Robust Waveform Generation via an Enhanced Location-Variable Convolution Network

****** ****[*]

School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong, 510006, China

[*]Corresponding author's e-mail: ********@mail2.sysu.edu.cn

Abstract. Audio synthesis faces significant challenges in generating high-quality waveforms while ensuring model efficiency. This paper proposes a novel learning network that involves two channels. One is the Location-variable convolution network (LVCNet) that generates time-domain features, the other channel uses a Fourier Transform as an input, processed through a convolutional neural network (CNN) to obtain frequency-domain features. The time-domain features generated by LVCNet and the frequency-domain features obtained via Fourier Transform are treated as two independent input channels that are concatenated after being processed by the CNN. Experimental evaluations indicate that the proposed network significantly improves waveform generation quality, stability, and inference speed when compared with competing methods. The results show that the combined use of time-domain and frequency-domain features enhances the performance of the synthesized waveforms, achieving superior outcomes in both subjective and objective assessments. These findings verify the effectiveness of integrating time and frequency domain characteristics for waveform synthesis, demonstrating the potential of this approach in advancing speech synthesis technologies.

Keywords: Speech synthesis, Waveform generation, Fourier transform, LVCNet, CNN

1. Introduction

Recent advances in neural network-based deep generative models have significantly improved speech synthesis, particularly in waveform generation. Traditional methods like concatenative speech synthesis (CSS) [1] and parametric models such as Hidden Markov Models (HMMs) [2] have been largely replaced by deep learning approaches. CSS stores multiple speech units for synthesis, while parametric models describe speech signals probabilistically but suffer from lower audio quality and higher complexity. Linear Predictive Coding (LPC) [3] models estimate future speech signal values for spectral features but struggle with nonlinear characteristics, leading to high-frequency distortions.

Waveform coding and synthesis models, including time-domain waveform synthesis [4, 5] and spectral synthesis, have shown high efficiency and naturalness. However, the introduction of statistical models and deep learning has further advanced speech synthesis. Neural network better capture temporal relationships in speech, resulting in more efficient, low-complexity, and high-quality audio generation. This highlights the superiority of deep learning techniques in modern speech synthesis systems.

Current waveform generation methods include parametric models [6][7], statistical models [8][9][10], GANs [11][12][13], autoregressive models [14][15][16], and flow-based models [17][18][19]. Each has unique

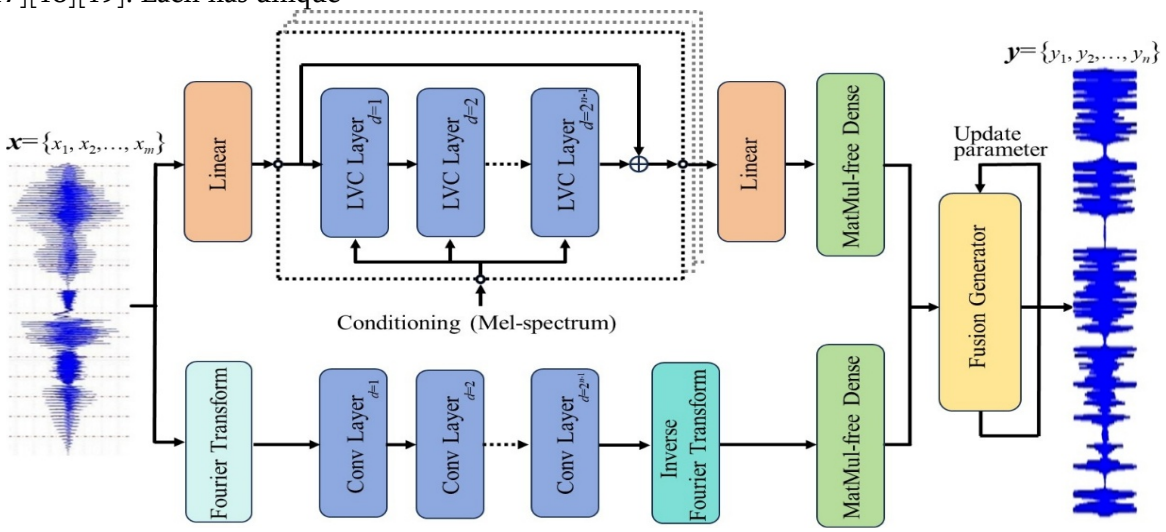


Figure 1. The proposed enhanced location-variable convolution network for waveform generation.

Strengths and weaknesses. Parametric models generate speech quickly but may lack naturalness and emotional expression. Statistical models like GMMs [20] and HMMs [21] effectively model temporal relationships but struggle with long-term dependencies. GANs produce high-quality, natural speech through adversarial training, though this can be unstable and vary in quality. Autoregressive models, such as WaveNet, capture short- and long-term dependencies but are slower. Flow models optimize reversible functions for diverse voice generation, though their training is complex.

Additionally, Variational Autoencoders (VAEs) [22][23] and end-to-end deep neural network models [24][25] generate high-quality speech under various conditions, despite challenges like low efficiency, high costs, and poor adaptability to conditional changes. These generative models collectively advance the field of speech synthesis, each contributing to improved audio quality and generation capabilities.

