Neural Wavelet Packet-Based Bidirectional Autoencoder for Multi-Resolution Speech Enhancement

Abstract

Speech enhancement is a critical challenge in signal processing, particularly in noisy environments where preserving intelligibility and perceptual quality is essential. Unlike conventional deep learning-based models that operate exclusively in either the time or frequency domain, we present an adaptive multi-resolution approach that enables superior noise suppression while meticulously preserving critical speech structures across diverse frequency bands. To this end, we introduce the *Neural Wavelet Packet-Based Bidirectional Autoencoder (NWPA)*, a novel framework for multi-resolution speech enhancement. NWPA leverages the Fast Discrete Wavelet Packet Transform with trainable filters that jointly decompose both approximation and detail sub-bands, capturing richer time-frequency features than traditional fixed-wavelet approaches. A bidirectional autoencoder design reduces parameter overhead by unifying the encoding and decoding stages, while an improved Learnable Asymmetric Hard Thresholding function adaptively suppresses noise in the wavelet domain. Furthermore, a Sparsity-Enforcing Loss Function balances reconstruction fidelity with wavelet sparsity, preserving critical speech components across multiple resolutions. Comprehensive evaluations on the VoiceBank-DEMAND dataset demonstrate NWPA's state-of-the-art performance, underscoring its effectiveness in both noise reduction and intelligibility preservation. These results highlight NWPA's potential as a robust and scalable solution for speech enhancement under diverse noise conditions. The source code is available at: This repository.

Keywords: Speech Enhancement, Bidirectional Autoencoder, Wavelet Packet Transform, Signal Analysis.

1. Introduction

Speech signal processing under challenging noise conditions is fundamental in applications such as hearing aids, teleconferencing, and automatic speech recognition. These tasks demand robust techniques that can isolate speech from diverse background noises while preserving essential timbre and intelligibility. Traditional approaches often rely on time-frequency decomposition methods, most notably the Short-Time Fourier Transform (STFT), which factorizes the signal into overlapping frames and performs frequency analysis on each. However, pure STFT-based methods can suffer from limited resolution trade-offs (i.e., wide time windows capturing narrow frequency details or short windows capturing coarser frequency details), thereby hampering optimal denoising across multiple scales [1].

To address these shortcomings, many recent deep learning frameworks have resorted to learned transformations or multi-scale architectures. For instance, time-domain approaches such as Conv-TasNet [2] and multi-stage U-Nets [3] aim to discover more flexible representations than fixed STFT windows. Nonetheless, these solutions can grow in complexity, and their learned filters may not provide interpretable, multi-resolution decompositions [4].

Wavelet-based methods have also been proposed for speech enhancement and analysis, offering simultaneous time-frequency localization that can capture transient speech cues more effectively than purely STFT-based methods [5]. Early works used fixed wavelet transforms, such as Daubechies or Haar wavelets. Such transforms improved noise suppression at multiple scales but lacked adaptive filters that could tailor the decomposition to particular speech and noise profiles. Meanwhile, advanced data-driven transforms such as Adaptive Wavelet Distillation [6] and Learnable Wavelet Transform [7, 8] demonstrated that wavelet filters

can be learned in tandem with neural networks, yielding more flexible multi-resolution representations. However, most of these studies still focus on "wavelet-only" decomposition or rely on single final approximation sub-bands, leaving more comprehensive wavelet packet structures underexplored.

Simultaneously, autoencoders, especially in deep learning, have demonstrated strong capabilities in feature extraction and denoising tasks [9, 10]. Nonetheless, classical autoencoders tend to be unidirectional, where separate networks are used for encoding and decoding, resulting in duplicated parameters and suboptimal use of memory. Recent advancements in bidirectional autoencoders (BAEs) offer a more efficient alternative by sharing the same weights for forward and backward passes [11]. This design halves the parameter count and can enhance computational efficiency while still capturing crucial latent representations.

In this paper, we present the *Neural Wavelet Packet-Based Bidirectional Autoencoder (NWPA)*, a novel end-to-end framework that bridges the strengths of wavelet packet transforms and bidirectional autoencoders for speech enhancement. Unlike standard wavelet transforms (e.g., DWT) or static wavelet packet methods, our approach:

- *Adapts wavelet packet filters*, allowing the decomposition to dynamically learn both approximation- and detail-like sub-bands best suited to speech signals.
- *Employs a bidirectional architecture*, reducing the memory footprint and complexity by using the same network for encoding and decoding.
- *Incorporates a Learnable Asymmetric Hard Thresholding (LAHT) mechanism*, applying subband-specific noise suppression that retains vital speech features.
- *Enforces sparsity on wavelet-domain via a Sparsity-Enforcing Loss Function loss (SELF)*, balancing time-domain reconstruction and multi-resolution denoising to handle various noise types effectively.

By combining these elements, NWPA not only leverages multi-scale speech representations through wavelet packet decomposition but also benefits from the parameter efficiency of bidirectional learning. Evaluations on the VoiceBank-DEMAND dataset show that our method outperforms state-of-the-art approaches. Moreover, the interpretability of wavelet-domain coefficients, coupled with the flexibility of learnable filters, makes NWPA well-suited for real-world speech processing tasks where noise patterns can be highly variable.

The remainder of this paper is organized as follows. Section 2 details the NWPA architecture, including learnable wavelet packet filters, LAHT thresholding, and the bidirectional design. Section 3 presents experimental setups, comparisons with state-of-the-art methods, and an in-depth discussion of results. Finally, Section 4 concludes the paper and highlights avenues for future research, including adaptive wavelet structures and real-time deployment.

