UNet-VITS: Elevating Single-Stage TTS Quality with Spectral Restoration and Post-Processing Optimization

Min Zheng[1*†], Danqing Liu[2†], Tengyue Yang[1], Haoyu Liu[1], Yanhui Guo[5]
{zhengmin_824@foxmail.com, liudanqing@cdut.edu.cn, 344744841@qq.com, 1756103099@qq.com, 13264233@qq.com}

[1]College of Computer, Qinghai Normal University, Xining, China
[2]College of Computer Science and Cyber Security, Chengdu University of Technology, Chengdu, China
[5]College of Artificial Intelligence, Shandong Women's University, Shandong, China

Abstract. Recent advances in single-stage text-to-speech (TTS) synthesis have demonstrated superior performance compared to conventional pipeline systems. However, state-of-the-art models like VITS2 still suffer from insufficient naturalness, such as prosodic breaks in long utterances, poor spectral accuracy due to the loss of high-frequency details, and strong speech dependency, which results in failure to handle unseen speaking styles. This study proposes an enhanced VITS2 architecture, UNet-VITS, to address these long-standing challenges through three synergistic technical improvements—a unique combination not seen in prior research. Our method initially employs a multi-scale fusion mechanism to enhance latent feature representation and integrates residual blocks with SnakeBeta activation to optimize gradient flow and increase model capacity—both contributing to improved latent feature quality. Subsequently, we introduce a U-Net-based network architecture that treats mel-spectrograms as acoustic images and utilizes multi-scale skip connections to further refine F0-informed spectral details, thereby achieving post-hoc enhancement of audio quality. Comprehensive evaluations demonstrate significant improvements in speech naturalness and speaker similarity, while substantially reducing reliance on explicit phoneme conversion and maintaining competitive training efficiency. Practically, this work provides a reusable "full-chain optimization" paradigm for single-stage TTS, serving as a valuable reference for future TTS research.

Keywords: Speech Synthesis, U-Net, SnakeBeta Activation, End-to-End Text-to-Speech

1 Introduction

In recent years, the field of speech synthesis has undergone a significant transformation, shifting from traditional two-stage systems to a single-stage paradigm. Early two-stage approaches,

[1]Corresponding author: Min Zheng (zhengmin_824@foxmail.com).
[2]Min Zheng and Danqing Liu contributed equally to this work.

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365268

typically exemplified by the Tacotron+WaveNet framework[1, 2]—where Tacotron generates mel-spectrograms [1] and WaveNet converts them to raw audio [2]—involved acoustic models like Tacotron generating intermediate acoustic representations, such as mel-spectrograms, from text. These representations were then converted into waveforms by neural vocoders like WaveNet. While this clearly divided architecture simplified the design and training of individual modules, its inherent limitations severely constrained further improvements in speech quality. Beyond the core issue of cross-stage error propagation, this approach faced fundamental bottlenecks, including acoustic feature information loss (e.g., detail loss due to time-frequency compression in mel-spectrograms) and the limitations of manually defined features[3]. Notably, later works like "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions" [4] attempted to optimize this pipeline, but still retained the two-stage structure. Ultimately, these factors made it difficult for synthesized speech to overcome the "mechanical feel" and meet the demands of high-fidelity human-computer interaction scenarios.

End-to-end single-stage models, exemplified by VITS [5] and its improved variant VITS2 [6], have successfully overcome the inherent limitations of two-stage systems through architectural innovations. This framework innovatively integrates conditional variational autoencoders, normalizing flows, and adversarial training techniques, unifying acoustic modeling and waveform generation within an end-to-end training paradigm, thereby fundamentally redefining the technical approach to speech synthesis. Its core advantages are evident in three key aspects: First, by employing latent variable modeling to directly learn the end-to-end mapping from text to waveforms, it completely eliminates the problem of error accumulation across modules typical of two-stage systems. Second, the introduction of a stochastic duration predictor enables implicit and accurate modeling of speech rhythm and prosody, significantly enhancing the naturalness of synthesized speech. Third, through joint optimization of all components, it substantially improves both training and inference efficiency while maintaining high-quality audio output. These groundbreaking advancements have not only accelerated the maturation of single-stage speech synthesis technology but also established VITS2 as the benchmark framework for current research and applications in the field.

