwon
Graduate School of Smart Convergence
Kwangwoon University
Seoul, South Korea
ksc0226@kw.ac.kr

***Abstract*—This paper introduces BemaGANv2, an advanced model built upon the BemaGAN architecture. BemaGANv2 enhances the Generator and extends the discriminator framework to more effectively capture long-term audio dependencies and periodicity. By applying the Snake function as the activation function in the Generator, the model improves its ability to handle audio extrapolation and periodicity. Additionally, the envelope extraction method in the MED has been refined, and the combination of the discriminator in the existing BemaGAN has been changed from the MED+MPD to the MED+MRD structure. Experiments with various discriminator combinations, including MSD+MED, MSD+MRD, and MPD+MED+MRD, validate the effectiveness of BemaGANv2. The final BemaGANv2 model serves as a vocoder in Text-to-Audio (TTA) or Text-to-Music (TTM) tasks to restore the original audio, aiming to enhance the fidelity and perceptual quality of generated long-term audio. Experimental results demonstrate that BemaGANv2 outperforms the previous model in both objective and subjective evaluation metrics, making it more suitable for long-term audio generation.**

***Index Terms*—Artificial Intelligence (AI), Bespoke envelope multi-discriminator adaptive GAN (BemaGAN), Multi-Period Discriminator (MPD), Multi-Scale Discriminator (MSD), Multi-Envelope Discriminator (MED), Multi-spectrogram Resolution Discriminator (MRD)**

I. Introduction

Recently, deep learning-based Text-to-Audio (TTA) and Text-to-Music (TTM) models have garnered significant attention in the field of AI music generation. A critical component in generating audio signals in Diffusion-based TTA and TTM models is the vocoder, which is responsible for converting intermediate representations, such as Mel-Spectrograms, into actual audio signals.

In this study, we develop and enhance a unique vocoder model named BemaGAN to optimize the performance of TTM models. The original BemaGAN, inspired by the HiFi-GAN structure, replaces the Multi-Scale Discriminator (MSD) with a Multi-Scale Envelope Discriminator (MED). The MED is designed to more sensitively detect the periodicity in audio, thereby improving both the fidelity and perceptual quality of long-term generated audio [1].

Although the initial version of BemaGAN demonstrated superior performance over HiFi-GAN, indicating its potential as a vocoder for TTA and TTM models, our study goes further by introducing BemaGANv2 to further enhance its performance. BemaGANv2 incorporates a new activation function, the snake function—previously utilized in BigVGAN—to improve the generator's ability to extrapolate periodicity [2]. Additionally, the frequency range has been extended from 8 kHz to 12 kHz, with the sampling rate adjusted to 24 kHz, in accordance with the Nyquist sampling theorem. The model's performance is further improved by increasing the number of envelope features extracted from the original MED structure and by replacing the Multi-Period Discriminator (MPD) with a Multi-Resolution Discriminator (MRD). These enhancements lead to improved performance across various objective and subjective metrics.

In this paper, we compare the performance of BemaGANv2 with existing models and different discriminator combinations, validating the efficacy of the MED and MRD combination through various evaluation metrics and experiments. We discuss the advantages of BemaGANv2 as a vocoder for TTA and TTM models and its broader applicability. The LJ Speech dataset was used for model training, with tests conducted using data randomly downloaded from freesound.org. All training and experiments were performed in a CUDA environment using the A100 GPU provided by Colab.

Identify applicable funding agency here. If none, delete this.

II. Models

A. Generator with Snake Function

The existing BemaGAN Generator adopts the structure of HiFi-GAN. In BemaGANv2, the snake activation function, originally used in BigVGAN, is introduced as a key differentiating factor. The equation for the snake function is given as follows:

$$f_\alpha(x) = x + \frac{1}{\alpha}\sin^2(\alpha x) \tag{1}$$

Here, α is a learnable parameter that controls the periodic components and frequency of the signal. This sine-based function introduces an inductive bias that enables the model to better learn and generate periodic characteristics in audio [3]. The snake function is specifically designed to enhance the model's extrapolation ability and capture periodicity more accurately. By integrating this function into the Generator, BemaGANv2 is able to generate more refined audio signals, effectively learning and reproducing subtle acoustic features that are often challenging for traditional activation functions to capture.