2. Proposed Method

2.1. Motivation

Effective speech enhancement requires suppressing noise while preserving transient high-frequency cues (e.g., consonant bursts) crucial for intelligibility. However, many wavelet-based methods favor low-frequency approximation paths and discard detail coefficients, limiting their ability to capture fine spectral variations. The Fast Discrete Wavelet Packet Transform (FDWPT) enriches this process by decomposing both approximation and detail sub-bands at each level, yet classical FDWPT frameworks typically rely on fixed filters that cannot adapt to varying speech and noise characteristics. We introduce NWPA to bridge this gap. By embedding a learnable FDWPT in a bidirectional autoencoder, our approach tailors time-frequency decomposition to the data, shares weights to reduce overhead, and integrates

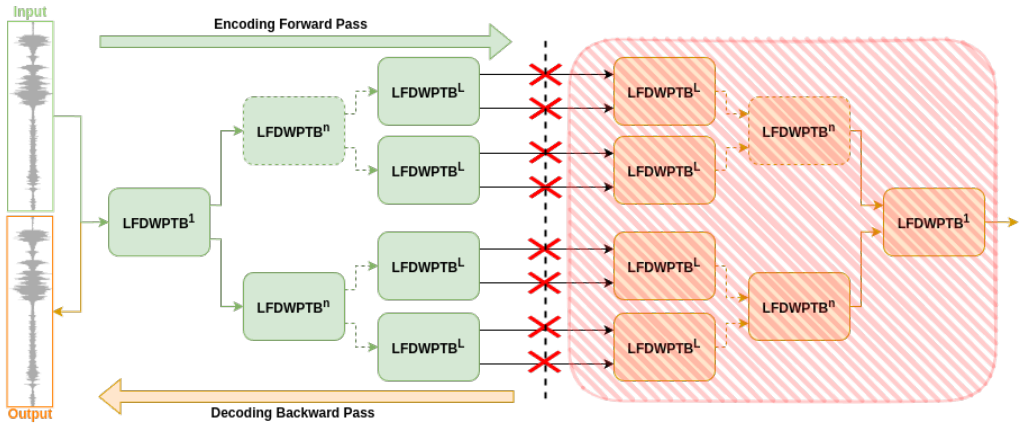


Figure 1. Overview of the NWPA architecture, integrating learnable wavelet decomposition within a bidirectional autoencoder.

adaptive thresholding (LAHT) plus a sparsity loss (SELF) for robust multi-resolution denoising. This combination preserves crucial high-frequency elements while effectively handling noise across diverse conditions.

2.2. **Method**

The proposed NWPA is an end-to-end deep learning framework for speech enhancement as illustrated in Fig. 1. It incorporates a Learnable FDWPT (LFDWPT) within a bidirectional autoencoder to perform adaptive multi-resolution analysis while ensuring efficient reconstruction. The LFDWPT allows the model to learn optimal wavelet bases that dynamically adjust to the speech signal, improving feature extraction and robustness to noise. NWPA enforces weight sharing between encoding and decoding through the Conjugate Quadrature Filter (CQF) constraint, ensuring that the same network is used for reconstruction. This reduces trainable parameters while maintaining structural consistency, enabling efficient bidirectional processing that captures both global and local speech features. To further enhance speech quality, NWPA integrates LAHT for adaptive noise suppression and SELF to promote compact and expressive representations. These components ensure selective noise reduction while preserving intelligibility. By unifying learnable wavelet decomposition, CQF-based bidirectionality, and adaptive filtering mechanisms, NWPA provides an efficient and robust approach to speech enhancement across various noise conditions.

2.2.1. **Learnable Fast Discrete Wavelet Packet Transform (LFDWPT)**

Our approach relies on the wavelet transform, a linear time-frequency tool that is particularly effective for speech enhancement and signal processing. According to the Uncertainty Principle, there is a fundamental trade-off between temporal and spectral localization. Mathematically, for a function f and its Discrete Fourier Transform Df, eq. (2.1) states:

$$\|f\|_0 \cdot \|Df\|_0 \geq n \tag{2.1}$$

where $\|f\|_0$ and $\|Df\|_0$ measure the sparsity of f and Df, respectively. Similarly, the Heisenberg Uncertainty Principle from quantum mechanics [12] is summarized by eq. (2.2):

$$\Delta x \cdot \Delta p \geq \frac{\hbar}{2} \tag{2.2}$$

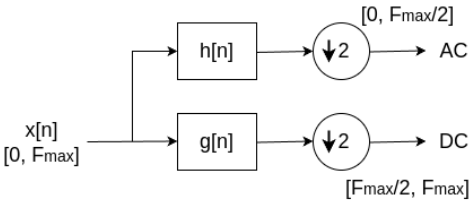


Figure 2. Traditional wavelet filter analysis (Fig. 2). The signal is iteratively split into low-pass and high-pass components at each stage, providing a multi-level decomposition.

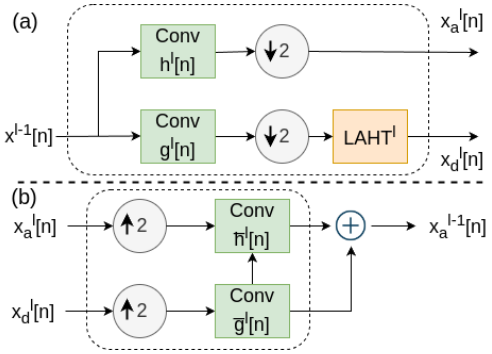


Figure 3. Learnable Fast Discrete Wavelet Transform Block (LFDWTB): (a) encoder unit, (b) decoder unit.

with Δx and Δp denoting the standard deviations of position and momentum, and $\hbar$ the reduced Planck's constant. Together, eqs. (2.1) and (2.2) highlight the difficulty of attaining perfect localization in both time and frequency domains.