To address the challenges outlined above, this study proposes an enhanced VITS2-based speech synthesis framework that achieves precise "problem-solution" alignment through targeted design. To overcome limitations in naturalness and architecture, we first develop a multi-scale feature fusion mechanism that integrates acoustic features across different network levels, thereby enhancing the model's ability to capture both local speech details and global prosody[7]. Simultaneously, we introduce the SnakeBeta periodic activation function[8], which leverages adjustable frequency parameters to significantly improve the network's capacity to represent speech harmonic structures, particularly optimizing the continuous modeling of the F0 trajectory. To further overcome bottlenecks in pitch modeling and detail restoration, we innovatively incorporate a U-Net architecture in the post-processing stage—originally proposed for biomedical image segmentation [9] and later adapted for speech tasks [10]. This module treats mel-spectrograms as time-frequency images for processing. Through a multi-scale skip-connection mechanism, it deeply integrates F0-derived excitation signals extracted from the original audio with spectral features, enabling accurate reconstruction of pitch characteristics and fine restoration of acoustic details[11]. Collectively, these improvements establish a novel single-stage speech synthesis solution that balances naturalness, applicability, and

efficiency.

2 Methodology

2.1 Overall Framework Architecture

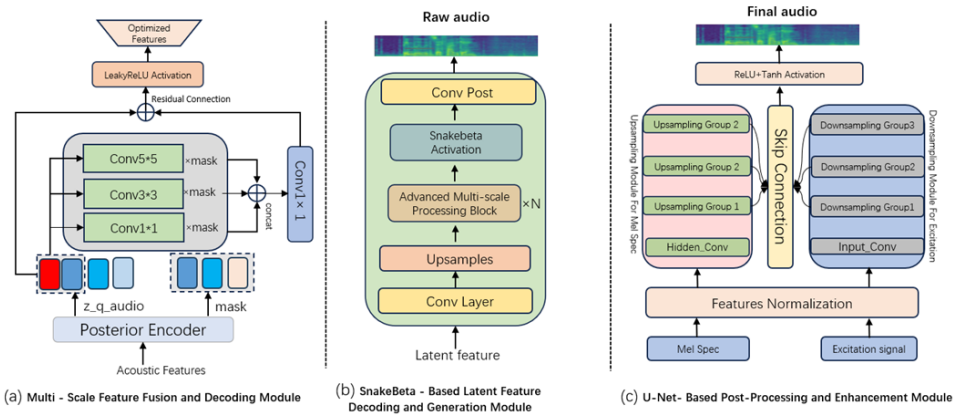

Fig. 1. Diagrams of three key modules in the enhanced speech synthesis framework: multi-scale feature fusion, SnakeBeta-based generation, and U-Net audio post-processing.

The framework proposed in this study establishes a core processing pipeline based on the VITS2 architecture (see the overall framework diagram in Fig. 1). First, through the inverse operation of the normalizing flow module, duration-related latent features and their statistics are converted into audio-domain latent features, with the corresponding mean and log standard deviation output simultaneously. Subsequently, these latent features are fed into the multi-scale feature fusion module, where deep separable convolutions extract features at different scales, masks filter out invalid regions, multi-scale features are concatenated along the channel dimension, 1×1 convolutions reduce the number of channels while preserving valid regions, and nonlinear activation combined with residual connections produces optimized latent features.

