Internet of Things 39 (2026) 102046

Contents lists available at ScienceDirect

Internet of Things

journal homepage: www.elsevier.com/locate/iot

Research article

Lightweight streaming speech enhancement for AIoT-enabled wearable hearing aids using parallel spiking mamba

Junkang Yang [a], Hiromitsu Nishizaki [a,*], Chee Siang Leow [a], Hongqing Liu [b], Lu Gan [c]

[a] Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Kofu, 400-8511, Yamanashi, Japan
[b] Chongqing Key Lab of Mobile Communications Technology, Chongqing University of Posts and Telecommunications, Chongqing, 400065, China
[c] College of Engineering, Design and Physical Science, Brunel University, London, UB8 3PH, United Kingdom

ARTICLE INFO

Keywords:
Speech enhancement
Spiking neural networks
State space models
Hearing aid
Real-time processing

ABSTRACT

Deploying deep learning-based speech enhancement in wearable hearing aids is challenging due to strict constraints on latency and real-time processing. Conventional models are often computationally intensive and unsuitable for resource-limited edge devices. To address this, we propose a lightweight and end-to-end streaming speech enhancement framework based on a parallel spiking Mamba architecture. The proposed model combines the computation-efficient, event-driven properties of spiking neural networks with the long-range sequence-modeling capabilities of state-space models. A parallel design enables simultaneous extraction of local acoustic features and global temporal dependencies, supporting efficient streaming inference. Experimental results show that the proposed method achieves 2.73 PESQ and 93.9% STOI in the streaming setting on the VoiceBank-DEMAND dataset, using only 0.26 M parameters and 0.11 G/s MACs. It also achieves an end-to-end algorithmic latency of 30 ms under the default configuration, demonstrating the potential of neuromorphic-inspired approaches for real-time speech processing in AIoT-enabled hearing aids.

1. Introduction

Speech enhancement is a critical front-end task that improves the intelligibility and quality of speech signals corrupted by background noise, with wide applications in real-time communication systems such as hearing aids [1], voice-controlled devices [2,3], and automatic speech recognition (ASR) front-ends [4]. Recent advances in deep learning, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer-based models, have significantly improved speech enhancement performance [5]. In particular, time-domain approaches such as Conv-TasNet [6], and its successors have achieved state-of-the-art results by directly modelling waveform signals.

However, deploying these models in AIoT-enabled wearable devices remains challenging due to strict constraints on latency, power consumption, and memory. Hearing aids and similar edge devices require efficient on-device processing to support real-time user experience, while conventional deep neural networks are often computationally intensive and unsuitable for resource-constrained environments.

* Corresponding author.
E-mail address: hnishi@yamanashi.ac.jp (H. Nishizaki).

https://doi.org/10.1016/j.iot.2026.102046
Received 3 April 2026; Received in revised form 23 June 2026; Accepted 27 July 2026
Available online 28 July 2026

Despite considerable progress, existing deep learning models for speech enhancement still face limitations in edge-oriented scenarios. Traditional CNNs and RNNs struggle to efficiently capture long-range temporal dependencies, often requiring large receptive fields that increase computational costs. Transformer-based models achieve strong sequence modelling capability but suffer from quadratic complexity with respect to input length, making them impractical for real-time streaming applications on low-power devices.

Recently, spiking neural networks (SNNs) [7] have emerged as a promising solution for energy-efficient computation due to their event-driven and sparse processing paradigm, which aligns well with neuromorphic and low-power edge hardware. However, SNNs alone often underperform on complex acoustic tasks due to their limited ability to model rich temporal dynamics. Meanwhile, state space models (SSMs), particularly the structured state space model S4 [8] and its efficient variant Mamba [9], enable long-range sequence modelling with near-linear complexity. Nevertheless, their direct application to streaming audio remains challenging due to the computational burden of processing dense waveform inputs.

From an application perspective, hearing aids represent a typical cyber-physical-human system in AIoT environments, where human auditory perception, physical acoustic signals, and intelligent processing are tightly coupled. According to the World Health Organisation [10], over 5% of the world's population, or 430 million people, require rehabilitation for disabling hearing loss, and this number is expected to grow significantly in the future. Although modern hearing aids have made progress in amplification and basic processing, limited speech intelligibility in noisy environments remains a major challenge, highlighting the need for efficient, real-time, and personalised speech enhancement solutions.

To address these challenges, we propose a lightweight and streaming speech enhancement framework, named Parallel Spiking Mamba for Speech Enhancement (PSMSE). The proposed approach integrates the energy-efficient properties of SNNs with Mamba's long-range sequence modelling capability within a unified architecture. Specifically, a parallel design is introduced, where a Spiking Mamba Convolution (SMC) module captures local temporal features, while a Spiking Mamba Recurrent (SMR) module models global dependencies. A key innovation is that the Mamba component operates directly on spike-based representations, significantly reducing sequence complexity compared to raw waveform processing. Designed with strict causality, the model enables low-latency streaming inference suitable for real-time applications.

Furthermore, targeting practical hearing aid systems, we incorporate a personalised fusion module that integrates user-specific auditory characteristics, such as audiograms, enabling adaptive, human-centred speech enhancement.

The main contributions of this work are summarised as follows:

- We propose PSMSE, a causal streaming speech enhancement framework whose main architectural novelty is the parallel integration of SMC and SMR modules for jointly modelling local acoustic features and global temporal dependencies.
- We design spike-based Mamba processing blocks, where PLIF neurons and sparse spike representations are incorporated into Mamba-based sequence modelling to reduce computational cost for real-time edge inference.
- We introduce audiogram-guided personalisation as an application-oriented extension for hearing aids, enabling the model to adapt enhancement behaviour to user-specific auditory characteristics.

The rest of this paper is organised as follows. Section 2 introduces the related work on speech enhancement and its application to hearing aids. Section 3 explains the foundations of SNNs and state space models. Section 4 details the proposed model's architecture. Section 5 presents experimental settings, followed by results and discussions in Section 6. Finally, Section 7 concludes the paper.

2. Related work

2.1. Speech enhancement (SE)

Speech enhancement addresses the problem of recovering a clean speech signal from its noisy observation. The fundamental acoustic model is expressed as:

$$y(t) = x(t) + n(t), \tag{1}$$

where $y(t)$ denotes the observed noisy speech signal, $x(t)$ represents the original clean speech signal and $n(t)$ is the additive noise. The mathematical objective is to derive an optimal estimator $\hat{x}(t)$ that minimises the distortion between the estimated and true clean speech signals. This is commonly formulated as a statistical estimation problem:

$$\hat{x}(t) = \arg\min_{\hat{x}} \mathbb{E}[\mathcal{L}(x(t), \hat{x}(t))] \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes a regression loss function.

Conventional speech enhancement methods commonly rely on time-frequency representations, which may introduce information loss during spectral decomposition and constrain the joint modeling of waveform structures. This limitation caused the emergence of end-to-end time-domain models. Conv-TasNet's encoder-separator-decoder architecture [6] does not need spectral decomposition, setting new benchmarks through learnable filters and temporal convolution. Subsequent innovations, such as DPRNN [11], split sequences into intra- and inter-block paths to capture both local and global dependencies. DCCRN [12] combined complex spectral operations with temporal modules to enhance the performance. But the cost is parameter growth, which is incompatible with edge constraints.

Recent years have witnessed a pattern shift in speech enhancement research toward computationally efficient architectures capable of real-time edge deployment. Transformer [13], and its variants further improved performance through self-attention mechanisms,

Table 1
Architectural comparison of representative spiking, Mamba-based, and streaming speech enhancement models in related work.