The wavelet transform addresses these limitations by providing a multi-scale decomposition that is well-suited for speech signals, which inherently contain important features across multiple time and frequency scales. As shown in eq. (2.3), it operates by dilating and translating a mother wavelet $\psi(t)$:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{|a|}}\psi\Big(\tfrac{t-b}{a}\Big) \tag{2.3}$$

where a and b are scale and translation parameters. By varying these parameters, the wavelet transform adapts to signal characteristics and circumvents the constraints set by eqs. (2.1) and (2.2).

We employ FDWPT for efficient signal decomposition [13]. FDWPT constructs an orthonormal family of wavelets by scaling $\psi(t)$ in powers of 2, as specified in eq. (2.4) and the discrete wavelet transform of a signal x is given by eq. (2.5):

$$\psi_j[n] = \frac{1}{2^j}\psi\Big(\tfrac{n}{2^j}\Big) \tag{2.4}$$

$$Wx\big[n, 2^j\big] = \sum_{m=0}^{N-1} x[m]\,\psi_j^*[m-n] \tag{2.5}$$

This operation yields low-pass (approximation) and high-pass (detail) coefficients, forming a reversible orthonormal basis in $\mathbf{L}^2(\mathbb{R})$. Fig. 2 illustrates a traditional wavelet filter analysis block diagram, showing how filters h (low-pass) and g (high-pass) split the signal across multiple levels.

The inverse transform reverses this decomposition via convolution with conjugate wavelets and interpolation. Its computational efficiency arises from the cascade algorithm iteratively applying filters h and g. Equations (2.6) and (2.7) define these filters linking wavelet coefficients to the original signal in a recursive manner:

$$h[n] = \Big\langle \tfrac{1}{\sqrt{2}}\phi\big(\tfrac{t}{2}\big),\, \phi(t-n)\Big\rangle \tag{2.6}$$

$$g[n] = \Big\langle \tfrac{1}{\sqrt{2}}\psi\big(\tfrac{t}{2}\big),\, \phi(t-n)\Big\rangle \tag{2.7}$$

In NWPA, we replace static wavelet bases with trainable filters, referred to as LFDWPT. As illustrated in Fig. 3, the LFDWPT recursively decomposes $x \in \mathbb{R}^T$ at each level l by convolving with h and g and downsampling by 2, as shown in eqs. (2.8) and (2.9):

$$a_{j+1}[p] = \sum_{n=-\infty}^{+\infty} h[2p-n]\,a_j[n], \tag{2.8}$$

$$d_{j+1}[p] = \sum_{n=-\infty}^{+\infty} g[2p-n]\,a_j[n] \tag{2.9}$$

where $h^{(l)}$ and $g^{(l)}$ are parameterized as learnable tensors $\theta_h^{(l)}$ and $\theta_g^{(l)}$. To preserve energy and ensure perfect reconstruction, we impose CQF constraints. Equations (2.10) and (2.11) express these constraints:

$$g[n] = (-1)^n \, h[-n] \tag{2.10}$$

$$\bar{h}[n] = h[-n], \quad \bar{g}[n] = (-1)^{(n+1)} \, h[n] \tag{2.11}$$

yielding a full binary decomposition tree that captures both low- and high-frequency sub-bands. By training these wavelet filters via backpropagation, the learned bases adapt directly to the statistical properties of speech signals, further enhancing multi-resolution analysis and noise robustness.

2.2.2. **Unidirectional vs. Bidirectional Autoencoders**

A. Unidirectional Autoencoders (UAE). UAEs are widely employed in signal and image processing, including speech enhancement. They consist of two distinct networks:

- *Encoder* N_θ, parameterized by θ, that maps an input signal $\mathbf{x} \in \mathbb{R}^T$ to a latent representation $\mathbf{z} \in \mathbb{R}^Z$:
$$\mathbf{z} = N_\theta(\mathbf{x}).$$
- *Decoder* N_ϕ, parameterized by ϕ, that reconstructs a target signal $\mathbf{y}$ from $\mathbf{z}$:
$$\hat{\mathbf{y}} = N_\phi(\mathbf{z}).$$

In the context of speech enhancement, $\mathbf{x}$ represents the noisy signal and $\mathbf{y}$ is the clean reference. UAEs typically optimize a reconstruction loss (e.g., $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$ or cross-entropy) to learn the mapping $\mathbf{x} \mapsto \hat{\mathbf{y}}$.

Probabilistic Perspective - Image Example. In image processing, the decoder's output can be modeled probabilistically, especially if pixel intensities are discrete and bounded. Following [14], one can treat each output neuron as a discretized Beta random variable:

$$p\big(y_k \mid \mathbf{x}, \theta, \phi\big) = \mathrm{Beta}\big(a_k^y; \alpha = 1 + y_k, \beta = 2 - y_k\big), \tag{2.12}$$

where a_k^y is the decoder's activation for the k^{th} output and $y_k \in [0, 1]$ is the normalized target. The negative log-likelihood simplifies to

$$-\ln p\big(y_k \mid \mathbf{x}, \theta, \phi\big) = (1 - y_k) \ln(1 - a_k^y) + y_k \ln a_k^y, \tag{2.13}$$

equivalent to a double cross-entropy. Although derived for image intensities, this perspective can also provide insights for normalized speech signals in $[0, 1]$.