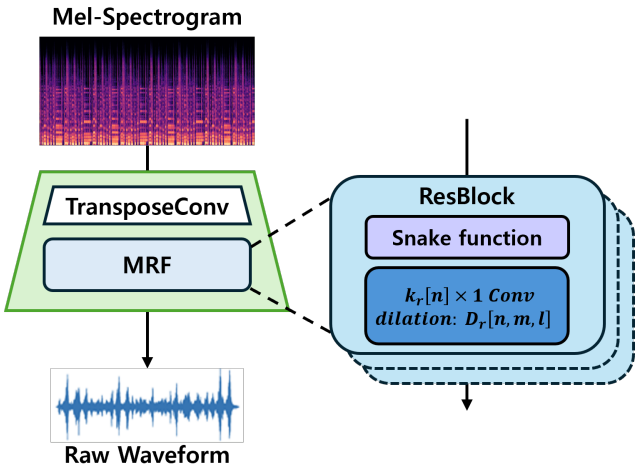

Fig. 1. Schematic diagram of BemaGANv2 generator. A structure that applies the snake function as an activation function to the multi-receptive field fusion (MRF) module in the generator structure of HiFi-GAN. [4]

B. Multi-Envelope Discriminator, MED

The MED is designed to extract various envelope features from audio signals and consists of five envelope extractors and a 1-Dimensional Convolutional Neural Network (CNN) [1]. During the preprocessing stage, the MED extracts the envelopes of the audio data. In the original BemaGAN, only the upper and lower envelopes were extracted. However, in BemaGANv2, additional low-pass filters at 300Hz and 500Hz are applied to capture the upper and lower envelope features at different cutoff frequencies.

These extracted envelopes are then processed through seven 1-D CNN layers, where various features of the audio signal are further extracted. These features are used to update the loss function, guiding the model to minimize the loss during training. Ultimately, these features and the refined loss function are used to evaluate the similarity between the original and restored audio.

C. BemaGANv2 Structure

Fig. 2 illustrates the overall structure of BemaGANv2. Based on the experimental results of this study, BemaGANv2 incorporates the MED and MRD discriminator structures. These two discriminators, along with the generator, are updated and trained in a complementary manner to minimize the loss.

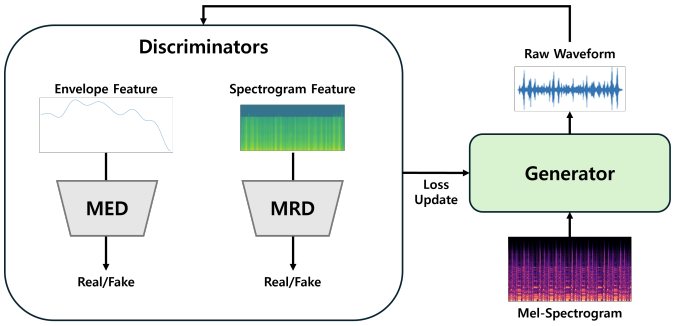

Fig. 2. Overall structure of BemaGANv2

III. Experiment with Combinations of Discriminators

In this section, we compare various combinations of discriminators using multiple evaluation metrics and visual analyses to verify the validity of the discriminator combination in BemaGANv2. We experimented with combinations from HiFi-GAN, BigVGAN, BemaGAN, and BemaGANv2, as well as specific combinations like MSD+MED and MSD+MRD. Additionally, we explored combinations involving three or more discriminators, presenting experimental results for the MED+MPD+MRD combination. Other combinations are omitted for reasons discussed later. Notably, all experimental models, except HiFi-GAN, utilized the snake function.

For this study, we employed four objective evaluation metrics and two subjective evaluation metrics.

A. Objective Evaluation Metrics

The objective evaluation metrics used in this study measure various distances between the ground truth and the restored audio. In the formulas, G and R represent Ground Truth and Restored audio, respectively. The metrics and their descriptions are as follows:

- FAD (Fréchet Audio Distance) - This metric is used to evaluate the quality of audio signals by measuring the difference in the distribution between two audio signals. [5]
- SSIM (Structural Similarity Index Measure) - In this paper, this metric evaluates the structural similarity of

the Mel-Spectrogram image between the Ground Truth and the restored audio. The SSIM is calculated for the three RGB channels, and the average value is taken. [6]