Model	Feature	Brief Description on Network Structure
SpikingS4 [21]	Spectrogram	An SNN-based encoder-spiking S4 layer stack-decoder network with structured state-space layers and LIF neurons.
DPSNN [22]	Waveform	A time-domain encoder-separator-decoder SNN that uses CNN and RNN modules to generate masks.
Spiking FullSubNet [20]	Spectrogram	An SNN-based full-band/sub-band fusion network with auditory-inspired frequency partitioning and adaptive spiking neurons.
MambaSEUNet [23]	Spectrogram	A U-Net-style time-frequency network that uses bidirectional TS-Mamba blocks with skip connections.
DeepFilterNet [24]	Spectrogram	A two-stage encoder-decoder framework that first applies ERB-scaled spectral gains and then uses a low-frequency deep filtering network.
FSPEN [15]	Spectrogram	A full-band/sub-band encoder-dual-path enhancer-decoder network.
LiSenNet [17]	Spectrogram	A lightweight encoder-decoder network with sub-band down/up-sampling, dual-path recurrent modelling, phase refinement, and optional noise detection.
UL-UNAS [18]	Spectrogram	A U-Net model whose efficient convolutional blocks, APReLU, cTFA, and overall architecture are optimised by MACS-aware neural architecture search.

with architectures like CleanUNet2 [14] achieving state-of-the-art results by modelling full-sequence dependencies. However, their $O(N^2)$ complexity remains prohibitive for streaming applications. To further simplify the network, several networks with ultra-low parameter counts have been proposed, such as FSPEN [15], GTCRN [16], LiSenNet [17], and UL-UNAS [18]. The number of parameters in these networks can be 23.7K to 79K, but their performance tends to degrade due to the limited capacity of their simplified designs. At the same time, some research explored SNNs as low-power alternatives. Recent breakthroughs, such as spiking transformer [19], have integrated spiking units with transformers to mitigate gradient issues via surrogate gradients. Spiking full-subnet [20] enhanced the performance of SNN on speech enhancement with a full-band and sub-band fusion approach. SpikingS4 [21] and DPSNN [22] also provided a light-weight solution for this issue. On the other hand, state space models (SSMs) [8], particularly mamba [9], achieve near-linear-time long-sequence modelling through input-dependent selective state transitions. Mamba's performance in language modelling has inspired related works [23], yet its application to raw waveforms remains limited by computational demands while processing dense, high-dimensional audio streams.

Table 1 summarises the differences of representative related models, highlighting the distinct design choices in input features and network structures.

2.2. SE in hearing aid

Compared to general speech enhancement, deploying SE algorithms in hearing aids (HAs) imposes constraints that are significantly stricter than general telecommunication scenarios. The primary challenge is low computational latency. This requirement mandates strictly causal processing, precluding the use of future frames for context modelling. HAs operate under extreme power and resource constraints. Powered by miniature batteries, the digital signal processor is restricted to a milliwatt-level budget. While large-scale DNNs offer superior denoising performance, they are computationally prohibitive for edge devices. Although lightweight architectures like DeepFilterNet [24] and model compression techniques have been proposed, they often compromise noise suppression performance in complex, non-stationary environments to meet hardware constraints.

Unlike universal SE tasks, HA processing must be personalised. Hearing impairment involves frequency-specific threshold shifts and reduced dynamic range. Consequently, SE strategies must align with fitting formulas (e.g., NAL-R [25]) to ensure the enhanced signal remains audible and intelligible for the specific user. Traditional pipelines typically treat noise reduction and hearing loss compensation (HLC) as separate, cascaded stages. However, this decoupling often leads to suboptimal interaction between modules and accumulated bias. Recent works have demonstrated the feasibility of joint optimisation, where denoising and amplification are performed within a unified framework [26,27]. Such end-to-end approaches have shown satisfactory intelligibility while effectively reducing the cumulative latency inherent in multi-stage processing. Despite this progress, efficiently fusing physiological priors (e.g., audiograms) into networks for personalised enhancement remains an active area of research.

To summarise, few works currently fully resolve the triple constraints of streaming, lightweight computation, and personalisable enhancement, which inspires our work to bridge these constraints by synergising mamba and SNNs.

3. Preliminaries

3.1. Selective state space model

Mamba [9] fundamentally rethinks sequence modelling through continuous system theory, implementing a selection mechanism that dynamically modulates state transitions based on input characteristics. The system is governed by linear time-invariant dynamics in continuous space:

$$\frac{d\mathbf{h}(t)}{dt} = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t), \tag{3}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t), \tag{4}$$

where $\mathbf{h}(t) \in \mathbb{R}^N$ represents the hidden state, $\mathbf{x}(t) \in \mathbb{R}$ is the input signal (e.g., audio waveform), $\mathbf{y}(t) \in \mathbb{R}$ is the output, and $\mathbf{A}$ is the system evolution matrix. The critical innovation lies in transforming static matrices $\mathbf{B}$ and $\mathbf{C}$ into input-dependent projections $\mathbf{B}(\mathbf{x}) = \text{Linear}_B(\mathbf{x})$ and $\mathbf{C}(\mathbf{x}) = \text{Linear}_C(\mathbf{x})$. This is also called the selection mechanism of mamba.

To integrate SSMs into deep learning architecture, the discretisation process of continuous-time SSMs is applied in advance. Specifically, a timescale parameter Δ controls the step size of the discretisation process. The discrete-time implementation leverages zero-order hold (ZOH) discretisation:

$$\overline{\mathbf{A}} = \exp(\Delta\mathbf{A}), \tag{5}$$

$$\overline{\mathbf{B}} = (\Delta\mathbf{A})^{-1}(\exp(\Delta\mathbf{A}) - \mathbf{I}) \cdot \Delta\mathbf{B}. \tag{6}$$

So the discretized version of Eqs. (3) and (4) can be described as:

$$\mathbf{h}_t = \overline{\mathbf{A}}\mathbf{h}_{t-1} + \overline{\mathbf{B}}\mathbf{x}_t, \tag{7}$$

$$\mathbf{y}_t = \mathbf{C}(\mathbf{x}_t)\mathbf{h}_t. \tag{8}$$

3.2. Spiking neural networks

SNNs are computational models inspired by the dynamic properties of biological neurons, characterised by their use of discrete spike signals for information transmission instead of continuous activations found in traditional artificial neural networks (ANNs). SNN neurons possess internal states, typically represented as membrane potentials $U[t]$, which evolve over time under the influence of input spikes and previous states, following an iterative dynamical mechanism. If the membrane potential exceeds a threshold, the neuron emits a spike and subsequently resets its potential. This sparse, asynchronous communication enables significant energy-efficiency advantages.

Commonly used neuron models in SNNs include the biologically detailed Hodgkin-Huxley model [28] and the computationally efficient Leaky-Integrate-and-Fire (LIF) model [29], with the latter being more prevalent in deep learning. The state update process of the LIF neuron at each time step t can be formalised into three consecutive stages: charging, firing, and resetting. First, during the charging stage, the neuron's membrane potential $U[t]$ is updated based on the potential at the previous time step $U[t-1]$ and the current input $X[t]$. For a standard LIF neuron, this dynamic process is governed by the state equation:

$$U[t] = U[t-1] + \tau \cdot (X[t] - (U[t-1] - V_{th})), \tag{9}$$

where V_{th} represents the resting potential and τ is the decay factor controlling the leakage rate of the membrane potential. In this model, both τ and V_{th} are typically initialised as preset constants.

Following the charging stage is the firing stage, in which the neuron determines whether to generate an output spike by comparing the current membrane potential to a threshold. The output $S[t]$ is defined by the Heaviside step function $\Theta(\cdot)$, expressed as:

$$S[t] = \Theta(U[t] - V_{th}). \tag{10}$$

If the membrane potential $U[t]$ exceeds the threshold V_{th} (i.e., $U[t] - V_{th} \geq 0$), the function outputs 1, signifying a spike event; otherwise, it outputs 0.

After determining the output state, the neuron enters the resetting stage. This stage simulates the refractory period of biological neurons. It is the process by which the membrane potential of a neuron rapidly repolarises after firing an action potential [30]. Specifically, if a spike is fired ($S[t] = 1$), the membrane potential is forcibly reset to V_{reset}; if no spike occurs, the potential retains its state from the charging phase. This hard reset mechanism is formulated as:

$$U'[t] = (1 - S[t]) \cdot U[t] + S[t] \cdot V_{reset}, \tag{11}$$

where $U'[t]$ is the membrane potential after updating.

However, because the Heaviside function is nondifferentiable at the firing stage, directly applying gradient-based backpropagation is intractable. Consequently, a sigmoid function is commonly employed as a surrogate gradient to approximate the derivative during practical training.

In this study, to further enhance the model's adaptability to complex features, we introduce the Parametric Leaky Integrate-and-Fire (PLIF) neuron [31] as an extension of the LIF model. Unlike the traditional LIF model, which utilises a fixed decay factor, the PLIF neuron parameterises τ as a trainable variable. Specifically, the decay factor is defined as a function of a learnable parameter a, given by:

$$\tau = \frac{1}{1 + e^{-a}}. \tag{12}$$

This design not only ensures that the value of τ remains within a reasonable range $(0, 1)$, but also allows the network to adaptively optimise the membrane time constant during training. In terms of network architecture, neurons within the same layer share the same τ parameter, aligning with biological observations that neighbouring neurons exhibit similar electrophysiological properties. Conversely, neurons across different layers possess independent τ values, thereby empowering the network to capture diverse phase and frequency response characteristics at different hierarchical levels.