Training Approach. Training a UAE involves a forward pass:

$$\mathbf{x} \;\mapsto\; \mathbf{z} = N_\theta(\mathbf{x}) \;\mapsto\; \hat{\mathbf{y}} = N_\phi(\mathbf{z}),$$

where $\mathbf{x}$ is fed through the encoder to produce $\mathbf{z}$, and the decoder reconstructs $\hat{\mathbf{y}}$. Parameters θ and ϕ are updated to minimize the total loss, often expressed as

$$-\ln \prod_k p\big(y_k \mid \mathbf{x}, \theta, \phi\big) = \sum_k -\ln p\big(y_k \mid \mathbf{x}, \theta, \phi\big). \tag{2.14}$$

B. Bidirectional Autoencoders (BAE). Unlike UAEs, BAEs employ a single network N_θ for both encoding and decoding, as illustrated in Fig. 4. This structure reduces parameter overhead and enhances computational efficiency by sharing weights:

$$\mathbf{z} = N_\theta(\mathbf{x}) \quad \text{(forward/encode)} \tag{2.15}$$

$$\hat{\mathbf{x}} = N_\theta^T(\mathbf{z}) \quad \text{(backward/decode)} \tag{2.16}$$

N_θ^T denotes the transposed weights of N_θ and the backward error in BAEs is defined by:

$$\mathcal{E}_M(\mathbf{x}, \theta) = -\frac{1}{M} \sum_{m=1}^{M} \sum_{i=1}^{I} \Big[x_i^{(m)} \ln a_i^{xb(m)} + \big(1 - x_i^{(m)}\big) \ln\big(1 - a_i^{xb(m)}\big) \Big], \tag{2.17}$$

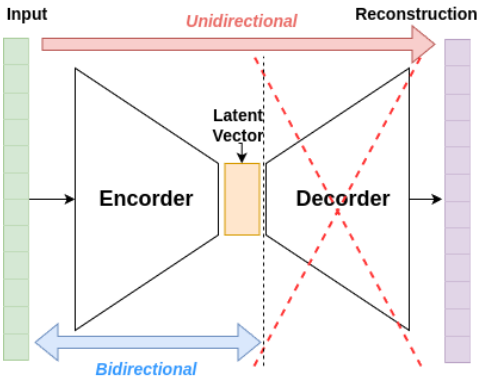


Figure 4. Comparison of Unidirectional (left) and Bidirectional (right) Autoencoders.

Algorithm 1 Bidirectional Training Procedure

Input: Dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})\}$, learning rate η, epochs M, batch size L.
Initialize: Network parameters $\theta^{(0)}$.
for epoch = 1 **to** M **do**
  for each mini-batch $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$ **do**
    $\mathbf{z} \leftarrow N_\theta(\mathbf{x})$ // forward
    $\hat{\mathbf{x}} \leftarrow N_\theta^T(\mathbf{z})$ // backward
    Compute backward error:

$$E_b(\theta) = -\frac{1}{L}\sum_{i=1}^{L}\Big[y_i \ln \hat{x}_i + (1-y_i)\ln(1-\hat{x}_i)\Big].$$

    $\theta \leftarrow \theta - \eta\,\nabla_\theta E_b(\theta)$
  end for
end for

where M is the number of samples, I the output neurons, $x_i^{(m)}$ the ground truth, and $a_i^{xb(m)}$ the backward-pass prediction for the i^{th} neuron in the m^{th} sample.

C. Integration of LFDWPT in BAEs. To further enhance speech processing, we embed LFDWPT into BAEs. This integration provides a multiscale decomposition that improves noise suppression and feature extraction:

- *Forward pass:* Uses learnable wavelet filters and thresholding for decomposition.
- *Backward pass:* Reconstructs the signal via transposed filters and upsampling.

By unifying wavelet-based denoising with a bidirectional architecture, the system can be trained end-to-end, preserving interpretability while increasing efficiency.

2.3. Recursive Encoding and Decoding via a Bidirectional Architecture

NWPA adopts a single network for both encoding (decomposition) and decoding (reconstruction), significantly reducing parameter usage. In the forward pass, the LFDWPT recursively decomposes the input x into approximation and detail coefficients. At each level l, the network computes:

$$a_{l+1} = \left(x * h^{(l)}\right) \downarrow 2 \qquad (2.18) \qquad\qquad d_{l+1} = \left(x * g^{(l)}\right) \downarrow 2 \qquad (2.19)$$

where $h^{(l)}$ and $g^{(l)}$ are the low-pass and high-pass filters, respectively. The detail coefficients d_{l+1} capture higher-frequency variations and are passed through the LAHT module (Section 2.4) to suppress noise while retaining essential features. This decomposition continues until a predefined number of levels L is reached, yielding a multi-resolution representation.

For reconstruction, the decoder employs the inverse wavelet packet transform (IDWPT) using transposed convolutions:

$$a_j[n] \;=\; \sum_{p=-\infty}^{+\infty} \bar{h}[\,n-2p\,]\,a_{j+1}[p] \;+\; \sum_{p=-\infty}^{+\infty} \bar{g}[\,n-2p\,]\,d_{j+1}[p], \qquad (2.20)$$

where $\bar{h}$ and $\bar{g}$ denote the conjugate (or transposed) versions of h and g. Thus, the approximation coefficients a_{j+1} (originally obtained via h) are convolved with $\bar{h}$, and the detail coefficients d_{j+1} (obtained via g) are convolved with $\bar{g}$. Since the same filters (or their

transposed counterparts) are used for both decomposition and synthesis, the bidirectional design ensures parameter efficiency while preserving accurate reconstruction.