- PLCC (Pearson Linear Correlation Coefficient) - This statistical index measures the linear correlation between two variables. In this paper, it indicates how well the restoration preserves the frequency domain characteristics. [7]
- MCD (Mel Cepstral Distortion) - This metric evaluates the distortion between two audio signals by calculating the difference in Mel-Frequency Cepstral Coefficients (MFCC) using dynamic time warping (DTW). [8]

$$\text{FAD} = \|\mu_G - \mu_R\|^2 + \text{Tr}\left(\Sigma_G + \Sigma_R - 2\sqrt{\Sigma_G \Sigma_R}\right) \quad (2)$$

$$\text{SSIM}_{\text{total}} = \frac{1}{3}\sum_{i=1}^{3} \text{SSIM}\left(\text{MelSpec}_G^{(i)}, \text{MelSpec}_R^{(i)}\right) \quad (3)$$

$$\text{PLCC}(G, R) = \frac{\text{Cov}(G, R)}{\sigma_G \sigma_R} \quad (4)$$

$$\text{MCD} = \frac{10 \cdot \sqrt{2}}{\ln(10)} \cdot \frac{1}{N} \sum_{(G,R)\in \text{path}} \sqrt{\sum_{k=1}^{K} \left(c_G^{(k)} - \hat{c}_R^{(k)}\right)^2} \quad (5)$$

B. Subjective Evaluation Metrics

For subjective evaluation, we conducted a Mean Opinion Score (MOS) evaluation for each discriminator combination and a Similarity MOS (SMOS) evaluation targeting prominent vocoder models such as HiFi-GAN, BigVGAN, BemaGAN, and BemaGANv2. The MOS evaluation involved randomly selected men and women in their 20s, while the SMOS evaluation was conducted with participants knowledgeable in audio signal processing or music. Both evaluations were carried out using a 5-point scale, with results presented alongside a 95 % confidence interval (CI).

In the MOS evaluation, participants listened to samples generated by the models without being provided with the Ground Truth, and rated the overall quality. In the SMOS evaluation, participants first listened to the Ground Truth and then rated the similarity of the generated sample to this reference. Both short and long audio samples were evaluated separately, and a variety of samples were provided to calculate an average score. To ensure fairness, the specific model from which each sample was generated was not disclosed to the participants.

C. Result

For short-term audio, HiFi-GAN still outperforms other models. This is because Leaky ReLU, used in HiFi-GAN, is a more suitable activation function than the Snake function for short-term audio, where periodicity is less prominent.

In long-term audio, BemaGANv2 outperforms other models in all objective metrics except SSIM. The lower SSIM score is likely due to BemaGANv2's inability to perfectly reproduce the loudness of the original audio. However, in the other key metrics—FAD, PLCC, and MCD—BemaGANv2 effectively reduces distortion in audio structure and frequency characteristics, strongly supporting its suitability as a vocoder for TTA or TTM models.

TABLE I
SHORT TERM AUDIO OBJECTIVE METRICS

Models	FAD ↓	SSIM ↑	PLCC ∼ 1	MCD ↓
HiFi-GAN	9.86	**0.81**	**0.995**	**0.52**
BigVGAN	**7.84**	0.79	0.989	0.92
BemaGAN	8.21	0.78	0.983	0.96
BemaGANv2	12.88	0.78	0.994	0.62
MSD + MED	15.00	0.68	0.985	0.92
MSD + MRD	13.67	0.65	0.975	0.99
MED + MPD + MRD	14.84	0.72	0.983	0.82

TABLE II
LONG TERM AUDIO OBJECTIVE METRICS

Models	FAD ↓	SSIM ↑	PLCC ∼ 1	MCD ↓
HiFi-GAN	15.76	**0.50**	0.753	0.98
BigVGAN	7.44	0.43	0.985	0.72
BemaGAN	9.33	0.45	0.984	0.67
BemaGANv2	**4.26**	0.37	**0.99**	**0.50**
MSD + MED	7.82	0.36	0.988	0.77
MSD + MRD	10.92	0.38	0.983	0.73
MED + MPD + MRD	7.23	0.38	0.988	0.74

The subjective evaluation metrics follow a similar trend to the objective metrics. For short-term audio, HiFi-GAN achieved the highest MOS, while BigVGAN achieved the best SMOS. For long-term audio, BemaGANv2 obtained the highest scores in both MOS and SMOS.