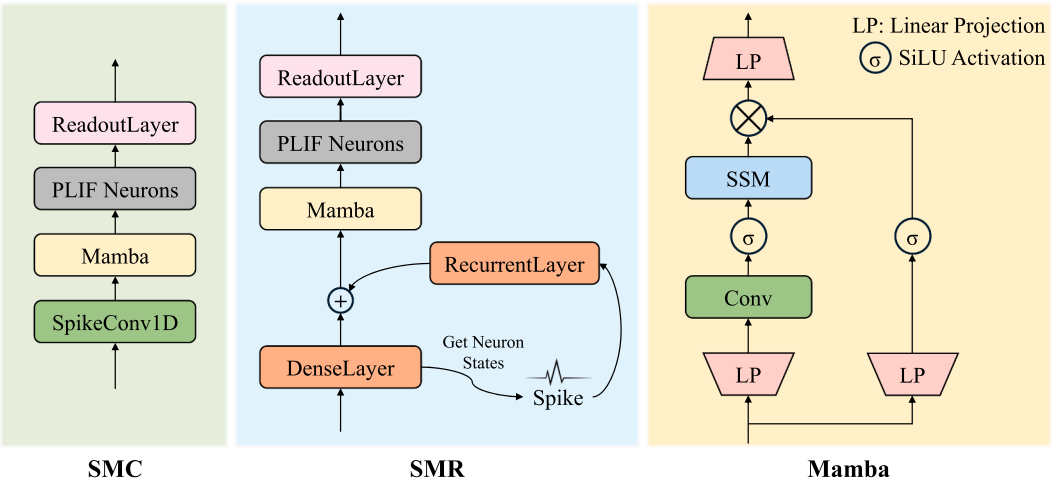

Fig. 1. Architecture of the proposed SMC (left), SMR (middle) module, and mamba (right).

4. Proposed method

4.1. Spiking mamba convolution (SMC) module

The structure of the proposed SMC module is shown in Fig. 1 (left). This is an innovative architecture designed to provide an efficient and effective solution for processing sequence data by combining SNNs with mamba's SSM. The module consists of several layers, including 1D convolution, mamba, PLIF neuron layers, and the readout layer.

In the SMC module, the data flow is as follows: first, the input data is fed into the 1D convolution layer (Conv1D). The main task of this layer is to perform convolution operations on the input data to extract local temporal features. The convolution layer processes the input signal using a sliding window, generating feature maps that capture the signal's local temporal information. Next, the output from the convolution layer is passed to the mamba layer, whose details are depicted in Fig. 1 (right). In the mamba layer, the input-dependent selective state transition mechanism is applied to handle sequence data. The mamba module dynamically adjusts state transitions using a structured state-space model to capture temporal data. Compared to traditional transformers, mamba can handle long sequences more efficiently and reduce computational complexity. Then the data processed by the mamba layer is passed to the PLIF neuron layer. In this layer, the PLIF neuron model is used to process the temporal signal. The PLIF neurons update their membrane potential based on the input (determined by the output from the previous layer). If the membrane potential exceeds a pre-set threshold, the neuron fires a spike. This process not only preserves temporal dependencies but also increases computational efficiency through its sparse computation properties, reducing unnecessary overhead. Finally, the output of the PLIF neuron layer is passed to the readout layer, whose structure follows that of the previous work [32]. Its task is to decode the spiking signals processed in the previous layers and convert them into the final output. This layer aggregates spiking signals from the neuron layer to generate outputs for subsequent tasks.

4.2. Spiking mamba recurrent (SMR) module

The SMR module is another key component proposed in our network, and its architecture is shown in Fig. 1 (middle). The SMR module aims to address both local and global temporal modelling of streaming data. It processes sequential inputs efficiently, capturing both short-term and long-term dependencies.

Initially, the input feature passes through a dense layer that prepares the data for further processing. This layer's role is to transform the raw input into a more suitable form for deeper analysis. The output of the dense layer will be divided into two branches, one for combining with the output of the recurrent layer and another for decoding into spikes, which are then fed into the recurrent layer.

The spike data is then processed by the recurrent layer, which is crucial for capturing contextual temporal dependencies within the sequence. This layer is specifically designed to handle recurrent interactions, a process vital for managing the global context in sequential data like speech signals. During this stage, neuron states are continuously updated, where each neuron's state reflects its temporal behaviour in the sequence. These states are calculated iteratively based on the previous neuron states and incoming data, ensuring the model maintains a consistent understanding of temporal dynamics.

The recurrent layer processes the data, updates the neuron states, and then generates spikes that represent critical events or thresholds within the sequence. This step ensures the model's sparsity by focusing only on significant data points. After the data has passed through the recurrent layer and PLIF neuron states have been updated, it flows to the readout Layer, where the spiking signals are decoded into the final output. Similarly, this readout layer aggregates these spikes and translates them back into usable data, making them suitable for subsequent tasks.

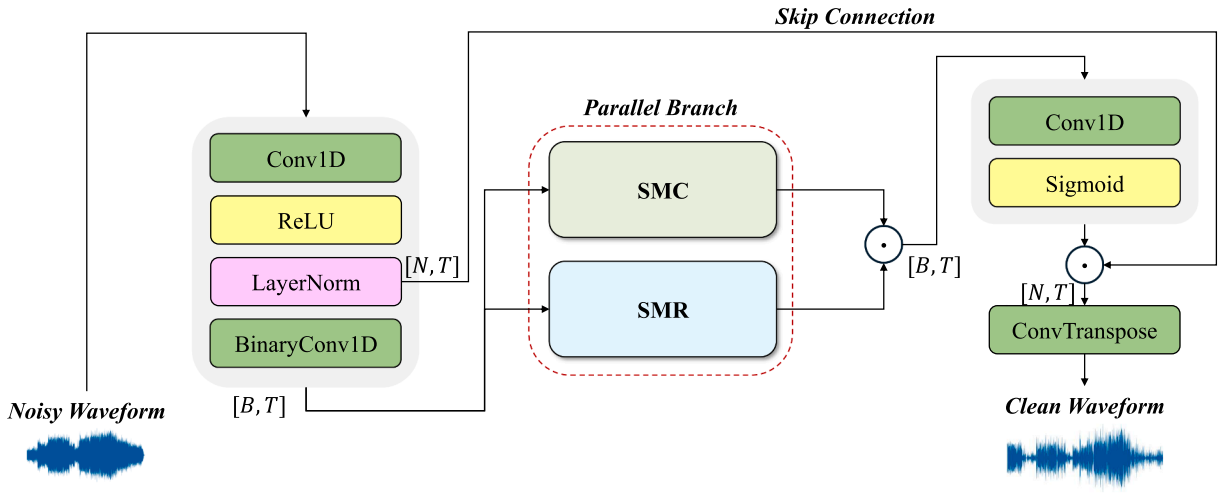

Fig. 2. Overall architecture of proposed PSMSE network.

4.3. Overall network architecture & implementation details

PSMSE is an end-to-end speech enhancement model based on spiking neural networks, integrating spatio-temporal feature extraction with spike dynamics. The core architecture follows a residual framework and innovatively incorporates multi-stage parallel spiking convolution and recurrent modules. Its details are illustrated in Fig. 2.

As the figure shows, the encoder utilises a 1D convolutional layer (Conv1D) to transform the input speech signal from the time domain to the feature domain, with kernel size $L = 20$, stride $stride = 10$, and output channels $N = 512$. The encoded features are processed by a ReLU activation and channel-wise normalisation before being projected into a bottleneck layer with $B = 256$ dimensions using binary convolution, which is an improved layer that adds a learnable threshold mechanism to the standard 1D convolution. It first performs regular convolution operations, then uses an adaptive activation function to apply non-linear filtering to features that exceed the threshold.

The bottleneck block consists of a parallel temporal processing module, containing two key components: SMC and SMR. SMC performs context feature extraction via depth-wise separable convolution and integrates temporal dynamics using PLIF neurons, where the membrane time constant, firing threshold, and reset value are initially set to $\tau = 2.0$, $V_{th} = 1.0$, and $V_{reset} = 0$, respectively. The SMR module also employs PLIF neurons. Mamba module is implemented with a model dimension of $d_{\text{model}} = 32$, a state-space dimension of $d_{\text{state}} = 16$, a depth-wise convolution kernel size of $d_{\text{conv}} = 4$, and an expansion factor of 2. The selective scan operation and parameter initialization follow the default implementation [9]. During inference, the input is processed causally using 30-ms frame chunks with 50% overlapping. Explicit Mamba state caching is not used, so the Mamba states are re-initialized on each call and not propagated across consecutive frames.