This paper proposes a novel waveform generation approach that builds upon the classical autoregressive model LVCNet architecture [26]. The main contributions are summarized as follows:

1) **Dual-Channel Architecture**. We propose a dual-channel framework that integrates time-domain features generated from the original LVCNet with frequency-domain features extracted via Fourier Transform, enabling simultaneous capture of both temporal and spectral information of audio signals, thereby improving waveform generation quality.

2) **Fourier Transform Layer**. By introducing a Fourier Transform process to handle frequency-domain features, our model leverages the frequency characteristics of audio signals while maintaining the temporal accuracy provided by the LVCNet model.
3) **Experimental Validation**. Experimental results demonstrate that our model outperforms existing methods in both subjective and objective evaluations, yielding higher-quality waveforms and faster inference speeds compared to autoregressive models.

2. The Waveform Generation Network

2.1 Problem Statement

Given an input waveform signal data $x=\{x_1, x_2,\ldots, x_m\}$, where $x\in\mathbb{R}$, m represents the dimensionality of the input features, this paper aims to explore an efficient network f to map this input x to an output sequence $f(x)=\{y_1, y_2,\ldots, y_n\}$, where n denotes the length of the time series. The challenge lies in that f need to effectively learn the intricate relationships within the input and generate high-quality waveform signals that closely approximate the ground-truth values. Specifically, the generated waveform signals should be as close as possible to the true, high-quality audio signals.

2.2 Overview of the waveform generation network

The framework is built upon an efficient dual-channel architecture based on MatMul-free technology, involving two key channels: one that integrates time-domain features generated by LVCNet with frequency-domain features extracted via Fourier Transform, and another that introduces a Fourier Convolution Network (FCN) to process frequency-domain characteristics (see Figure 1). The motivation is to simultaneously capture both temporal and spectral information of audio signals, as waveform generation tasks require both fine-grained temporal structures and long-range spectral dependencies. Specifically, the time-domain features are produced from LVCNet, which effectively models the temporal dynamics of speech signals, while the frequency-domain features are processed through the FCN, leveraging the frequency characteristics of audio signals to enhance the model's understanding and representation.

In different application scenarios, these two types of features have distinct emphases: in speech synthesis, time-domain features ensure the preservation of phase information and transient details, contributing to the smoothness of pronunciation, while frequency-domain features capture harmonic structures and formant distributions, enhancing vowel clarity. In music synthesis, the frequency-domain channel plays a crucial role in maintaining timbral consistency and spectral stability across different notes, whereas the time-domain features preserve note onsets, rhythmic accuracy, and expressive dynamics. In environmental sound generation (such as rain or wind noise), time-domain modeling captures stochastic characteristics, while frequency-domain features help maintain spectral textures. The interaction between the two enables the model to capture both random textures and structured spectral patterns.

The synergy between these feature representations forms a complementary enhancement mechanism where time-domain features excel at capturing local details while frequency-domain features provide global spectral structures. The model dynamically adjusts the contribution weights of both channels via attention mechanisms, adaptively emphasizing different representations based on audio characteristics—relying more on time-domain information for plosive sounds and frequency-domain channels for sustained vowels. This adaptive fusion strategy overcomes single-

domain modeling limitations, while information exchange between channels enables the model to learn intrinsic time-frequency relationship constraints. By merging these feature streams, the architecture improves expressiveness, fidelity, and robustness of waveform generation across various scenarios while maintaining computational efficiency through its MatMul-free design.

2.3 LVCNet-based channel

The LVCNet channel is built by stacking multiple Hierarchical Position-Variable Convolution (HPVC) layers. Each HPVC forms an LVCNet block, and multiple blocks make up the entire LVCNet architecture. Linear layers at the input and output facilitate channel transformation. In each position-variable convolution structure, adaptive kernel coefficients allow elements in the local conditioning sequence $\boldsymbol{h} = \{h_1, h_2, \ldots, h_m\}$ to associate with continuous intervals of the input $\boldsymbol{x} = \{x_1, x_2, \ldots, x_n\}$. Features are extracted from different input intervals using corresponding convolution kernels, and the output sequence is formed by concatenating these convolution results. In traditional linear predictive vocoders, the waveform generation process can be represented as:

$$y_t = \sum_{k=1}^{N} a_k x_{t-k} + b \,, \tag{1}$$

where y_t denotes the output at frame t, x_{t-k} represents the input from the previous N frames, a_k is the linear prediction coefficients, and b is the bias term. Given an input sequence $\boldsymbol{x}$ and a local conditioning sequence $\boldsymbol{h}$, the convolution process for each interval is defined as the following representation.

$$y(i) = \sigma\left(W_f(i) * x(i) + b_f(i)\right) \odot tanh\left(W_g(i) * x(i) + b_g(i)\right) , \tag{2}$$

where $x(i)$ denotes the input sequence interval associated with hi, $Wf(i)$ and $Wg(i)$ represent the filter and gating convolution kernels used for $x(i)$, respectively. Here, $\sigma(\cdot)$ signifies the Sigmoid activation function, $\odot$denotes the element-wise multiplication, and $tanh(\cdot)$ is the hyperbolic tangent activation function. The kernel predictor consists of multiple residual linear layers equipped with Leaky ReLU activation functions. It predicts multiple sets of convolution kernels based on the local conditioning sequence $\boldsymbol{h}$, where each set is employed for convolving different intervals of the input sequence. By generating convolution kernels tailored to different input intervals, the position-variable convolution effectively enhances the capability to model long-term dependencies during waveform generation.