Next, these optimized latent features are input into the generator module: they first undergo a pre-convolution to map to the initial number of upsampling channels, then pass through a cyclic process involving multiple rounds of upsampling and adaptive multi-scale processing to adjust the feature channels and temporal dimensions. Afterward, the SnakeBeta activation function applies nonlinear adjustment, the output convolution layer reduces the number of channels, and finally, an activation function maps the result to a specified range to obtain the initial audio signal tensor.

Finally, a feature extractor derives mel-spectrograms and fundamental frequency (F0) signals from the initial audio signal, which are then input into the U-Net network. The two types of features are first standardized separately; the encoding phase then performs multiple rounds of residual block processing and downsampling, retaining features from each stage for subsequent skip connections.

In the decoding phase, the bottleneck layer output is fused with the skip connection features from the encoding phase, and multiple rounds of upsampling and residual block processing gradually restore the temporal dimension and reduce the number of channels. Ultimately, the output layer processes the features to generate the optimized audio signal.

The complete process can be formulated as:

$$\hat{\mathrm{Y}}_{\text{final}} = \mathrm{V}(\mathrm{U}(\mathrm{M}(\mathrm{G}(\mathrm{T})), \mathrm{F}(\mathrm{G}(\mathrm{T})))) \tag{1}$$

where T denotes the input text feature sequence, G represents the enhanced generator, M extracts mel-spectrograms, F extracts F0 contours, U denotes the U-Net post-processor, and V is the neural vocoder.

2.2 Multi-Scale Feature Fusion

The multi-scale feature fusion module enhances the model's ability to capture temporal dependencies across various resolutions. It utilizes parallel depth-wise separable convolutions with kernel sizes of 1, 3, and 5, chosen to align with the hierarchical temporal characteristics of speech: 1×1 convolutions focus on local phoneme-level details (e.g., smooth transitions between /s/ and /a/), 3×1 convolutions model syllable-level rhythm, and 5×1 convolutions capture phrase-level prosody. This approach overcomes the limitations of single-scale methods, such as missing prosodic cues or blurred phoneme representations, by enabling feature extraction across multiple temporal contexts.

The fusion process combines these multi-scale features through concatenation followed by projection:

$$\mathrm{H}_{\text{out}} = \mathrm{LeakyReLU}(\mathrm{Proj}([\mathrm{H1}; \mathrm{H3}; \mathrm{H5}])) + \mathrm{H}_{\text{in}} \tag{2}$$

where $H1, H3, H5$ represent features from different convolutional branches, and the residual connection preserves original feature information while injecting multi-scale context[12].

2.3 SnakeBeta Activation Function

We systematically replace conventional activation functions with the advanced SnakeBeta activation in all residual blocks throughout the generator architecture. SnakeBeta is a specialized periodic activation function explicitly designed for audio signal processing, demonstrating superior capability in modeling the complex oscillatory nature of speech waveforms compared to traditional ReLU activations[13]. The SnakeBeta function includes learnable parameters (α for frequency regulation, β for magnitude adjustment) that dynamically regulate the frequency and magnitude of its periodic component—we initialize $\alpha = 1.0$ (to match the 200–500 Hz harmonic frequency range of speech in LJ Speech) and $\beta = 2.0$ (to balance feature modulation intensity and training stability, avoiding gradient explosion or linear degeneration), as described by the following formulation:

$$\mathrm{SnakeBeta}(x) = x + \frac{1}{\beta + 10^{-6}} \cdot \sin^2(\alpha \cdot x) \tag{3}$$

where parameters α and β control frequency and magnitude of periodic oscillations, adapting to speech characteristics during training.