The decoder reconstructs the waveform in the time domain using transposed convolution (ConvTranspose1D) with a residual connection to the original input waveform.

The loss function combines time-domain scale-invariant signal-to-noise ratio (SI-SNR), mean squared error (MSE), and sparsity constraints on spiking activity (L1 regularisation). The final objective is formulated as:

$$\begin{aligned} L = {} & 0.001 \times L_{MSE} + L_{SI-SNR} \\ & + 0.001(||e_{proj}||_1 + ||e_{readout}||_1), \end{aligned} \tag{13}$$

where e_{proj} is the spike activations in SMC (output of spike convolution), while $e_{readout}$ is the spike before the decoding layer. They are also named events. Given a time step t, it can be formulated as:

$$\text{event}(t) = \begin{cases} 1 & \text{if membrane potential } V(t) \geq \text{threshold} \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

These events are sparse, event-driven, discrete signals, analogous to the action potentials emitted by biological neurons.

4.4. Streaming training & inference framework

Fig. 3 describes the streaming framework utilised in our system, which involves dividing the input audio stream into frames, buffering, enhancing the frames, and outputting the enhanced audio stream. Initially, the raw audio stream is divided into multiple overlapping frames. On the time axis, these frames represent consecutive audio windows, each containing localised audio information.

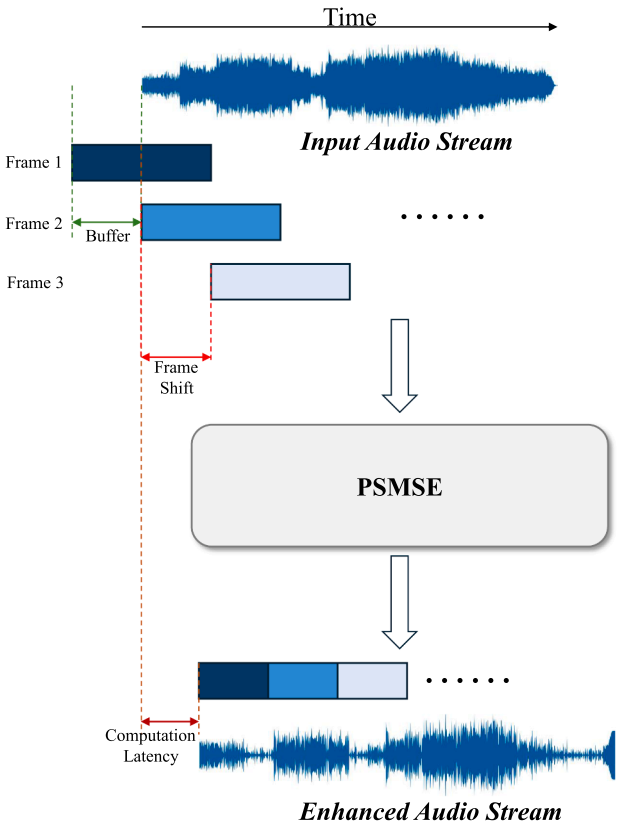

Fig. 3. Streaming processing of the proposed system.

To ensure continuous processing, the system buffers the input stream to ensure that each audio frame has enough data for processing. Let the time sequence of the audio stream be represented by $x[n]$, and each frame of audio is represented by $x[n_i]$, where n_i refers to the sample indices within the audio window.

For each frame, the audio signal undergoes a frame shift, meaning its starting position is offset relative to the previous frame. Let the frame shift be denoted as Δn, so each frame $x[n_i]$ is processed at $n_i + \Delta n$, ensuring continuity and smooth transition of the audio signal. During the processing of each frame, the system applies certain algorithms to enhance the audio. This process can be represented by a speech enhancement function $\hat{x}[n] = f(x[n])$, where $f(x[n])$ is the enhancement operation applied to the input audio signal, and $\hat{x}[n]$ is the enhanced output audio signal.

In the streaming process, each frame requires a certain amount of computation time, leading to computational latency. The computation latency depends on the size of the audio frame and the complexity of the algorithms used. The enhanced audio stream is output in sync with the original. Let the original audio stream be $x[n]$, and the enhanced audio stream be $\hat{x}[n]$, where $\hat{x}[n]$ represents the processed audio signal, which has been significantly improved in quality. Through this streaming framework, the system achieves fast, continuous speech enhancement, ensuring real-time audio quality improvement.

4.5. Personalized task optimization

In real-world communication systems, personalised speech enhancement performs targeted optimisation by analysing user characteristics. For example, it can automatically amplify the speaker's voice and suppress keyboard typing sounds in a video conference, or compensate for specific frequency bands in smart headphones based on the user's hearing curve. It can also adapt to personal pronunciation habits when interacting with voice assistants, significantly improving voice clarity and intelligibility. Therefore, in this section, we take personalised hearing aids as an example to explore the potential of the proposed PSMSE in personalised real-time speech enhancement.

Modern medicine often uses pure-tone audiometry to quantify hearing loss [33]. A standardised audiometer is used to test an individual's minimum audible threshold (in dB HL) for pure tones of different frequencies (250 Hz - 8000 Hz) in a soundproof environment. The resulting audiogram can intuitively display the degree of hearing loss at each frequency point. The audiograms of a normal person and a hearing-loss patient are shown in Fig. 4. These data directly determine the initial form of the hearing aid

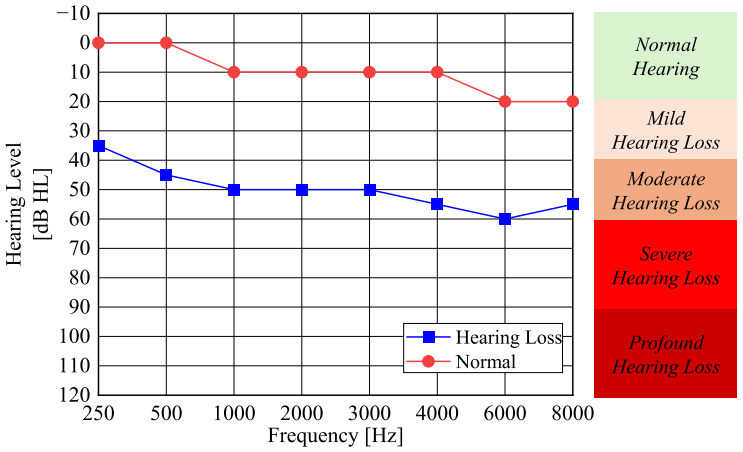

Fig. 4. Normal and hearing-loss audiograms (monaural).

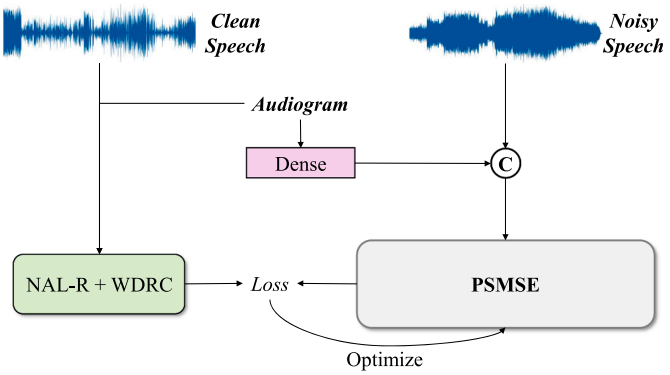

Fig. 5. Optimized system for personalized hearing aids.

frequency response compensation curve. For example, for a threshold defect of 60 dB at 4 kHz, the hearing aid will set a higher gain in this frequency band and, at the same time, classify the overall degree of loss, providing an objective basis for subsequent personalised debugging.