2.4 FCN-based Channel

The Fourier-based channel is designed to analyze the frequency components of signals through Fourier Transform. In the proposed dual-channel architecture, we introduce Fourier Transform as a second input pathway, converting raw waveform data into frequency-domain featurescvia Fast Fourier Transform (FFT). It allows to exploit the rich information embedded in the frequency domain, complementing the captured temporal dynamics. The Fourier Transform of an input waveform $x(t)$ is given with:

$$\boldsymbol{X}(f) = \sum_{t=0}^{T-1} x(t) e^{-j2\pi ft} \,, \tag{3}$$

$\boldsymbol{X}(f)$ is the complex spectrum in the frequency domain, $x(t)$ is the time-domain signal, f is the frequency, T is the length of the time-domain signal. To obtain the magnitude of each frequency component, let:

$$|\boldsymbol{X}(f)| = \sqrt{\Re(X(f))^2 + \Im(X(f))^2} \,. \tag{4}$$

where $|\boldsymbol{X}(f)|$ serves as an additional feature channel that is integrated into the network alongside the time-domain features. During the forward pass of the model, the input x is transformed into its spectral features $\boldsymbol{X}(f)$ via the Fourier Transform. These spectral features are then concatenated with

the time-domain features to form a composite feature channel. The processed spectral features and the original features obtained from the LVCNet are concatenated along the channel dimension, resulting in a more comprehensive representation that captures both temporal and spectral characteristics of the waveform.

Integrating Fourier Transform into neural network architectures requires careful consideration of computational efficiency. The Discrete Fourier Transform (DFT) has a complexity of $O(T^2)$, which is impractical for large-scale waveform data. By leveraging the Fast Fourier Transform (FFT), this is reduced to $O(T \log T)$, making real-time processing feasible. Compared to other frequency-domain feature extraction methods, FFT balances efficiency and interpretability: the Discrete Wavelet Transform (DWT) offers $O(T)$ complexity but requires careful wavelet selection; Mel-Frequency Cepstral Coefficients (MFCC) introduce additional computational overhead due to the mel-scale filter bank and Discrete Cosine Transform (DCT); while DCT shares $O(T \log T)$ complexity with FFT, it lacks phase information crucial for waveform characteristics. Given these trade-offs, FFT is chosen for its computational feasibility and effectiveness in spectral feature extraction.

2.5 MatMul-free Dense Layer

To enhance the efficiency of the proposed network, we introduce a MatMul-free dense layer [27], which replaces traditional matrix multiplication operations with more computationally efficient alternatives. In conventional neural networks, dense layers compute the weighted sum of inputs using matrix multiplication between an input vector and a weight matrix. Given an input vector $\boldsymbol{a} \in \mathbb{R}^{1\times n}$ and a weight matrix $\boldsymbol{V} \in \mathbb{R}^{n\times p}$, the standard matrix multiplication can be expressed as:

$$\boldsymbol{b} = \boldsymbol{a}\boldsymbol{V} = \sum_{k=1}^{n} a_k V_{kj} \quad \text{for} \quad j=1, 2,\ldots, p\,, \tag{5}$$

where $\boldsymbol{b} \in \mathbb{R}^{1\times p}$ represents the output. While this method is straightforward and effective, matrix multiplication typically incurs significant computational costs. To optimize the computation process and reduce hardware resource consumption, we propose the adoption of BitNet, which utilizes binary linear modules (BLMs) to replace traditional dense layer computations. BitNet leverages ternary weights (i.e., weights taking values of -1, 0, or +1) to simplify matrix multiplication into addition operations. In this layer, traditional matrix multiplication is replaced by ternary matrix operations. Let $\widetilde{\boldsymbol{V}}$ be a ternary weight matrix where each element $\widetilde{V}_{kj} \in \{-1,0,+1\} \in \{-1,0,+1\}$. The output can be represented with:

$$\widetilde{\boldsymbol{b}} = \boldsymbol{a} \odot \widetilde{\boldsymbol{V}} = \sum_{k=1}^{n} a_k \widetilde{V}_{kj}\,, \quad \widetilde{V}_{kj} \in \{-1, 0, +1\}, \quad \text{for} \quad j=1, 2,\ldots, p\,, \tag{6}$$

where $\widetilde{\boldsymbol{b}} \in \mathbb{R}^{1\times p}$ is the output vector, and $\odot$ denotes the ternary matrix operation. The elements of the ternary weight matrix $\widetilde{V}_{kj}$ can only take values of -1, 0, or +1. Therefore, during each multiplication operation, the traditional multiplication can be directly substituted with addition or subtraction:

$$a_k \widetilde{V}_{kj} = \begin{cases} a_k, & \text{if } \widetilde{V}_{kj} = +1 \\ 0, & \text{if } \widetilde{V}_{kj} = 0 \\ -a_k, & \text{if } \widetilde{V}_{kj} = -1 \end{cases} \tag{7}$$