2.4. **Learnable Asymmetric Hard Thresholding (LAHT)**

To further improve noise suppression, we incorporate an LAHT function applied to the detail coefficients at each decomposition level. Unlike conventional thresholding methods that rely on static thresholds, LAHT employs trainable parameters to adapt threshold values for each subband, as per eq. 2.21:

$$\mathcal{T}(d) = d \cdot \Big[\sigma\big(\alpha\,(d-\tau^{+})\big) + \sigma\big(-\beta\,(d+\tau^{-})\big)\Big], \tag{2.21}$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, τ^{+} and τ^{-} are learnable thresholds, and α and β control transition steepness. By selectively attenuating coefficients that likely represent noise, LAHT preserves vital signal details crucial for speech intelligibility. Its adaptive nature allows the network to handle diverse noise conditions more effectively than fixed-threshold approaches.

2.4.1. **Sparsity-Enforcing Loss Function (SELF)**

We adopt a two-term objective function that balances time-domain fidelity and wavelet-domain sparsity, ensuring both remain significant throughout training. In FDWPT with depth L, there are multiple "approximation-like" and "detail-like" sub-bands at the final decomposition step defined as follows:

$$\mathcal{S}_{\text{final}} = \Big\{\mathbf{a}^{(L,k)},\, \mathbf{d}^{(L,k)} \;\Big|\; k = 1,\ldots,2^{L}\Big\}, \tag{2.22}$$

denote the approximation and detail sub-bands exclusively at level L. We define $\text{Card}(\mathcal{S}_{\text{final}})$ as the total number of wavelet coefficients at that final layer. Let y be the clean target signal with $\text{Card}(y)$ samples. Our two-term loss is then:

$$\begin{aligned}\mathcal{L} = \;& \lambda \cdot \frac{1}{\text{Card}(y)}\Big\|y-\tilde{y}\Big\|_1 \;+\; \gamma \cdot \frac{1}{\text{Card}\big(\mathcal{S}_{\text{final}}\big)} \sum_{k=1}^{2^{L}} \Big(\|\mathbf{a}^{(L,k)}\|_1 + \|\mathbf{d}^{(L,k)}\|_1\Big) \\ & \text{subject to:}\quad 0 \;\le\; \lambda \;\le\; 1, \quad 0 \;\le\; \gamma \;\le\; 1, \quad 1 \;\le\; \lambda+\gamma \;\le\; 2.\end{aligned} \tag{2.23}$$

The first term, $\|y-\tilde{y}\|/\text{Card}(y)$, promotes accurate time-domain reconstruction, while the second term enforces ℓ_1-based sparsity specifically on the final FDWPT sub-bands. By restricting the penalty to $\mathbf{a}^{(L,k)}$ and $\mathbf{d}^{(L,k)}$, we concentrate on sparsifying the terminal layer's approximation-like and detail-like outputs, which effectively captures the signal's multi-resolution structure. We linearly schedule λ and γ across E epochs to transition smoothly between fidelity and sparsity:

$$\begin{aligned}\lambda_e =\; & \lambda_{\text{start}} + \Big(\lambda_{\text{end}} - \lambda_{\text{start}}\Big)\frac{e-1}{\max(1,\,E-1)}, \\ \gamma_e =\; & \gamma_{\text{start}} + \Big(\gamma_{\text{end}} - \gamma_{\text{start}}\Big)\frac{e-1}{\max(1,\,E-1)}, \\ & \text{with}\quad 1 \;\le\; \lambda_e+\gamma_e \;\le\; 2, \quad 0 \;\le\; \lambda_e, \gamma_e \;\le\; 1.\end{aligned} \tag{2.24}$$

This triangular constraint $0 \le \lambda, \gamma \le 1$ and $\lambda + \gamma \ge 1$ ensures neither reconstruction nor sparsity is overlooked. Early epochs ($\lambda \approx 1$) emphasize learning robust speech representations, while gradually increasing γ strengthens the ℓ_1 regularization on the final sub-bands, further refining denoising at multiple resolution scales. Consequently, this SELF formulation unifies wavelet packet decomposition, threshold-based denoising, and a flexible balance between time-domain accuracy and wavelet sparsity—suiting the broader aim of enhancing speech under diverse noise conditions.

2.5. Integration and End-to-End Training

All components of the NWPA framework—including the LFDWPT filters, LAHT, and SELF—are trained jointly in a single end-to-end network. During each forward pass, the wavelet filters adaptively decompose the input into multiple resolution levels, while LAHT applies trainable thresholds for noise suppression. The backward pass reconstructs the signal using transposed operations, ensuring gradient propagation flows through both the decomposition and reconstruction stages. This unified procedure updates not only the wavelet bases and thresholding parameters but also the model's reconstruction weights to maintain fidelity in the time domain. By avoiding separate or staged training for each module, the network learns cohesive representations of speech signals that effectively balance denoising and fidelity. Gradients are backpropagated end-to-end, allowing wavelet analysis, thresholding, and reconstruction to co-evolve for improved perceptual quality and speech intelligibility, as validated by our experiments.