As shown in Fig. 5, in our optimised system for hearing aid users with different audiograms, we still primarily rely on PSMSE to enhance speech quality. Unlike the general model, the user's audiogram will be used as prior knowledge to fuse features with the noisy input speech. Specifically, during the training phase, the audiogram (an 1×8 tensor) is processed through a dense layer to form a feature tensor having the same dimension as the waveform feature, which is then fused with it in concatenation to form the input of PSMSE. This dimension is related to the frame length; for example, it equals 160 when the frame length is 10 ms. This operation exists in each frame. Besides, the NAL-R [25] and wide dynamic range compression (WDRC) are used for pre-processing of reference clean speech. This module combines linear gain (NAL-R) and wide-dynamic-range compression (WDRC) to adjust the sound based on the user's audiogram. The purpose is to ensure that the target speech better adapts to the individual's hearing needs, thereby achieving a more ideal speech perception effect. Therefore, the entire system is optimised using a loss function between the PSMSE output and the NAL-R + WDRC module, with full use of the personal audiogram.

In addition to audiograms, features from other scenarios can also be used to optimise the system in a personalised way, such as feature vectors representing the sound range in vehicle acoustic scenarios and the speaker structure in mobile devices. This enables our model to adapt to a wider range of personalised tasks and scenarios.

5. Experiment settings

5.1. Training details

The hyperparameters of layers in our network are depicted in Section 4.3. During the training phase, we adopted an end-to-end framework. We process audio at a sampling rate of 16 kHz, and each frame contains 0.03 s of context information. The training batch size is set to 32 per GPU and executed across 2 GPUs using a distributed data parallel strategy, yielding a total effective batch size of

64. The Adam optimiser is used, with an initial learning rate of 3×10^{-4} and a cosine annealing schedule. The hyperparameters of the optimiser are set to $\beta_1 = 0.99, \beta_2 = 0.999$. The training cycle is capped at 500 epochs, and the model is saved using a top-k strategy, retaining the best 3 checkpoints based on the validation set. We use PyTorch as the training framework and train on a dual NVIDIA RTX 3090 GPU setup. A single full training session takes approximately 13 h. The CPU of our system is Intel(R) Xeon(R) Silver 4314.

5.2. Data and baselines

In this work, we use multiple datasets to improve the model's generalisation. For the clean speech in the training and validation sets, we select data from the DNS Challenge [34], which consists of audio-book-reading speech with a total duration of 550 h. Another part of the clean speech comes from the CommonVoice 11.0 [35] English portion and the VCTK reading speech [36], with data durations of 550 h and 80 h, respectively. On the other hand, noise data is also collected from multiple sources, including noise data provided by DNS challenge [34], the WHAM! dataset [37], and NoiseX92 [38].

Before simulation, a voice activity detection (VAD) algorithm from SpeechBrain [39] is used to identify speech samples that are actually non-speech or mostly silence, which are then removed from the data. After this, the non-intrusive DNSMOS [40] scores are calculated for each remaining speech sample. This allows us to filter out noisy and low-quality speech samples via thresholding each score. The final data will be used for simulating the training and validation data in the next step. The details of this pre-processing method are listed in Algorithm 1. We refer to the previous work [41] to set the threshold.

Algorithm 1 Data preprocessing and filtering strategy.

Input: Original dataset $\mathcal{D}_{raw} = \{x_1, \ldots, x_N\}$; VAD threshold $\tau_{vad} = 0.5$; DNSMOS thresholds $\tau_{dnsmos} = 3.0$.
Output: High-quality dataset $\mathcal{D}_{hq}$.

Initialize $\mathcal{D}_{hq} \leftarrow \emptyset$
for each clip $x_i \in \mathcal{D}_{raw}$ **do**
  // Step 1: VAD-Based Filtering
  Obtain VAD sequence $V_i \leftarrow \text{VAD}(x_i)$, where V_i is a sequence of $\{0, 1\}$
  Calculate active speech ratio $r_i \leftarrow \frac{\sum V_i}{\text{length}(V_i)}$
  if $r_i < \tau_{vad}$ **then**
    continue
  end if
  // Step 2: DNSMOS-Based Filtering
  Obtain quality scores $s_{dnsmos} \leftarrow \text{DNSMOS}(x_i)$
  if $s_{dnsmos} < \tau_{dnsmos}$ **then**
    continue
  end if
  $\mathcal{D}_{hq} \leftarrow \mathcal{D}_{hq} \cup \{x_i\}$
end for
return $\mathcal{D}_{hq}$

After pre-processing, we simulate the training and validation sets by randomly mixing the pre-processed clean corpus with noise, and all samples are normalised to 16 kHz and 5 s. The SNR of each sample was a random integer in the range $[-5, 20]$ dB. The total training set duration is about 600 h, of which 10% is used for validation.

For testing and comparing the performance of our method with baseline models, we use the general benchmark, including the VoiceBank-DEMAND [42] and the DNS challenge 2020 no-reverb test set [43]. We also use unseen speech and noises, including AISHELL-1 [44] and IEEE ICME 2024 Grand Challenge data [45], to simulate speech with SNR $\in [-5, 0, 5]$ to further evaluate the model performance. Details about the testing data in each experiment are discussed in Section 6.

Additionally, all audiograms in this study are provided by Clarity challenge [46]. They are all real-world data from hearing-aided listeners. The total number is 100; we use 85% for training and the rest for evaluation. Train/test audiograms are disjoint.

As for baseline models, we select those with the same number of parameters and good performance, including state-of-the-art methods. All baselines follow the original setting in their papers. For the baseline comparisons, the results of the models listed in Tables 2–4 are taken from the metrics reported in their original papers, while the remaining baseline models are retrained using the same training data as PSMSE for fair comparison. The models evaluated in Tables 3, 7, and Fig. 14 are implemented under a causal inference setting, whereas the other comparisons follow a non-causal inference setting.

5.3. Assessment metrics

Speech quality metrics:

- PESQ [47] (-0.5 to 4.5): a speech quality assessment metric based on human auditory perception, simulating subjective listening experiences. Higher values indicate speech that is closer to the original clarity.

Table 2
Offline evaluation results on VoiceBank-DEMAND test data.

Models	Year	Para [M]	PESQ	STOI [%]	CSIG	CBAK	COVL	MACS [G/s]
Noisy	—	—	1.97	79.0	3.34	2.82	2.74	—
DCCRN [12]	2020	3.67	2.67	93.7	3.53	2.43	3.08	24.69
DPCRN [54]	2021	0.67	2.51	93.5	2.94	3.25	2.70	4.81
TSTNN [55]	2021	0.92	2.75	94.0	3.93	3.44	3.34	19.95
DeepFilterNet [24]	2022	1.78	2.81	94.2	4.14	3.31	3.46	0.35
FFC-AE-V0 [56]	2022	0.42	2.67	86.5	4.06	3.30	3.37	4.06
CCFNet+ [57]	2023	0.62	3.03	95.0	**4.27**	3.55	**3.61**	1.47
aTENNuate [58]	2025	0.84	**3.14**	94.2	4.30	2.62	3.73	0.33
L-DPSB [59]	2025	1.16	3.00	**96.0**	3.96	3.75	3.52	49.90
HiFi-Stream [60]	2025	1.17	2.70	86.5	4.03	3.22	3.37	1.90
HiFi-Stream-2D [60]	2025	0.50	2.65	86.4	4.02	3.17	3.35	1.97
PSMSE	2025	**0.26**	2.85	95.1	4.03	**3.77**	**3.61**	**0.11**

- STOI [48] (0 - 100%): objective intelligibility metric quantifying speech comprehensibility. Predicts human recognition of speech content through time-frequency correlation analysis.
- CSIG, CBAK, COVL [49] (0 to 5): Metrics reflect speech naturalness, denoising performance, and overall acceptability, respectively.

Complexity metrics:

- Number of parameters: model parameter count, directly reflecting model size and memory footprint.
- MACS: multiply-accumulate operations required per second of audio processing.

Downstream task metrics:

- Speaker similarity: calculating the speaker embedding using the pre-trained model in ESPNet [50] and calculating the similarity using the VERSA toolkit [51].
- Emotion similarity: transforming the model output and ground truth into vectors based on Emo2Vec [52] and calculating the similarity using the VERSA toolkit [51].
- Character accuracy: doing ASR on the speech and calculating the character accuracy compared to the reference text based on the ESPNet [50] toolkit.

Personalised metrics for hearing aid application:

- HASPI [53]: predicts speech intelligibility for hearing-impaired listeners (0–1, higher values indicate clearer speech) based on auditory neural spectrum correlation.
- HASQI [53]: evaluates naturalness of assisted hearing sound quality (0–1, higher values indicate more natural sound) through harmonic distortion analysis.