Thus, the ternary matrix multiplication can be further simplified as:

$$\widetilde{b}_j = \sum_{k=1}^{n} a_k \widetilde{V}_{kj} = \sum_{\widetilde{V}_{kj}=+1} a_k - \sum_{\widetilde{V}_{kj}=-1} a_k\,, \quad \text{for} \quad j=1, 2,\ldots, p\,. \tag{8}$$

By transforming the original matrix multiplication into ternary additions and subtractions, the method not only enhances computational efficiency but also reduces the complexity of multiplication operations. Using ternary weights in dense layers significantly lowers energy consumption and storage needs for resource-constrained devices without sacrificing performance.

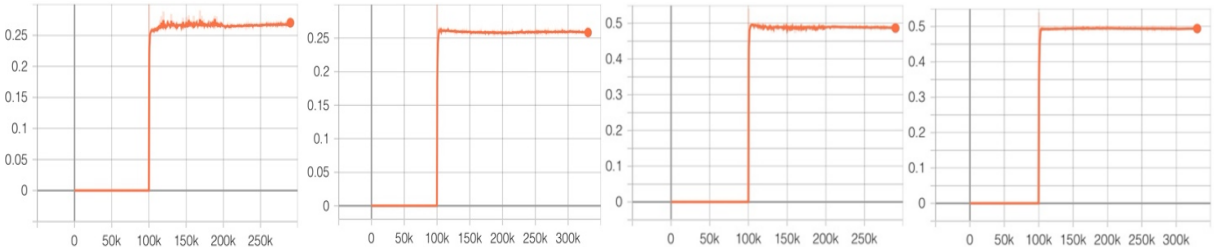
3. Experimental Evaluation

3.1. Experiment design

The model was trained on the LJSpeech dataset[28], divided into 12,600 training, 400 validation, and 100 testing samples at a 22,050 Hz sampling rate. Mel-spectrograms were computed using STFT with a Hann window (FFT size: 1024, window size: 1024, hop size: 256) and converted to the Mel scale via an 80-channel filterbank covering 80 Hz to 7.6 kHz. Our enhanced LVCNet generator integrates Fourier Transform and original LVCNet channels, with a Parallel WaveGAN-based discriminator[29]. The generator has three LVCNet blocks, each with 10 LVC layers and 8 residual channels. Frequency-domain features are processed through a Fourier Convolution Network (FCN) with 16 channels and concatenated with time-domain features, using weight normalization. The training was conducted on a single NVIDIA V100-32GB GPU with PyTorch 1.7.0 and CUDA 11.0, employing a batch size of 8, Adam optimizer with learning rates of 0.0001 (generator) and 0.00005 (discriminator), and learning rate decay by a factor of 0.5 every 200,000 steps. The model was trained using a combination of multi-resolution STFT loss with three different FFT sizes (512, 1024, 2048) and adversarial loss weighted with a coefficient of 4.0. We evaluated performance using both objective metrics (RTF, MCD, STOI, PESQ) and subjective metrics (MOS, CMOS). The enhanced LVCNet was compared with the original LVCNet under identical conditions, including the same dataset partitioning, preprocessing methods, and training steps. By monitoring training losses, we confirmed that our dual-channel approach significantly improved synthesis quality and efficiency compared to the original single-channel LVCNet. The integration of frequency-domain processing with traditional time-domain methods proved particularly beneficial for capturing both local temporal details and global spectral characteristics in waveform generation.

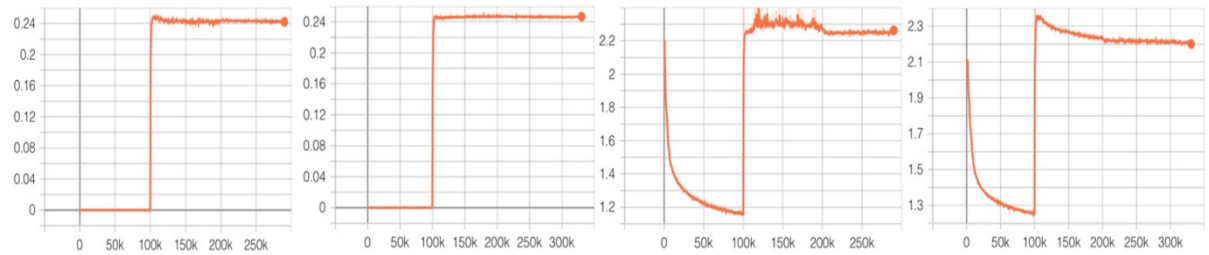
3.2 Result analysis

The trained loss for the original LVCNet and the proposed waveform generation network are shown in Figures 2~5. Key metrics during adversarial training include adversarial loss, discriminator loss for real vs. fake samples, generator loss for fake sample quality, log STFT magnitude loss, and spectral convergence loss measuring proximity to real audio spectra. These metrics comprehensively assess the performance, highlighting improvements by the proposed network. From the results, the proposed network outperforms the original LVCNet in waveform quality and training stability. It reduces losses.