The MED+MPD+MRD combination demonstrated strong performance in objective metrics for long-term audio, securing the second-best scores in FAD and PLCC. However, it received the second-lowest MOS score, likely due to overfitting despite the experiments being conducted with the same number of epochs. Additionally, it is important to note the risk of mode collapse inherent in GAN models when using multiple discriminators [10] [11]. For this reason, further experiments involving combinations of three or four different discriminators were not pursued.

TABLE III
MEAN OPINION SCORE

Models	Short Audio MOS ↑	Long Audio MOS ↑
HiFi-GAN	**2.64**(± 0.09)	1.36(± 0.14)
BigVGAN	1.96(± 0.08)	3.11(± 0.14)
BemaGAN	1.79(± 0.10)	3.07(± 0.13)
BemaGANv2	2.14(± 0.09)	**3.38**(± 0.11)
MSD + MED	2.14(± 0.09)	2.43(± 0.09)
MSD + MRD	2.14(± 0.10)	2.74(± 0.09)
MED + MPD + MRD	2.18(± 0.07)	2.21(± 0.08)

As illustrated by the Mel-spectrogram in Figure 3, BemaGANv2 shows a significant improvement in reconstructing audio signals, especially in long-term audio. The visual analysis reveals that BemaGANv2 effectively captures long-term

TABLE IV
SIMILARITY MEAN OPINION SCORE

Models	Short Audio SMOS ↑	Long Audio SMOS ↑
HiFi-GAN	2.7(± 0.09)	1.15(± 0.08)
BigVGAN	**2.95**(± 0.10)	3.3(± 0.11)
BemaGAN	2.48(± 0.10)	3.28(± 0.13)
BemaGANv2	2.53(± 0.09)	**3.58**(± 0.09)

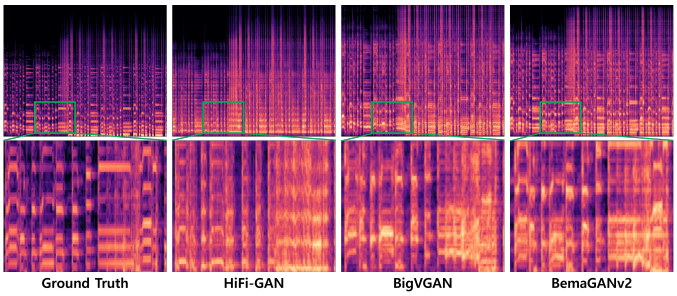

Fig. 3. Mel-Spectrogram visualization of samples form Ground Truth, Hifi-GAN, BigVGAN, and BemaGANv2 models trained on LibriTTS

characteristics of the audio, preserving intricate patterns and structures essential for high-fidelity audio generation. Notably, BemaGANv2 exhibits reduced noise artifacts compared to HiFi-GAN and BigVGAN, indicating a more refined and accurate synthesis process. This reduction in noise not only enhances the perceptual quality of the generated audio but also suggests that the model is better at distinguishing and reproducing subtle audio details, further validating the architectural enhancements implemented in BemaGANv2.

For details on the loss functions and learning rates used in training, please refer to Appendix A.

IV. CONCLUSION AND FUTURE WORK

A. Conclusion

In this study, BemaGANv2 has been demonstrated to be a highly suitable vocoder for wide frequency bands and long-term audio. By introducing the snake activation function, BemaGANv2 enhances its ability to capture periodicity and improve extrapolation, surpassing the performance of the original BemaGAN model, particularly in restoring long-term audio. In objective evaluation metrics, BemaGANv2 achieves excellent results in key measures such as FAD, PLCC, and MCD, significantly broadening its applicability as a vocoder in TTA and TTM models.

Furthermore, this study explored various discriminator combinations to identify the optimal configuration. The resulting design of BemaGANv2, informed by these experiments, demonstrates competitive performance in both objective and subjective evaluation metrics, showcasing its effectiveness in audio generation with a focus on real-world user experience.

B. Future Work

Recently, many Diffusion-based audio generation AIs have used HiFi-GAN as a vocoder. Accordingly, we plan to develop a new generative AI that applies BemaGANv2 as a vocoder. Based on the excellent long-term audio processing performance and excellent restoration ability in a wide frequency band of BemaGANv2, this new AI model is expected to enable more natural and high-quality audio and music generation.