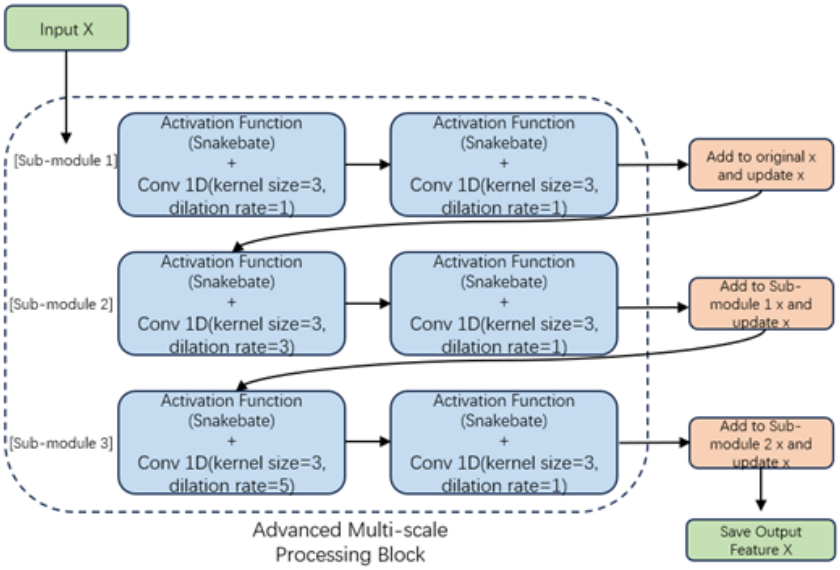

Fig. 2. Execution Flow of the AMPBlock (Adaptive Multi-scale Periodic Block) Module. The feature propagates through sub-modules and undergoes residual updates to generate the optimized output.

2.4 U-Net Post-Processing Module

The U-Net post-processing module employs an encoder-decoder architecture, treating mel-spectrograms as time-frequency images for processing while incorporating excitation signals derived from fundamental frequency (F0) to facilitate feature interaction with mel-spectrograms through skip connections. Specifically, the extracted mel-spectrogram features are concatenated and fused with standardized F0 excitation signal features at corresponding stages. By leveraging the high-resolution information transmission capability of skip connections, prosodic and spectral structure information collaborate throughout the process, thereby preserving feature integrity across the entire network and enabling accurate alignment of these two key types of information[14].

The dual inputs of mel-spectrograms and F0 signals are designed to overcome the limitations of single-feature inputs, based on two primary considerations: first, mel-spectrograms simulate human auditory characteristics, preserving speech spectral structure details (such as vowel formants) to correct high-frequency distortions in the initial audio, and their 2D time-frequency morphology aligns well with U-Net convolution's strength in extracting spatial features; second, F0 signals directly characterize fundamental frequency variations and are essential for conveying prosodic information (such as intonation fluctuations), and their derived excitation signals guide the U-Net to optimize harmonic structures and reduce prosodic breaks—effects unattainable by mel-spectrograms alone.

The core operation of this module can be expressed as:

$$\hat{Y}_{\text{final}} = U(M_{\text{init}}, F0) \tag{4}$$

where the U-Net processor U takes both the initial mel-spectrogram M_{init} and fundamental frequency information $F0$ as inputs and directly outputs the final audio waveform $\hat{Y}_{\text{final}}$.

In the encoder pathway, the network implements three successive downsampling stages with carefully selected reduction factors of $[4,4,2]$, employing kernel sizes of $[8,8,4]$, respectively. Each stage incorporates Multi-Receptive Field (MRF) residual blocks that process features using depth-wise separable convolutions with a kernel size of 3 and dilation patterns of $[1,3,5]$[15]. The encoder operation at each level can be mathematically represented as:

$$H_l = \text{Downsample}_l \left(\frac{1}{N} \sum_{j=1}^{N} \text{ResidualBlock}_j^{(l)}(H_{l-1}) \right) \tag{5}$$

where H_l denotes the feature map at level l, the downsampling operation reduces the temporal resolution while doubling the channel dimensionality—from an initial 256 channels to 2048 channels at the bottleneck. The architecture supports both causal and non-causal convolution modes through the implementation of CausalConv1d and CausalConvTranspose1d operations, making it suitable for both real-time streaming and offline processing scenarios. Weight normalization is applied throughout the network to enhance training stability.