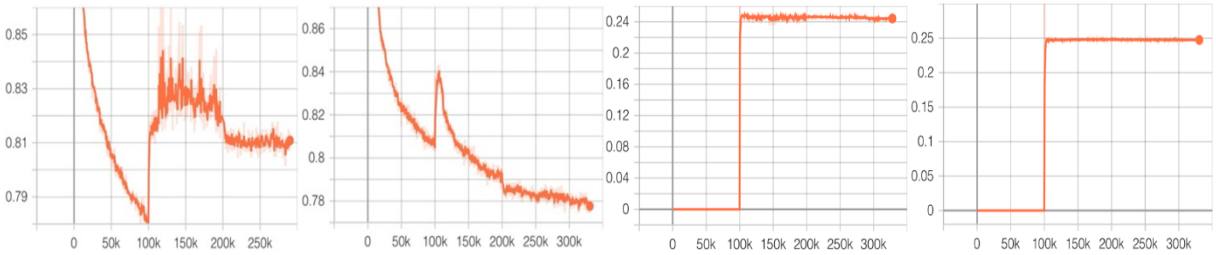
(*a*) Original LVCNet (*b*) Proposed method (*c*) Original LVCNet (*d*) Proposed method

Figure 2. Comparison of the trained *adversarial loss* ((*a*)(*b*)), and trained *discriminator loss*((*c*)(*d*)).



(*a*) Original LVCNet (*b*) Proposed method (*c*) Original LVCNet (*d*) Proposed method

Figure 3. Comparison of the trained *fake loss* ((*a*)(*b*)), and that trained *generator loss* ((*c*)(*d*)).



(*a*) Original LVCNet (*b*) Proposed method (*c*) Original LVCNet (*d*) Proposed method

Figure 4. Comparison of the trained *magnitude loss* ((*a*)(*b*)), and that trained *real loss* ((*c*)(*d*)).

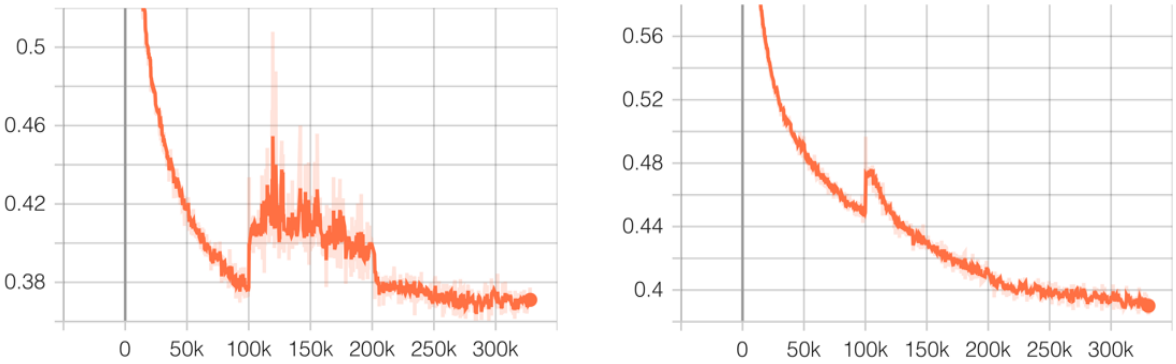
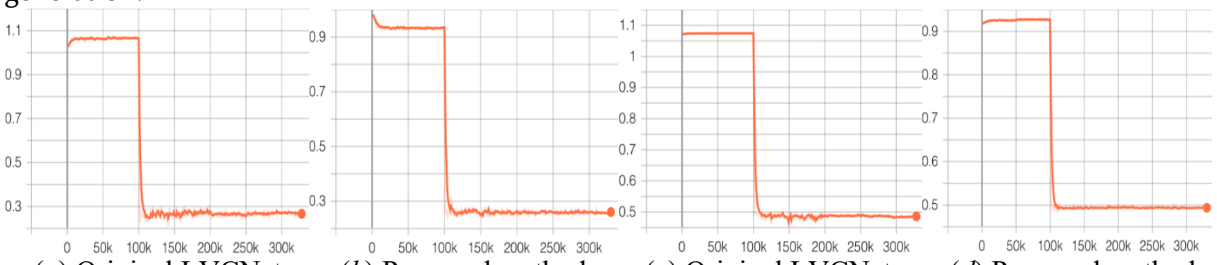


Figure 5. The trained *convergence loss* of the original LVCNet (left) and proposed method (right).