3. Experiments and Results

3.1. Dataset

We conducted our experiments using the publicly available VoiceBank-DEMAND dataset [15], a widely recognized benchmark for evaluating speech enhancement methods. The dataset is divided into two subsets. The **Training Set** consists of 28 speakers and 11,572 utterances, each generated by mixing clean speech with various noise sources, while the **Test Set** has 2 speakers and 824 utterances specifically designed to assess generalization under unseen noisy conditions. All noise signals are taken from the DEMAND corpus [16] as well as speech-shaped and babble noise. During training, noise from 10 different sources is used, whereas the test set incorporates 5 unique DEMAND noises, creating a challenging evaluation framework for speech enhancement models.

3.2. Experimental Setup

To standardize processing and reduce computational overhead, all audio signals were resampled to 16 kHz. We followed the standard protocol in [15] for splitting the dataset into training and testing subsets. Our proposed NWPA model was trained for 100 epochs with a batch size of 64 using the Adam optimizer at a learning rate of 1×10^{-4}. The weighting factors λ and γ that balance reconstruction fidelity against sparsity constraints were gradually updated during training: λ decreased from 1.0 to 0.8, and γ increased from 0.5 to 1.0. The network architecture employs 15-layer LFDWPT blocks with a kernel size of 40, designed to emulate the Daubechies-20 wavelet family [17], which is often used for multiresolution signal analysis. We trained the model on NVIDIA GPUs, validating performance at the end of each epoch to monitor convergence. Hyperparameters were tuned by minimizing the validation loss, ensuring that the trained model generalized effectively to previously unseen conditions. The final evaluation was conducted on the designated test set.

3.3. Metrics

We evaluated the model's performance using five well-established metrics commonly employed in speech enhancement studies:

(1) *Perceptual Evaluation of Speech Quality (PESQ)* [18], which measures the perceptual quality of enhanced speech relative to a clean reference.
(2) *Short-Time Objective Intelligibility (STOI)* [19], which captures how intelligible the enhanced signal is, especially in noisy environments.

Model	COVL	CBAK	CSIG	STOI	PESQ
Noisy	2.63	2.44	3.35	0.91	1.97
MUSE [21]	4.10	3.80	4.63	0.95	3.37
MetricGAN+ [22]	3.64	3.16	4.14	-	3.15
SEMamba [23]	4.26	3.98	4.75	0.96	3.52
TSTNN [24]	3.67	3.53	4.33	0.95	2.96
DPCFCS-Net [25]	4.15	3.88	4.71	0.96	3.42
CMGAN [26]	4.12	3.94	4.63	0.96	3.41
S4ND U-Net [27]	3.85	3.62	4.52	-	3.15
Mamba-SEUNet [28]	4.32	4.02	4.80	0.96	3.59
MP-SENet [29]	4.22	3.95	4.73	0.96	3.50
NWPA (Ours)	**4.40**	**4.20**	**4.85**	**0.97**	**3.78**

Table 1. Comparison of NWPA with state-of-the-art models on the VoiceBank-DEMAND dataset across COVL, CBAK, CSIG, STOI, and PESQ metrics.

(3) The *Mean Opinion Score (MOS) predictor of signal distortion (CSIG)* [20], gauging the level of distortion perceived in the enhanced signal.
(4) The MOS predictor of background noise intrusiveness (CBAK) [20], quantifying how much residual noise affects perceived quality.
(5) The MOS predictor for overall signal quality (COVL) [20], providing a comprehensive measure of the enhancement's subjective quality.

Collectively, these metrics offer a balanced assessment of both perceptual and objective dimensions of speech enhancement.

3.4. **Results and Comparison with State-of-the-Art**

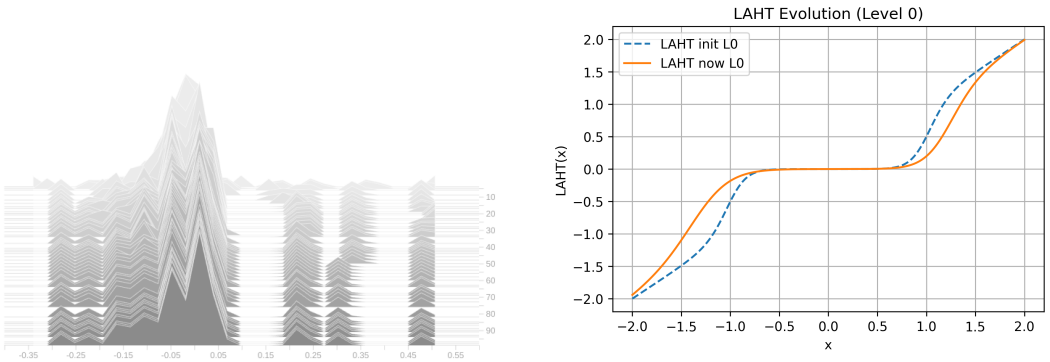
Table 1 summarizes the performance of our method, NWPA, compared to several state-of-the-art speech enhancement methods. All results are evaluated on the VoiceBank-DEMAND dataset, capturing a range of perceptual and intelligibility metrics. The Noisy baseline shows the scores before any enhancement, highlighting how background noise degrades speech quality.

Overall Quality (COVL). NWPA achieves a COVL of 4.40, exceeding the previously highest score of 4.32 by Mamba-SEUNet. This underscores the balanced improvement across multiple dimensions of speech quality, including clarity, distortion, and noise suppression. The higher COVL value indicates that NWPA yields a more natural and less fatiguing listening experience.