6. Results & analysis

6.1. Results on VoiceBank-DEMAND

As presented in Table 2, our proposed model is evaluated on the VoiceBank-DEMAND test dataset against a range of recent and state-of-the-art speech enhancement models. The results demonstrate that PSMSE achieves competitive or superior performance in speech quality and intelligibility while maintaining high computational efficiency.

Specifically, in terms of objective evaluation metrics, our model showcases a remarkable balance. PSMSE achieves the highest score for CBAK, with a value of 3.77, indicating its superior ability to suppress background noise without introducing significant artefacts. Furthermore, it achieves the highest COVL score of 3.61, matching the performance of recent CCFNet+. For speech intelligibility, as measured by STOI, our model scores a high 95.1%, which is highly competitive and surpasses most of the listed models, falling only slightly behind L-DPSB (96.0%). In the perceptual evaluation of speech quality (PESQ), PSMSE scores 2.85. While some models, like aTENNUate, achieve a higher PESQ score (3.14), our model's performance remains robust and comparable to other leading methods.

The most significant advantage of the PSMSE model lies in its unparalleled efficiency. As depicted in Fig. 6, with only 0.26 million parameters, it is the lightest-weight model with the best performance in CSIG, CBAK, and COVL. It has fewer than half the parameters of the next-smallest model, FFC-AE-V0 (0.42M). This efficiency advantage is even more pronounced in terms of computational complexity, where PSMSE requires 0.11 G/s MACS. This is much lower than computationally intensive models like L-DPSB (49.90 G/s) and DCCRN (24.69 G/s), and is also significantly more efficient than other light-weight models such as DeepFilterNet (0.35 G/s) and aTENNUate (0.33 G/s).

To evaluate the practical applicability of our model for real-time inference, we conducted tests using a streaming framework on the VoiceBank-DEMAND dataset, and the results are detailed in Table 3.

As shown by the results, our proposed method achieves strong performance, outperforming all comparative models across most evaluation metrics. PSMSE obtains the highest PESQ score of 2.73. This is a significant improvement over the unprocessed noisy audio score of 1.97 and surpasses the next-best model, DEMUCS [61], which scores 2.66, demonstrating our model's effectiveness in producing speech that is perceived as cleaner and more natural. As for STOI, PSMSE achieves a leading score of 93.9. This represents

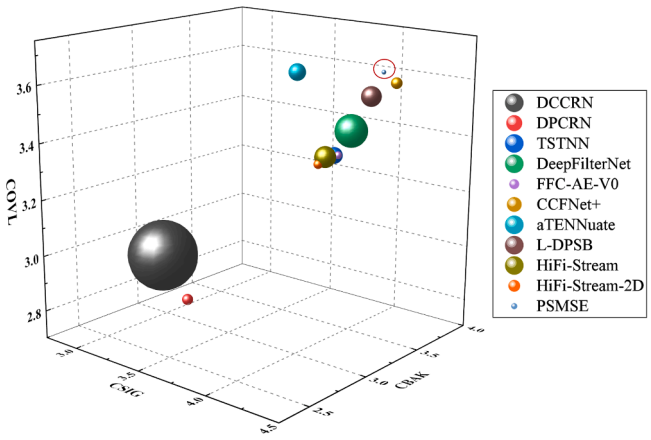

Fig. 6. Comparison of parameter number and performance. Our method is marked by a circle.

Table 3
Streaming evaluation results on VoiceBank-DEMAND test data.

Models	PESQ	STOI [%]	CSIG	CBAK	COVL
Noisy	1.97	79.0	3.34	2.82	2.74
DEMUCS [61]	2.66	88.1	3.82	3.23	3.24
FFC-AE-V0 [56]	2.00	78.9	3.50	2.80	2.74
FFC-AE-V1 [56]	2.42	86.0	3.90	3.21	3.17
FFC-Unet [56]	2.45	86.1	**3.97**	3.22	3.22
HiFi++ [62]	2.37	93.0	3.84	3.16	3.10
HiFi-Stream [60]	2.30	84.5	3.67	2.86	2.98
HiFi-Stream-2D [60]	2.29	84.7	3.68	2.93	2.98
PSMSE	**2.73**	**93.9**	3.85	**3.29**	**3.31**

Table 4
Evaluation results on DNS test data. "MACs" here is the total MAC operations processing one sample.

Models	Para [M]	PESQ	STOI [%]	MACs [G]
Noisy	—	1.38	59.0	—
NSNet [63]	1.3	2.18	—	7.24
DCCRN [12]	3.67	1.95	93.2	140.53
DPCRN [54]	0.67	2.16	93.3	27.61
CA DU-Net [64]	26.0	1.99	83.0	—
FasNet TAC [65]	2.9	2.26	87.0	55.3
TPARN [66]	3.2	2.75	92.0	71.8
ADCN [67]	9.3	2.84	92.0	72.8
DRC-Net [68]	2.6	2.79	89.0	13.9
DeFT-AN [69]	2.7	**3.01**	92.0	95.6
PSMSE	**0.26**	2.72	**94.6**	**1.05**

a substantial gain over strong competitors like HiFi++ (93.0). Furthermore, PSMSE scores highest in overall quality (COVL score of 3.31) and in background noise suppression (CBAK score of 3.29), confirming its balanced ability to reduce artefacts while effectively removing unwanted noise. While FFC-Unet achieves a slightly higher CSIG score of 3.97 compared to our model's 3.85, PSMSE's dominant performance across the other 4 key metrics, especially in both perceptual quality and objective intelligibility, robustly establishes its overall superiority and effectiveness for the streaming speech enhancement task.

6.2. Results on DNS test data

We also conduct an evaluation on the DNS challenge test set. We compare our model, PSMSE, against a range of contemporary speech enhancement models. The quantitative results are presented in Table 4. The proposed PSMSE model demonstrates better performance across all evaluation metrics compared to the other listed methods. PSMSE achieves the highest PESQ score of 2.72. This result shows a significant improvement over unprocessed noisy audio, which has a PESQ score of 1.38, and surpasses most baseline models. Additionally, our method achieves the highest STOI score. It achieves a STOI of 94.6%, a considerable improvement over the

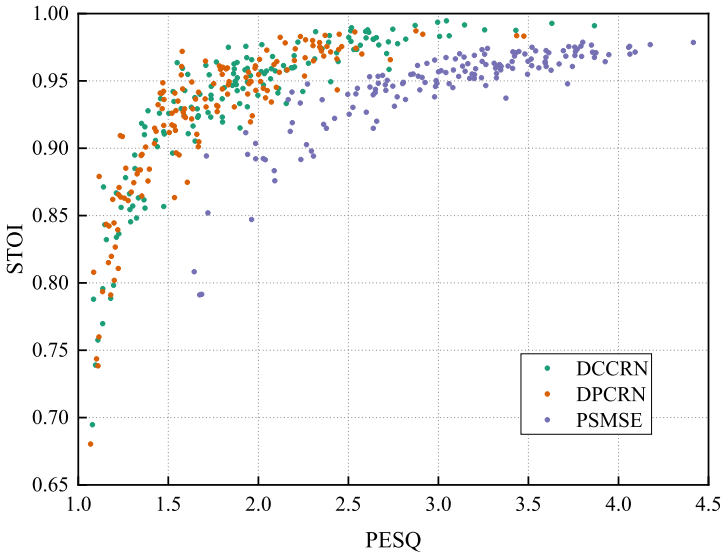

Fig. 7. Comparison of distribution of PESQ and STOI scores for all test samples.

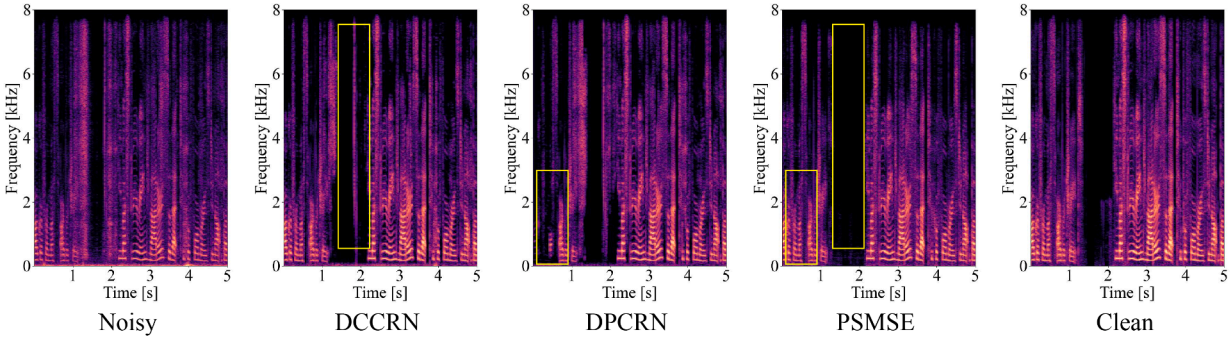

Fig. 8. Visualization of the enhanced speech samples.

noisy baseline of 59.0%. This score is notably higher than that of all other competing models, including DCCRN and DPCRN, which reached 93.2% and 93.3%, respectively. This indicates that the speech processed by our model is clearer and easier to understand. Additionally, our model remains the lightest, with only 0.26 million parameters and a total MAC count of 1.05G, making it the most compact model in the comparison.