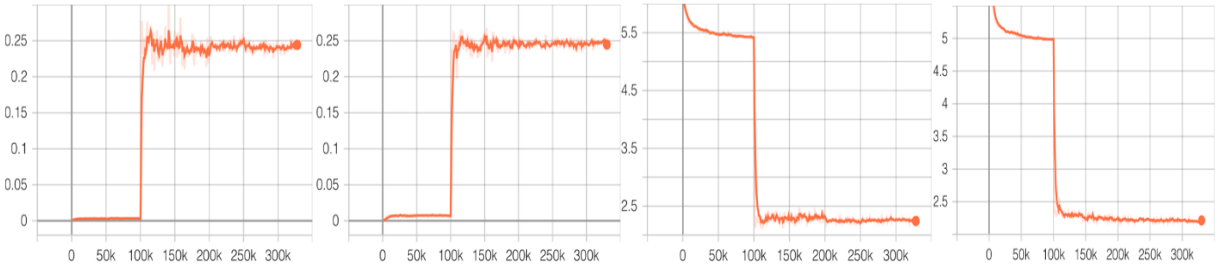
More rapidly with better monotonicity and a more stable decreasing trend, showing reduced fluctuations. Based on the loss curves from both training and evaluation phases, the improved network demonstrates superior performance compared to the original LVCNet in multiple aspects. Particularly in terms of waveform generation quality and training stability, the enhanced network reduces losses more quickly and exhibits better monotonicity than the original LVCNet model. The loss function curves show a more consistently decreasing trend with significantly reduced oscillations, indicating stronger stability in audio generation quality, enhanced robustness against interference, and reduced training randomness. Specifically, the log STFT magnitude loss curve reveals substantial improvement in amplitude differences between generated and real waveforms in the frequency domain. The improved model enables the log STFT magnitude loss to decrease more steadily with increasing training iterations, without many obvious and intense oscillations observed in the original model. This indicates that after integrating the Fourier Transform network channel, the original LVCNet model generates waveforms that are closer to real waveforms in the frequency domain, resulting in relatively higher quality. Figure 6˜10 demonstrate the evaluated loss from the comparative methods. The proposed method demonstrates superior performance compared to the original LVCNet, as evidenced by faster convergence and lower loss values across adversarial, discriminator, fake, generator, magnitude,

real, and convergence losses. Overall, our network surpasses the original LVCNet in both training and evaluation, yielding superior audio quality and a smoother, more efficient training process, underscoring the effectiveness and potential of the proposed improvement in waveform generation.



(*a*) Original LVCNet (*b*) Proposed method (*c*) Original LVCNet (*d*) Proposed method

Figure 6. Comparison of evaluated *adversarial loss* ((*a*)(*b*)), and evaluated *discriminator loss*((*c*)(*d*)).



(*a*) Original LVCNet (*b*) Proposed method (*c*) Original LVCNet (*d*) Proposed method

Figure 7. Comparison of the evaluated *fake loss* ((*a*)(*b*)), and that evaluated *generator loss* ((*c*)(*d*)).

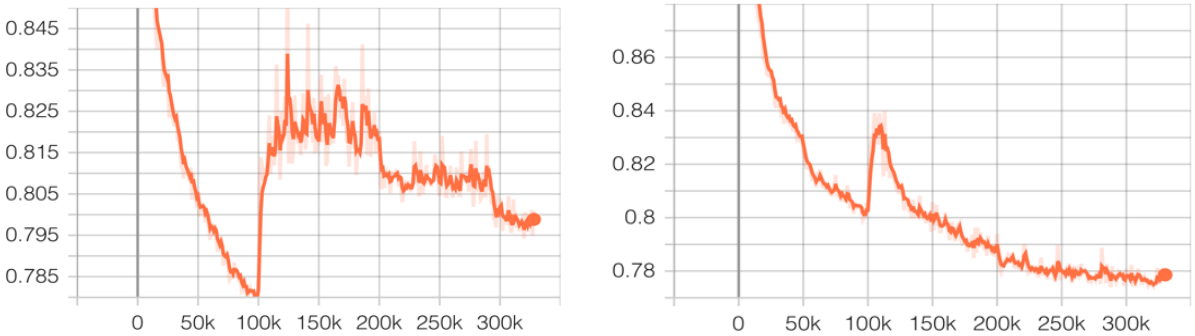


Figure 8. The evaluated *magnitude loss* of original LVCNet (left) and the proposed method (right).