Background Noise Intrusiveness (CBAK). For CBAK, NWPA attains 4.20, outperforming Mamba-SEUNet's 4.02 and SEMamba's 3.98. This metric is particularly sensitive to how much residual noise remains noticeable in the enhanced speech. NWPA's superior CBAK suggests that our approach more effectively suppresses intrusive background noise, making the resulting signal more pleasant and less distracting for listeners.

Signal Distortion (CSIG). NWPA obtains a CSIG of 4.85, surpassing Mamba-SEUNet's 4.80 and SEMamba's 4.75. A higher CSIG indicates lower perceived distortion in the enhanced speech signal. NWPA's result confirms that our method retains the critical speech components while removing artifacts and interference introduced by the noise suppression process.

Speech Intelligibility (STOI). With an STOI score of 0.97, NWPA demonstrates a measurable boost in intelligibility compared to top-performing models such as CMGAN and SEMamba (both at 0.96). Maintaining or improving intelligibility under noisy conditions is crucial for real-world applications like hearing aids, teleconferencing, and automatic speech recognition systems. NWPA's higher STOI signals its ability to preserve essential speech cues required for comprehension.



(a) Wavelet Filter Evolution *(b)* First Level LAHT Evolution

Figure 5. Training evolution of learnable wavelet filters and LAHT.

Perceptual Quality (PESQ). PESQ is often considered one of the most critical measures for perceptual evaluation. NWPA achieves 3.78, outperforming Mamba-SEUNet's 3.59 and SEMamba's 3.52. This improvement highlights NWPA's effectiveness in reducing noise without introducing warping, tonal artifacts, or excessive attenuation.
Discussion. NWPA consistently outperforms other state-of-the-art (SOTA) methods across multiple dimensions, including the challenging measures of background noise intrusiveness (CBAK) and signal distortion (CSIG). These gains can be attributed to the joint use of:

- *Learnable wavelet packet filters*, which adapt the time-frequency basis to speech signals more effectively than fixed wavelets.
- *Bidirectional autoencoder design*, reducing parameter redundancy and ensuring coherent reconstruction across decomposition and synthesis.
- *Learnable Asymmetric Hard Thresholding (LAHT)*, which leverages subband-specific thresholds to suppress noise selectively, preserving crucial speech details.
- *Sparsity-Enforcing Loss Function (SELF)*, balancing time-domain reconstruction with wavelet-domain sparsity to eliminate extraneous coefficients and effectively handle diverse noise conditions.

By integrating these components into an end-to-end trainable framework, NWPA avoids segmented optimization of each module, thereby achieving greater synergy among decomposition, thresholding, and reconstruction. NWPA's scores imply that it not only improves perceptual quality but also maintains critical intelligibility and acoustic fidelity, making it a robust and practical solution for real-world speech enhancement tasks.
Training Dynamics. Figures 5a and 5b offer a closer look at how the learnable wavelet filters and LAHT thresholds evolve over the course of training. In Figure 5a, the wavelet filters gradually transition from broad, generic shapes to more refined, feature-specific kernels that better capture speech structures. Meanwhile, Figure 5b shows how the LAHT thresholds dynamically adjust, striking an optimal balance between noise suppression and preservation of key speech components. Observing these adaptations provides insight into NWPA's end-to-end learning process: the wavelet filters and threshold parameters co-evolve to enhance speech intelligibility and reduce artifacts, ultimately contributing to NWPA's good performance.

4. **Conclusion**

We presented NWPA, a novel framework that unifies LFDWPT with bidirectional autoencoding for robust speech enhancement. By enabling adaptive multi-resolution analysis

through learnable wavelet packet filters and refining noise suppression via LAHT, NWPA effectively preserves critical speech cues while mitigating noise artifacts. SELF encourages compact representations in the wavelet domain, promoting both efficiency and intelligibility across multiple frequency bands. Moreover, the bidirectional architecture consolidates encoding and decoding processes, reducing memory overhead and complexity. Extensive evaluations on the VoiceBank-DEMAND dataset verified NWPA's state-of-the-art performance, outperforming existing methods in key metrics such as PESQ, STOI, CSIG, CBAK, and COVL. These results underline NWPA's capacity to not only remove background noise but also maintain essential speech characteristics under diverse and challenging conditions. Future work may explore applying NWPA to other audio processing tasks that require fine-grained time-frequency analysis, such as music denoising, acoustic scene classification, or low-resource speech recognition. Additionally, investigating more advanced wavelet packet decompositions (e.g., with adaptive level selection) and hardware-optimized implementations (e.g., for mobile or embedded devices) could further widen NWPA's real-world impact. From a broader perspective, NWPA's end-to-end learning process, grounded in adaptive wavelet decomposition and bidirectional autoencoding, offers a powerful and scalable approach to speech enhancement in noisy environments.

References

[1] D. Wang and J. Chen. "Supervised speech separation based on deep learning: An overview". In: *IEEE/ACM transactions on audio, speech, and language processing* 26.10 (2018), pp. 1702–1726.

[2] Y. Luo and N. Mesgarani. "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation". In: *IEEE/ACM transactions on audio, speech, and language processing* 27.8 (2019), pp. 1256–1266.

[3] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde. "Singing voice separation with deep u-net convolutional networks". In: (2017).

[4] S. Pascual, A. Bonafonte, and J. Serrà. "SEGAN: Speech Enhancement Generative Adversarial Network". In: *Interspeech 2017*. 2017, pp. 3642–3646. DOI: 10.21437/Interspeech.2017-1428.