In future work, we will focus on integrating BemaGANv2 into a Latent Diffusion-based model, especially maximizing its performance in long-term audio and wide frequency band music generation. To this end, we will optimize the structure and parameters of BemaGANv2 with the Latent Diffusion model to achieve better learning efficiency and generation quality. In addition, we will explore the applicability of BemaGANv2 in various application fields and verify its performance in difficult environments such as real-time audio generation.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2024-00352526)

REFERENCES

[1] Mun-Gwi Jeong, Na-Rae Kim, Jun-Young Kim, Min-Gyu Park, Tae-Soo Park, and Ji-Sang Yoo, "BemaGAN: Generative Adversarial Networks with Multi Envelope Discriminator," in Summer Annual Conference of IEIE, pp. 1992-1995, July 2024
[2] 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, Jun 2022.
[3] L. Ziyin, T. Hartwig, and M. Ueda, "Neural networks fail to learn periodic functions and how to fix it," in Advances in Neural Information Processing Systems, vol. 33, pp. 1583–1594, Jun 2020.
[4] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in Advances in Neural Information Processing Systems, vol. 33, pp. 17022–17033, October 2020.
[5] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fre'chet audio distance: A metric for evaluating music enhancement algorithms," arXiv preprint arXiv:1812.08466, December 2018.
[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, April 2004.
[7] P. Sedgwick, "Pearson's correlation coefficient," BMJ, vol. 345, 2012.
[8] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in Proc. IEEE Pacific Rim Conf. Commun. Comput. Signal Process., vol. 1, pp. 125–128, May 1993.
[9] W. Jang, D. Lim, and J. Yoon, "Universal MelGAN: A robust neural vocoder for high-fidelity waveform generation in multiple domains," arXiv preprint arXiv:2011.09631, 2020.
[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems, vol. 27, Jun 2014.
[11] D. Berthelot, T. Schumm, L. Metz "BEGAN: Boundary Equilibrium Generative Adversarial Networks," arXiv preprint arXiv:1703.10717, May 2017.
[12] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in Proc. IEEE Int. Conf. Comput. Vis., pp. 2794–2802, 2017.

APPENDIX

A. Training Loss

The loss of BemaGANv2 is the same as that used in HiFi-GAN.

- **Gen Loss**: This model adopts the approach of LSGAN [12], where the binary cross-entropy loss function used in the original GAN objective is replaced with a least squares loss function. This substitution helps maintain

a non-vanishing gradient flow. The discriminator (D) is trained to classify real samples as 1 and synthetic samples generated by the generator (G) as 0. The generator (G) is then optimized to improve the quality of synthetic samples so that they are classified by the discriminator as close to 1 as possible. The GAN losses for the generator and discriminator are defined in Equations (6) and (7), where x represents the input condition (the real audio), and s is the Mel-Spectrogram of the real audio.

$$\mathcal{L}_{\text{Adv}}(D;G) = \mathbb{E}_{(x,s)}\left[(D(x)-1)^2 + (D(G(s)))^2\right] \quad (6)$$

$$\mathcal{L}_{\text{Adv}}(G;D) = \mathbb{E}_{s}\left[(D(G(s))-1)^2\right] \quad (7)$$

- **Mel-Spectrogram Loss**: In addition to the GAN objective, we incorporate a Mel-Spectrogram Loss to further enhance the training efficiency of the Generator (G) and improve the fidelity of the generated audio. The Mel-Spectrogram Loss is defined in (8) as the L1 distance between the Mel-spectrograms of the waveforms synthesized by the Generator (G) and the real waveforms. This loss function aids in synthesizing realistic waveforms that correspond to the input conditions and also helps stabilize the adversarial training process in its early stages. Here, Φ is a function that transforms the waveform into the corresponding mel spectrogram.

$$\mathcal{L}_{\text{Mel}}(G) = \mathbb{E}_{(x,s)}\left[\|\phi(x) - \phi(G(s))\|_1\right] \quad (8)$$

- **Feature Matching Loss**:This loss is a learned similarity metric measured by the difference between features in the Discriminator (D). It extracts each intermediate feature of the Discriminator between real and generated samples, and calculates the L1 distance between real samples and conditionally generated samples in each feature space. This loss is based on the similarity in the Discriminator's feature space and is used as an additional loss in training the Generator. It is defined in (9). Here, T represents the number of layers of the discriminator. D_i and N_i represent the number of features and features in the i-th layer of the discriminator, respectively.