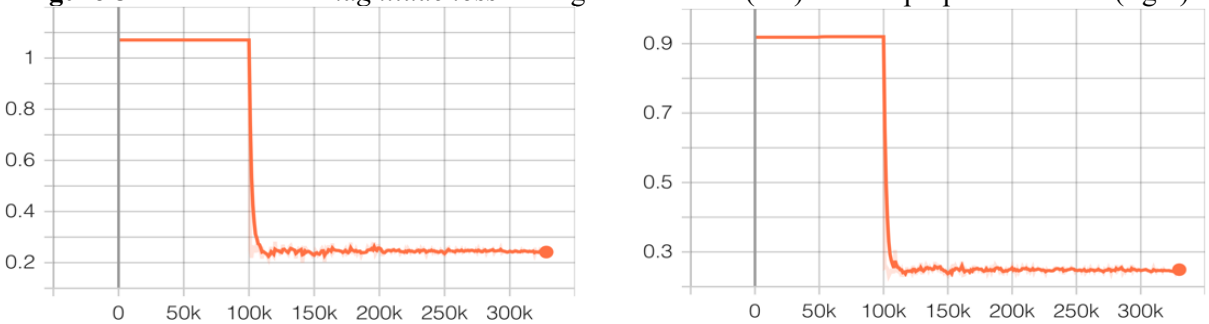


Figure 9. The evaluated *real loss* of original LVCNet (left) and the proposed method (right).

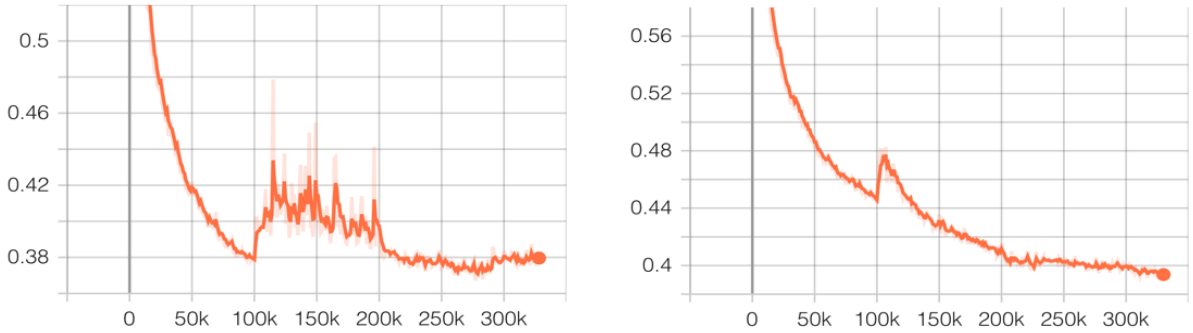


Figure 10. The evaluated *convergence loss* of original LVCNet (left) and the proposed method (right).

The integration of time-domain and frequency-domain features provides significant advantages over the original LVCNet's time-domain-only approach. Spectral analysis of the generated samples reveals that our dual-channel architecture achieves better harmonic representation in sustained vowels and improved spectral consistency during phoneme transitions. This improvement is primarily attributed to the FCN's ability to model global frequency relationships that complement LVCNet's local temporal modeling. For consonant sounds, spectrogram comparisons show that the original LVCNet often struggles to reproduce accurate spectral transitions, exhibiting energy smearing across frequency bands, whereas our enhanced model maintains clearer temporal boundaries while preserving more accurate spectral distribution. The time-domain channel effectively preserves phase information essential for natural articulation, while the frequency-domain channel simultaneously enhances formant structure definition. This complementary modeling approach addresses the fundamental limitations of single-domain processing: where time-domain features excel at capturing transient details and local dependencies, frequency-domain features ensure global spectral consistency, resulting in more coherent and natural waveforms. The objective and subjective evaluation results confirm these theoretical advantages, demonstrating measurable improvements in both perceptual quality and intelligibility metrics compared to the single-channel LVCNet baseline.

4. Conclusion

This paper presents a dual-channel architecture that integrates both time-domain and frequency-domain features for waveform generation in audio synthesis tasks. By augmenting the LVCNet model with a Fourier Transform module, we leverage both temporal and spectral characteristics to enhance the quality and efficiency of audio synthesis. Experimental results demonstrate that the proposed network significantly outperforms previous methods in terms of waveform accuracy, stability, and inference speed. Specifically, the dual-channel structure excels in generating high-quality waveforms and accelerating processing times, overcoming the limitations of traditional models. This innovation offers new directions for future advancements in audio synthesis technology and holds broad application potential across various audio generation tasks.

References

[1] Khan, R.A., Chitode, J.S. (2016) Concatenative Speech Synthesis: A Review. Int. J. Comput. Appl., 136(3): 1–6.

[2] Zen, H., Tokuda, K., Black, A.W. (2009) Statistical parametric speech synthesis. Speech Commun., 51(11): 1039–1064.