To provide a more detailed analysis of our model's performance, we visualised the distributions of PESQ and STOI scores for each sample in the DNS test set. Fig. 7 presents a scatter plot comparing our proposed PSMSE with two baseline models, DCCRN and DPCRN. Each point on the plot represents an individual test sample, with its horizontal position indicating the PESQ score and its vertical position representing the STOI score.

The visual plot clearly illustrates the superiority of the PSMSE model. The cluster of points representing PSMSE is located distinctly in the upper-right region of the graph. This placement indicates that our model achieves higher scores in both speech quality and intelligibility across the vast majority of test samples. Compared to DCCRN and DPCRN, PSMSE's results show a trend toward PESQ scores generally above 2.5 and STOI scores frequently above 0.95. In contrast, the distributions for DCCRN and DPCRN are situated more towards the lower-left area of the plot. Many of their test samples fall within the PESQ range of 1.5 to 2.5 and the STOI range of 0.85 to 0.95. Furthermore, these baseline models exhibit greater variance, with a notable number of samples scoring below 1.5 on PESQ and 0.85 on STOI, indicating less consistent performance across the entire test set. The PSMSE model, however, demonstrates a more concentrated and stable distribution in the high-performance region of the chart.

In Fig. 8, we present a visual comparison of the spectrograms of a sample in test data. The figure displays the enhancement results for this speech sample, showing the original noisy signal, the outputs from DCCRN, DPCRN, and PSMSE, and the clean reference signal. From the spectrograms, we can observe differences in the enhancement capabilities of each model. The DCCRN, while removing significant noise, still leaves residual noise artefacts. This is particularly evident in the highlighted yellow box, where vertical patterns of noise present in the noisy input remain visible in the output. On the other hand, the DPCRN model appears to be overly aggressive in its noise suppression. The area within its yellow box shows that parts of the harmonic structure of the speech

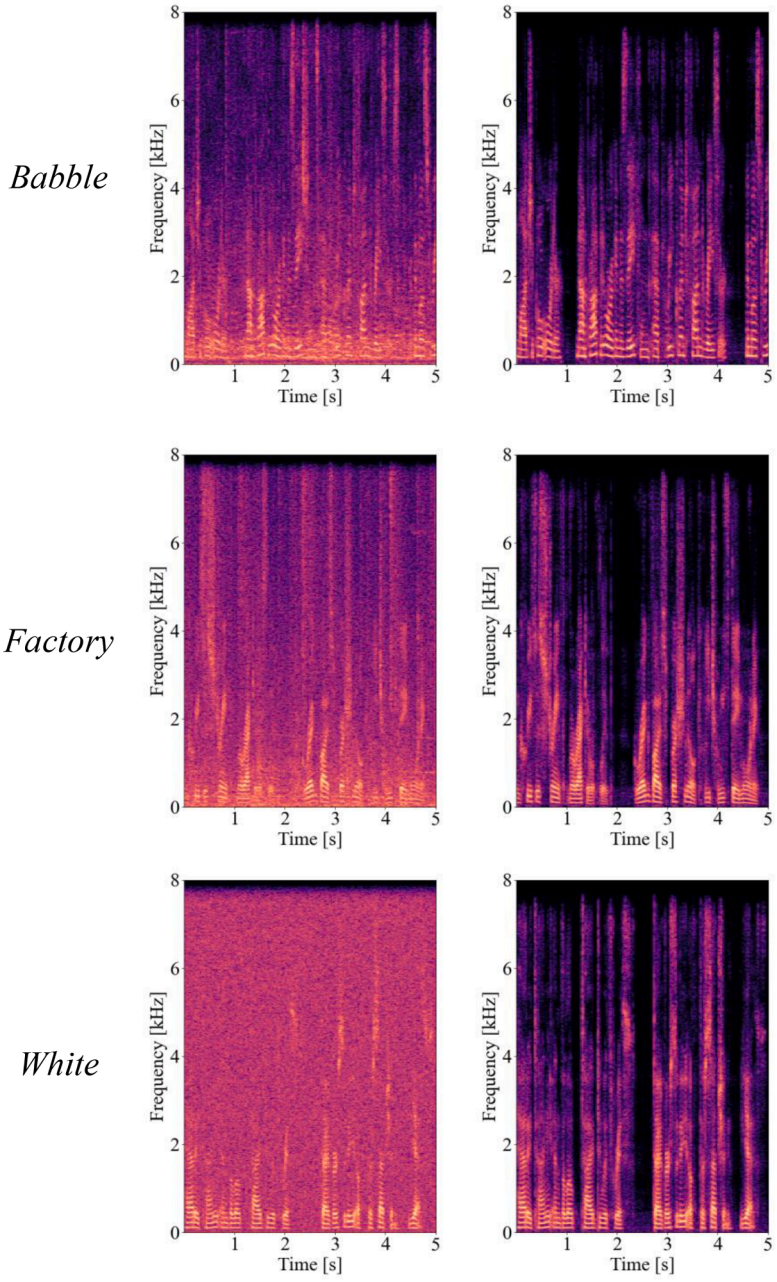

Fig. 9. Visualisation of the speech enhanced by our proposed model in 3 kinds of noises (right) and its noisy input (left).

signal have been mistakenly removed along with the noise, leading to speech distortion and a potential loss of fidelity. Instead, our proposed model demonstrates a better balance between noise suppression and speech preservation. The spectrogram for PSMSE shows that within the highlighted region, the background noise has been effectively eliminated. At the same time, the essential speech components have been accurately retained without the obvious distortion seen in the DPCRN's result. Compared to the other models, PSMSE's output most closely resembles the spectrogram of the clean reference signal.

6.3. Performance under different conditions of noise

To evaluate the model performance under different noise types and SNR levels. We conduct the experiment based on clean speech from the DNS test set [43] and noises from the NoiseX92 dataset [38]. Each clean speech clip is mixed with a certain type of noise at

Table 5
Model performance comparison under 3 types of noises and different SNR levels.

Noise Type	Babble (Speech Based Noise)						Factory (Non-Speech Based Noise)						White (Non-Speech Based Noise)					
Metrics	PESQ			STOI [%]			PESQ			STOI [%]			PESQ			STOI [%]		
SNR (dB)	-5	0	5	-5	0	5	-5	0	5	-5	0	5	-5	0	5	-5	0	5
Noisy	1.06	1.09	1.18	57.3	70.9	83.0	1.06	1.05	1.11	57.2	70.3	82.4	1.03	1.03	1.05	65.9	75.6	84.8
DCCRN [12]	1.08	1.24	1.59	58.8	77.8	89.2	1.07	1.20	1.46	57.0	75.4	87.8	1.08	1.16	1.34	62.8	74.3	85.6
DPCRN [54]	1.05	1.13	1.31	52.4	74.9	87.7	1.05	1.11	1.22	57.5	76.3	86.9	1.03	1.04	1.09	53.0	70.3	80.9
DEMUCS [61]	1.11	**1.33**	**1.75**	**67.2**	**85.5**	**93.2**	1.13	1.31	1.64	**73.3**	**85.8**	**92.2**	1.25	1.37	1.68	79.0	85.5	91.3
DeepFilterNet [24]	1.12	1.32	1.71	61.8	80.8	90.9	1.18	1.37	1.68	68.7	82.5	90.5	**1.36**	1.54	1.83	77.2	83.8	89.9
PSMSE	**1.13**	**1.33**	**1.75**	63.8	81.9	91.5	**1.19**	**1.41**	**1.78**	70.4	84.1	91.6	1.35	**1.55**	**1.93**	**80.0**	**86.7**	**92.1**