[5] B. Nasih and P. D. Assitant. "Application of Wavelet Transform and Its Advantages Compared To Fourier Transform". In: (2016).

[6] W. Ha, C. Singh, F. Lanusse, E. Song, S. Dang, K. He, S. Upadhyayula, and B. Yu. "Adaptive wavelet distillation from neural networks through interpretations". In: (2021), pp. 20669–20682.

[7] G. Frusque and O. Fink. "Learnable Wavelet Packet Transform for Data-Adapted Spectrograms". In: *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2022), pp. 3119–3123. DOI: 10.1109/ICASSP43922.2022.9747491.

[8] A. Nfissi, W. Bouachir, N. Bouguila, and B. Mishara. "Deep Multiresolution Wavelet Transform for Speech Emotion Assessment of High-Risk Suicide Callers". In: *IAPR Workshop on Artificial Neural Networks in Pattern Recognition*. Springer. 2024, pp. 256–268.

[9] S. Nagar, A. Kumar, and M. Swamy. "Orthogonal Features-based EEG Signal Denoising using Fractionally Compressed AutoEncoder". In: *Signal Process.* 188 (2021), p. 108225. DOI: 10.1016/j.sigpro.2021.108225.

[10] G. S. Martín, E. L. Droguett, V. Meruane, and M. das Chagas Moura. "Deep variational auto-encoders: A promising tool for dimensionality reduction and ball bearing elements fault diagnosis". In: *Structural Health Monitoring* 18 (2018), pp. 1092 –1128. DOI: 10.1177/1475921718788299.

[11] O. Adigun and B. Kosko. "Bidirectional Backpropagation Autoencoding Networks for Image Compression and Denoising". In: *2023 International Conference on Machine Learning and Applications (ICMLA)* (2023), pp. 730–737. DOI: 10.1109/ICMLA58977.2023.00107.

[12] P. Busch, T. Heinonen, and P. Lahti. "Heisenberg's uncertainty principle". In: *Physics reports* 452.6 (2007), pp. 155–176.

[13] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier Science, 2008. ISBN: 9780080922027. URL: https://books.google.ca/books?id=5qzeLJljuLoC.

[14] Z. Ma and A. Leijon. "Beta mixture models and the application to image classification". In: *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE. 2009, pp. 2045–2048.

[15] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi. "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech." In: *SSW*. 2016, pp. 146–152.

[16] C. Veaux, J. Yamagishi, and S. King. "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database". In: *2013 international conference oriental COCOSDA held jointly with 2013 conference on Asian spoken language research and evaluation (O-COCOSDA/CASLRE)*. IEEE. 2013, pp. 1–4.

[17] S. Mallat. "A theory for multiresolution signal decomposition: the wavelet representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11.7 (1989), pp. 674–693.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra. "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs". In: *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*. Vol. 2. IEEE. 2001, pp. 749–752.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. "A short-time objective intelligibility measure for time-frequency weighted noisy speech". In: *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2010, pp. 4214–4217.

[20] Y. Hu and P. C. Loizou. "Evaluation of objective quality measures for speech enhancement". In: *IEEE Transactions on audio, speech, and language processing* 16.1 (2007), pp. 229–238.

[21] Z. Lin, X. Chen, and J. Wang. "MUSE: Flexible Voiceprint Receptive Fields and Multi-Path Fusion Enhanced Taylor Transformer for U-Net-based Speech Enhancement". In: *Interspeech 2024*. 2024, pp. 672–676. DOI: 10.21437/Interspeech.2024-1017.

[22] S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao. "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement". In: *Interspeech 2021*. 2021, pp. 201–205. DOI: 10.21437/Interspeech.2021-599.

[23] R. Chao, W.-H. Cheng, M. La Quatra, S. M. Siniscalchi, C.-H. H. Yang, S.-W. Fu, and Y. Tsao. "An Investigation of Incorporating Mamba for Speech Enhancement". In: *arXiv preprint arXiv:2405.06573* (2024).

[24] K. Wang, B. He, and W.-P. Zhu. "TSTNN: Two-stage transformer based neural network for speech enhancement in the time domain". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 7098–7102.

[25] J. Wang. "Efficient Encoder-Decoder and Dual-Path Conformer for Comprehensive Feature Learning in Speech Enhancement". In: *Interspeech 2023*. 2023, pp. 2853–2857. DOI: 10.21437/Interspeech.2023-815.

[26] R. Cao, S. Abdulatif, and B. Yang. "CMGAN: Conformer-based Metric GAN for Speech Enhancement". In: *Interspeech 2022*. 2022, pp. 936–940. DOI: 10.21437/Interspeech.2022-517.

[27] P.-J. Ku, C.-H. H. Yang, S. Siniscalchi, and C.-H. Lee. "A Multi-dimensional Deep Structured State Space Approach to Speech Enhancement Using Small-footprint Models". In: *Interspeech 2023*. 2023, pp. 2453–2457. DOI: 10.21437/Interspeech.2023-1084.

[28] J. Wang, Z. Lin, T. Wang, M. Ge, L. Wang, and J. Dang. "Mamba-SEUNet: Mamba UNet for Monaural Speech Enhancement". In: *arXiv preprint arXiv:2412.16626* (2024).

[29] Y.-X. Lu, Y. Ai, and Z.-H. Ling. "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra". In: *Interspeech 2023*. 2023, pp. 3834–3838. DOI: 10.21437/Interspeech.2023-1441.