[3] Yu, C.-Y., Fazekas, G. (2024) Differentiable Time-Varying Linear Prediction in the Context of End-to-End Analysis-by-Synthesis. In: Interspeech, Queen Mary University of London, UK. pp. 1820–1824.
[4] Zarazaga, P.P., Malisz, Z., Henter, G.E., Juvela, L. (2023) Speaker-independent neural formant synthesis. arXiv:2306.01957.
[5] Koffi, E., Petzold, M. (2022) A tutorial on formant-based speech synthesis for the documentation of critically endangered languages. Linguist Portf., 11: 26–55.
[6] Wang, X., Takaki, S., Yamagishi, J. (2020) Neural Source-Filter Waveform Models for Statistical Parametric Speech Synthesis. IEEE/ACM Trans. Audio Speech Lang. Process., 28: 402–415.
[7] Wang, X., Takaki, S., Yamagishi, J. (2019) Neural Source-filter-based Waveform Model for Statistical Parametric Speech Synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton. pp. 5916–5920.
[8] Saito, Y., Takamichi, S., Saruwatari, H. (2018) Statistical Parametric Speech Synthesis Incorporating Generative Adversarial Networks. IEEE/ACM Trans. Audio Speech Lang. Process., 26(1): 84–96.
[9] Singh, A.K., Singh, P. (2021) Detection of AI-Synthesized Speech Using Cepstral & Bispectral Statistics. arXiv:2009.01934.
[10] Zen, H., Tokuda, K., Black, A.W. (2009) Statistical parametric speech synthesis. Speech Commun., 51(11): 1039–1064.
[11] Baas, M., Kamper, H. (2023) GAN You Hear Me? Reclaiming Unconditional Speech Synthesis from Diffusion Models. In: IEEE Spoken Language Technology Workshop, Doha.. pp. 906–911.
[12] Kong, J., Kim, J., Bae, J. (2020) HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In: Conference on Neural Information Processing Systems (NeurIPS), Vancouver. pp. 17022–17033.
[13] Sualiheen, S., Kim, D.-H. (2025) EMGVox-GAN: A transformative approach to EMG-based speech synthesis, enhancing clarity, and efficiency via extensive dataset utilization. Comput. Speech Lang., 92: 101754.
[14] Meng, L., Zhou, L., Liu, S., Chen, S., Han, B., Hu, S., Liu, Y., Li, J., Zhao, S., Wu, X., Meng, H., Wei, F. (2024) Autoregressive Speech Synthesis without Vector Quantization. arXiv:2407.08551.
[15] Zhu, X., Tian, W., Xie, L. (2024) Autoregressive Speech Synthesis with Next-Distribution Prediction. arXiv:2412.16846.
[16] Liu, Z., Wang, S., Inoue, S., Bai, Q., Li, H. (2024) Autoregressive Diffusion Transformer for Text-to-Speech Synthesis. arXiv:2406.05551.
[17] Strauss, M., Edler, B. (2021) A Flow-Based Neural Network for Time Domain Speech Enhancement. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5754–5758.
[18] Klapsas, K., Nikitaras, K., Ellinas, N., Sung, J. S., Hwang, I., Raptis, S., Chalamandaris, A., Tsiakoulis, P. (2022) Predicting phoneme-level prosody latents using AR and flow-based Prior Networks for expressive speech synthesis. arXiv:2211.01327.
[19] He, J., Zhao, Z., Ren, Y., Liu, J., Huai, B., Yuan, N. (2022) Flow-Based Unconstrained Lip to Speech Generation. In: AAAI Conference on Artificial Intelligence (AAAI-22), 36(1): Technical Tracks 1.

[20] Wen, Y., Lei, Z., Yang, Y., Liu, C., Ma, M. (2022) Multi-Path GMM-MobileNet Based on Attack Algorithms and Codecs for Synthetic Speech and Deepfake Detection. In: Interspeech, ISCA, Incheon. pp. 4795–4799.
[21] Bhushan, U., Malipatil, K., Vishruth Patil, V., Anilkumar, V., Ananya, S., Bharath, K. P. (2024) HMM and Concatenative Synthesis based Text-to-Speech Synthesis. In: International Conference on Software, Systems and Information Technology (SSITCON), Tumkur. pp. 1–6.
[22] Melechovsky, J., Mehrish, A., Herremans, D., Sisman, B. (2023) Learning Accent Representation with Multi-Level VAE Towards Controllable Speech Synthesis. In: IEEE Spoken Language Technology Workshop (SLT), Doha. pp. 928–935.
[23] Yang, F., Luan, J., Wang, Y. (2022) Improving Emotional Speech Synthesis by Using SUS-Constrained VAE and Text Encoder Aggregation. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, pp. 8302–8306.
[24] Amrouche, A., Bentrcia, Y., Boubakeur, K. N. (2022) DNN-Based arabic speech synthesis. International Conference on Electrical and Electronics Engineering. Alanya. pp. 378–382.
[25] Karpov, A., Potapova, R., Eds. (2021) Speech and Computer: International Conference, Springer International Publishing, St. Petersburg.
[26] Zeng, Z., Wang, J., Cheng, N., Xiao, J. (2021) LVCNet: Efficient Condition-Dependent Modeling Network for Waveform Generation. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto. pp. 6054–6058.
[27] Zhu, R.J., Zhang, Y., Sifferman, E., Sheaves, T., Wang, Y.Q., Richmond, D., Zhou, P., Eshraghian, J.K. (2024) Scalable MatMul-free Language Modeling. https://doi.org/10.48550/arXiv.2406.02528.
[28] Ito, K. (2017) The LJ Speech Dataset. https://keithito.com/LJ-Speech-Dataset/.
[29] Yamamoto, R., Song, E., Kim, J.-M. (2020) Parallel Wavegan: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain: IEEE, pp. 6199–6203.