The core innovation of our U-Net architecture lies in its sophisticated skip connection mechanism. During the decoding phase, at each stage l, features from the corresponding encoder level $4-l$ are concatenated with upsampled features from the previous decoder level:

$$H_l^{\text{dec}} = \text{Concat}(\text{Upsample}(H_{l-1}^{\text{dec}}), H_{4-l}^{\text{enc}}) \tag{6}$$

This concatenation operation ensures that high-resolution information from early encoding stages remains directly accessible during reconstruction. The upsampling process employs transposed convolutions with kernel sizes $[4,8,8]$ and strides $[2,4,4]$, symmetrically reversing the encoder's downsampling pattern.

The concatenated features then undergo refinement through MRF processing:

$$H_l^{\text{dec}} = \frac{1}{N} \sum_{j=1}^{N} \text{ResidualBlock}_j^{(l)}(H_l^{\text{dec}}) \tag{7}$$

The bottleneck layer serves as a critical integration point where the deepest encoded features merge with the normalized mel-spectrogram input through a hidden convolutional operation:

$$H_{\text{bottleneck}} = \text{HiddenConv}(M_{\text{norm}}) + H_3 \tag{8}$$

The final audio output is generated by processing the refined decoder features through an output convolutional layer with a kernel size of 7, followed by a tanh activation function that directly produces the final audio waveform:

$$\hat{Y}_{\text{final}} = \tanh(\text{OutputConv}(H_3^{\text{dec}})) \tag{9}$$

This architectural approach effectively balances F0-guided processing for harmonic enhancement with spatial-spectral detail preservation through skip connections. The entire U-Net processing pipeline facilitates significant improvements in output naturalness and audio quality through the

coordinated application of these mathematical transformations. The module's capability to directly generate final audio waveforms, along with its robust handling of variable-length sequences and support for both causal and non-causal processing, makes it especially well-suited for practical speech synthesis applications[16].

3 Experiments

3.1 Datasets and Experimental Setting

To evaluate the performance of the proposed model, we utilized two public speech datasets: the LJ Speech single-speaker dataset [17] and VCTK multi-speaker dataset[18]. Samples were randomly split into training and test sets in a 9:1 ratio. Audio data were quantized at 16 bits and sampled at 22.05 kHz.

Mel-spectrogram features were extracted using the Short-Time Fourier Transform (STFT) with parameters where the FFT size was set to 2048, hop length to 256, and window size to 1024, retaining 80 mel bands. Model training was performed on an RTX 4090 GPU with 24 GB of memory.

Key training hyperparameters were configured as follows: initial learning rate 2.0×10^{-4}; learning rate decay coefficient 0.999875; batch size 8; audio segment length 4096 samples ($\approx$0.186 s, based on 22.05 kHz sampling rate); mel loss weight 45; duration KL loss weight 2; audio KL loss weight 0.05.

To comprehensively assess the quality of synthesized speech, we employed four evaluation metrics: Mel Cepstral Distance (MCD) – lower values indicate better spectral similarity; Mean Opinion Score (MOS) – higher values indicate better subjective naturalness; Spectral Convergence (Spec.C) – lower values indicate better spectral consistency; Pitch Similarity (Pitch.S) – higher values indicate better prosodic matching.

3.2 Comparison With Other Methods

The proposed UNet-VITS model was evaluated against five commonly used speech synthesis techniques: Baseline (VITS2), HiFiGAN (used as the post-decoder for VITS2)[19], MB-iSTFT-ViT (denoted as MB-ViT)[20], Glow-TTS[21], and FastSpeech[22].

3.2.1 Results on LJ Speech Dataset

UNet-VITS achieved the lowest MCD (5.98±0.11) and Spec.C (1.0929), and the highest MOS (3.524±0.04) and Pitch.S (0.961159) among all compared methods, demonstrating superior spectral similarity, subjective naturalness, and prosodic matching.