$$\mathcal{L}_{\text{FM}}(G;D) = \mathbb{E}_{(x,s)}\left[\sum_{i=1}^{T}\frac{1}{N_i}\left\|D^i(x) - D^i(G(s))\right\|_1\right] \quad (9)$$

- **Final Loss**: The ultimate objectives of the Generator and Discriminator are defined in (10) and (10), where the coefficients are set as $\lambda_{fm} = 2$ and $\lambda_{mel} = 45$, following the configuration of HiFi-GAN. Since the Discriminator is composed of a set of sub-discriminators, MPD and MSD, Equations (10) and (11) can be expanded (12) and (13) for each sub-discriminator. Here, D_k represents the k-th sub-discriminator in both MPD and MSD.

$$\mathcal{L}_G = \mathcal{L}_{\text{Adv}}(G;D) + \lambda_{\text{fm}}\mathcal{L}_{\text{FM}}(G;D) + \lambda_{\text{mel}}\mathcal{L}_{\text{Mel}}(G) \quad (10)$$

$$\mathcal{L}_D = \mathcal{L}_{\text{Adv}}(D;G) \quad (11)$$

$$\mathcal{L}_G = \sum_{k=1}^{K}\left[\mathcal{L}_{\text{Adv}}(G;D_k) + \lambda_{\text{fm}}\mathcal{L}_{\text{FM}}(G;D_k)\right] + \lambda_{\text{mel}}\mathcal{L}_{\text{Mel}}(G) \quad (12)$$

$$\mathcal{L}_D = \sum_{k=1}^{K}\mathcal{L}_{\text{Adv}}(D_k;G) \quad (13)$$

B. Learning Rate

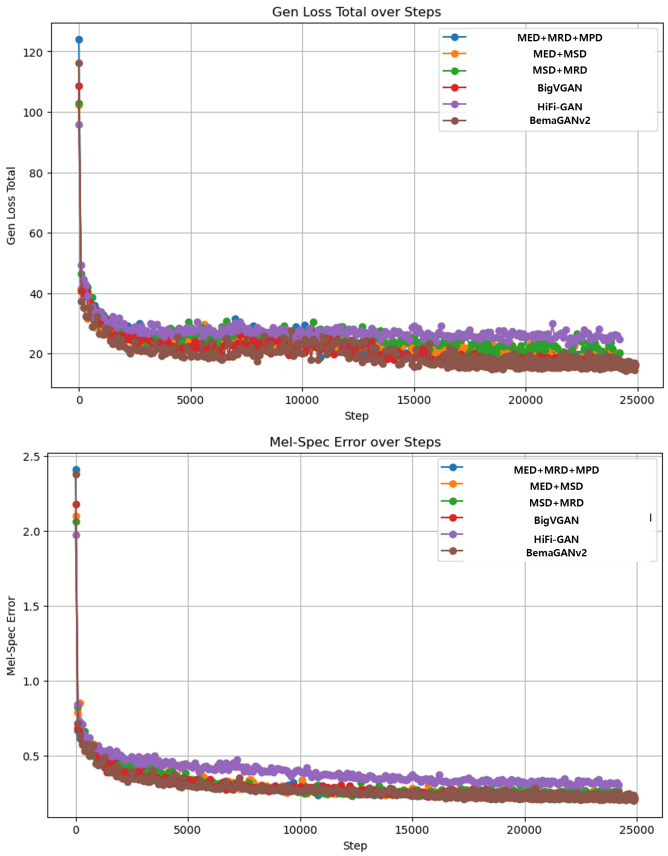

Fig. 4. Visualization of gradient norm for various combination of discriminators with Gen Loss Total and Mel-Spectrogram Error

BemaGANv2 stands out for its exceptional convergence efficiency, enabling the model to achieve high performance with significantly fewer training epochs compared to other models. While models like HiFi-GAN and BigVGAN require extensive training periods to reach optimal results, BemaGANv2's Generator completes the learning process in a reduced number of epochs. This efficiency not only decreases the computational resources and time required for training but also provides practical advantages in scenarios where rapid model deployment is crucial. These features make BemaGANv2 an ideal choice for applications that demand high-quality audio generation under tight training constraints.