3.2.2 Results on VCTK Dataset

UNet-VITS maintained its competitive advantage with the lowest MCD (5.66±0.09) and Spec.C (1.0112), and the highest Pitch.S (0.971159). While FastSpeech achieved a slightly higher

Table 1: Comparison of MCD, MOS, Spec.C, and Pitch.S for Different Speech Synthesis Models on LJ Speech

Method	MCD(↓)	MOS(↑)	Spec.C(↓)	Pitch.S(↑)
Baseline	7.62±0.12	3.413±0.07	1.5545	0.860613
FastSpeech	6.12±0.14	3.489±0.04	1.1002	0.955427
HiFiGAN	6.33±0.09	3.391±0.12	1.2753	0.909669
Glow-TTS	6.43±0.10	3.421±0.04	1.6025	0.910057
MB-ViT	6.67±0.07	3.424±0.09	1.3856	0.942213
UNet-VITS	**5.98±0.11**	**3.524±0.04**	**1.0929**	**0.961159**

Table 2: Comparison of MCD, MOS, Spec.C, and Pitch.S for Different Speech Synthesis Models on VCTK

Method	MCD(↓)	MOS(↑)	Spec.C(↓)	Pitch.S(↑)
Baseline	6.81±0.21	3.481±0.12	1.3289	0.910331
FastSpeech	5.73±0.31	**3.689±0.19**	1.0472	0.968132
HiFiGAN	6.17±0.18	3.591±0.53	1.1233	0.939327
Glow-TTS	6.24±0.08	3.652±0.28	1.4963	0.953361
MB-ViT	6.52±0.24	3.595±0.14	1.3856	0.938241
UNet-VITS	**5.66±0.09**	3.641±0.04	**1.0112**	**0.971159**

MOS (3.689±0.19), UNet-VITS's MOS (3.641±0.04) was more stable (smaller standard deviation), indicating better consistency across multiple speakers.

3.2.3 Analysis of Results

On the LJ Speech dataset, UNet-VITS outperformed all other methods across all key metrics, significantly reducing the "mechanical feel" of synthesized speech compared to the baseline VITS2 model. The multi-scale feature fusion module effectively captured both local phoneme details and global prosody, while the SnakeBeta activation function optimized the modeling of speech harmonic structures and F0 trajectories.

On the multi-speaker VCTK dataset, UNet-VITS exhibited strong generalization capability with minimal performance degradation, maintaining clear advantages in MCD, Spec.C, and Pitch.S. The U-Net post-processing module's fusion of F0-derived excitation signals with spectral features enabled accurate pitch reconstruction even for diverse speaker voices, addressing the common issue of prosodic inconsistency in multi-speaker synthesis.

These consistent findings across datasets underscore the impact of UNet-VITS's architectural innovations—including multi-scale fusion, SnakeBeta activation, and U-Net post-processing—in delivering state-of-the-art speech synthesis performance for both single- and multi-speaker scenarios.

4 Conclusion

In this study, we propose an enhanced single-stage speech synthesis framework named UNet-VITS to address the issues of insufficient naturalness and missing fine-grained spectral details in existing end-to-end TTS systems. We introduce three core improvements: a multi-scale feature fusion mechanism to simultaneously capture local phonetic details and global prosodic structures, the SnakeBeta periodic activation function with learnable frequency parameters to improve the modeling of speech harmonics and continuous F0 trajectories, and a U-Net-based post-processing module that effectively fuses F0-related excitation signals with spectral features to achieve accurate pitch reconstruction and high-frequency detail restoration. Experimental results on both single-speaker and multi-speaker datasets demonstrate that our framework significantly improves speech naturalness, spectral fidelity, and pitch consistency while maintaining competitive training and inference efficiency. Although challenges still exist in cross-lingual synthesis, low-resource speaker adaptation, and extremely low-latency streaming scenarios, this work provides a feasible full-chain optimization paradigm for single-stage TTS and establishes a solid foundation for further research toward more robust, adaptive, and human-like speech generation systems.

Acknowledgments

The authors thank the anonymous reviewers for their valuable comments and suggestions.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT-4 in order to: Grammar and spelling check of the abstract and introduction. After using this tool, the author(s) reviewed and edited all content and took full responsibility for the publication's content.

References

[1] Wang Y, Skerry-Ryan RJ, Stanton D, Wu Y, Weiss RJ, Jaitly N, et al.. Tacotron: Towards end-to-end speech synthesis; 2017. arXiv preprint arXiv:1703.10135.

[2] Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, et al.. Wavenet: A generative model for raw audio; 2016. arXiv preprint arXiv:1609.03499.

[3] Li N, Liu S, Liu Y, Zhao S, Liu M, Zhou M. Close to human quality TTS with transformer; 2018. arXiv preprint arXiv:1809.08895.

[4] Shen J, Pang R, Weiss RJ, Schuster M, Jaitly N, Yang Z, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2018. p. 4779-83.

[5] Kim J, Kong J, Son J. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: International Conference on Machine Learning. PMLR; 2021. p. 5530-40.

[6] Kong J, Park J, Kim B, Kim J, Kong D, Kim S. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design; 2023. arXiv preprint arXiv:2307.16430.

[7] Luo J, Yang J. An Efficient Emotional Speech Synthesis Approach via Multi-scale Feature Generation. In: International Conference on Intelligent Computing. Singapore: Springer Nature Singapore; 2025. p. 3-16.

[8] Sitzmann V, Martel J, Bergman A, Lindell D, Wetzstein G. Implicit neural representations with periodic activation functions. Advances in Neural Information Processing Systems. 2020;33:7462-73.

[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing; 2015. p. 234-41.

[10] Cenik Y. Phase-Aware Speech Super Resolution Using U-Net Architecture with Lattice Topology. Middle East Technical University (Turkey); 2024.

[11] Sun C, Jiang W, Leng Y, Chen F. A new speech enhancement method based on Swin-UNet model. Noise Control Engineering Journal. 2023;71(4):258-67.

[12] Zhao J, Li R, Tian M, An W. Multi-view self-supervised learning and multi-scale feature fusion for automatic speech recognition. Neural Processing Letters. 2024;56(3):168.

[13] Lee SH, Choi HY, Kim SB, Lee SW. Hierspeech++: Bridging the gap between semantic and acoustic representation of speech by hierarchical variational inference for zero-shot speech synthesis. IEEE Transactions on Neural Networks and Learning Systems. 2025.

[14] Kaneko T, Tanaka K, Kameoka H, Seki S. iSTFTNet: Fast and lightweight mel-spectrogram vocoder incorporating inverse short-time Fourier transform. In: ICASSP 2022–2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2022. p. 6207-11.

[15] Zhao S, Zhou K, Pan Z, Ma Y, Zhang C, Ma B. HiFi-SR: A Unified Generative Transformer-Convolutional Adversarial Network for High-Fidelity Speech Super-Resolution. In: ICASSP 2025–2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2025. p. 1-5.

[16] Stoller D, Ewert S, Dixon S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation; 2018. arXiv preprint arXiv:1806.03185.

[17] Ito K, Johnson L. The lj speech dataset; 2017.

[18] Yamagishi J, Veaux C, MacDonald K. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92); 2019.

[19] Kong J, Kim J, Bae J. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in Neural Information Processing Systems. 2020;33:17022-33.

[20] Kawamura M, Shirahata Y, Yamamoto R, Tachibana K. Lightweight and high-fidelity end-to-end text-to-speech with multi-band generation and inverse short-time fourier transform. In: ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2023. p. 1-5.

[21] Kim J, Kim S, Kong J, Yoon S. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. Advances in Neural Information Processing Systems. 2020;33:8067-77.

[22] Ren Y, Ruan Y, Tan X, Qin T, Zhao S, Zhao Z, et al. Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems. 2019;32.