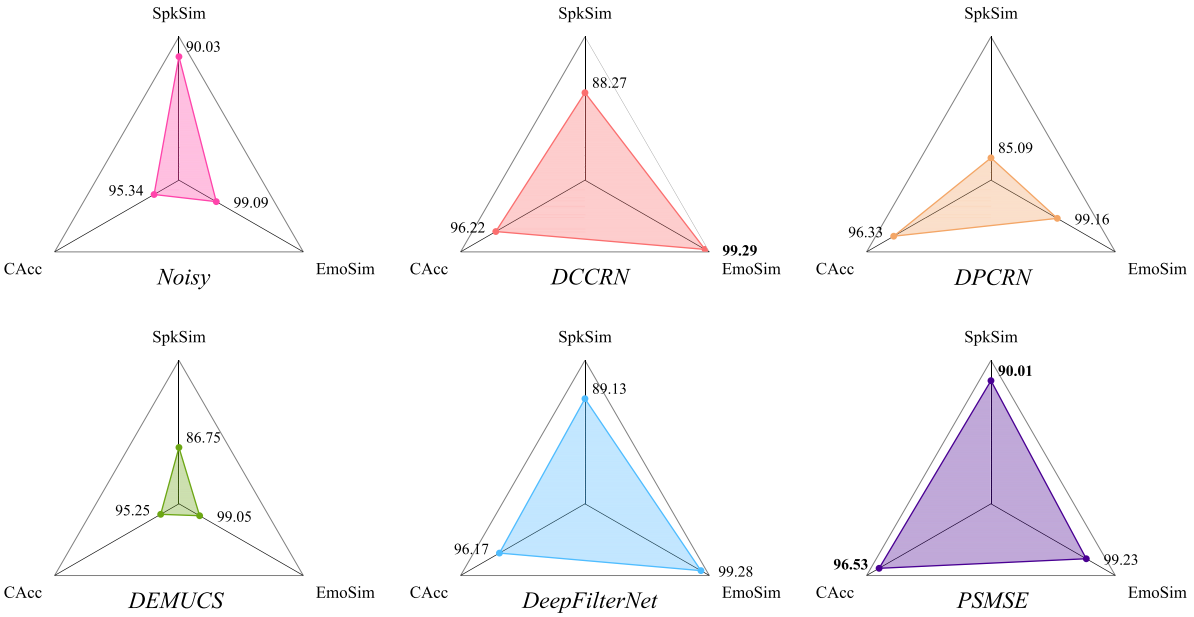

Fig. 10. Accuracy comparison of downstream tasks using enhanced speech from different models. The unit for all metrics is "%".

a fixed SNR level. The comparison results are listed in Table 5, and the visualisation of our model outputs in this test is depicted in Fig. 9.

As shown in the table, the proposed model demonstrates better overall performance under various noise types (including babble, factory, and white noises) and different SNR levels.

In detail, our model achieves optimal or near-optimal performance under all noise types and SNR conditions in the PESQ test. In the most challenging low-SNR (-5 dB) environment, PSMSE achieves the highest PESQ scores of 1.13 and 1.19 under babble noise and factory noise, respectively, significantly outperforming the second-best DEMUCS (1.11 and 1.14). At a high SNR level (5 dB), PSMSE continues to lead across both noise types, achieving PESQ scores of 1.75 and 1.78, respectively. In the presence of white noise, PSMSE achieves the highest PESQs of 1.55 and 1.93 at SNRs of 0 and 5 dB, respectively, demonstrating its robustness to non-speech noise.

In terms of the objective speech intelligibility metric STOI, the PSMSE model also demonstrates strong competitiveness. In white noise environments, PSMSE achieves an STOI of 92.1% at an SNR of 5 dB, while other models are slightly lower. Furthermore, PSMSE achieved a STOI of 70.4% under a very low signal-to-noise ratio (-5 dB) in factory noise, demonstrating its ability to maintain speech intelligibility under demanding listening conditions. Although the DEMUCS model's STOI is higher than PSMSE's (93.2% vs 91.5%) under specific conditions, our model's overall performance confirms its comprehensive optimisation capabilities in improving speech quality and intelligibility.

6.4. Performance for downstream tasks

To evaluate the effectiveness of speech enhancement models in practical applications, relying solely on objective auditory quality metrics is insufficient. Therefore, we further tested the performance of each model's enhanced speech in key downstream tasks. We evaluated the enhanced speech on the VoiceBank-DEMAND dataset [42] across multiple downstream tasks, covering speech

Table 6
Models' performance on unseen data.

Models	PESQ	STOI [%]	CAcc
Noisy	1.44	73.6	80.24
DCCRN [12]	1.69	76.5	76.30
DPCRN [54]	1.80	76.9	79.44
DEMUCS [61]	2.09	81.5	80.94
DeepFilterNet [24]	2.11	83.4	82.03
PSMSE	**2.33**	**85.9**	**82.57**

Table 7
Models' performance on hearing aid. The "HLC" denotes hearing loss compensation, which is implemented by NAL-R and WDRC.

Methods	Type	HASPI	HASQI
NAL-R	HLC-Only	0.35	0.28
NAL-NL2	HLC-Only	0.38	0.25
WDRC	HLC-Only	0.41	0.31
DCCRN + HLC	2 Stage	0.84	0.46
DPCRN + HLC	2 Stage	0.85	0.49
DEMUCS + HLC	2 Stage	0.87	0.53
DeepFilterNet + HLC	2 Stage	0.89	0.56
PSMSE	End-to-End	**0.90**	**0.60**

recognition, speaker recognition, and speech emotion recognition. The experimental results are shown in Fig. 10. SpkSim, Emosim, and CAcc in the figure denote speaker similarity, emotion similarity, and character accuracy, respectively.

In the speech recognition task, the PSMSE model achieved the highest CAcc score of 96.53%, significantly higher than the unprocessed Noisy baseline and DEMUCS model, and also outperforming the well-performing DPCRN and DCCRN. This demonstrates that the model effectively preserves speech content while removing noise, thereby improving recognition accuracy.

In the crucial task of speaker recognition, speech enhancement algorithms often suffer from speaker feature loss due to the introduction of nonlinear distortion. For example, the SpkSim scores for DPCRN and DEMUCS drop to 85.09% and 86.75%, respectively, which are lower than the Noisy baseline of 90.03%. In contrast, the PSMSE model demonstrates excellent feature preservation in this metric, achieving a SpkSim score of 90.01%, nearly matching that of the original speech. This indicates that PSMSE effectively preserves the speaker's timbre and identity features during noise reduction, overcoming the drawback of traditional models, which equate noise reduction with distortion.

PSMSE also performs robustly in the speech emotion recognition task, achieving an EmoSim score of 99.23%, outperforming Noisy and DPCRN, and on par with DCCRN and DeepFilterNet. Given the proposed model's lower parameter count and computational cost, this result is reasonable and acceptable.

6.5. Performance under domain shift

Domain shift is a challenging issue for deep neural models. To measure the model's generalisation performance and robustness, in this sub-section, we conduct comparative experiments on a test set not used during training, using PESQ, STOI, and downstream speech recognition accuracy (CAcc) as evaluation metrics. We use the clean speech from AISHELL-1 [44] and the noises from the IEEE ICME ASC challenge data [45]. The speech samples from AISHELL are all in Mandarin, which did not appear during the training phase of all models. The IEEE ICME ASC challenge dataset contains environmental sounds from 10 real-world scenes, and it is also not seen by all models. To ensure the fair comparison with the seen data, we randomly mix the speech and noises with the same range of SNR in Section 6.2's data, which is [-5, 20] dB. The total number of speech clips is also kept the same, which is 150.

The test results are listed in Table 6. PSMSE achieved PESQ and STOI scores of 2.33 and 85.9%, respectively, both significantly surpassing the noisy inputs' 1.44 PESQ score and 73.6% STOI score. It also significantly outperforms DeepFilterNet, which achieves PESQ and STOI scores of 2.11 and 83.4%, respectively. It is worth noting that while traditional DCCRN and DPCRN models implement noise reduction to some extent, the introduction of severe nonlinear distortion or speech loss reduces their speech recognition accuracy on unseen data to 76.30% and 79.44%, respectively, even below the 80.24% achieved on noisy speech. However, PSMSE not only successfully avoids the negative effects of this over-enhancement but also improves the CAcc to 82.57%.

On the other hand, a visual comparison in Fig. 11 shows that, due to differences in data distribution, all models experienced a degree of performance decline on unseen data compared to seen data. However, the PSMSE model consistently maintained a significant lead in both speech quality and intelligibility, and its performance degradation was more controllable than that of other baseline models.

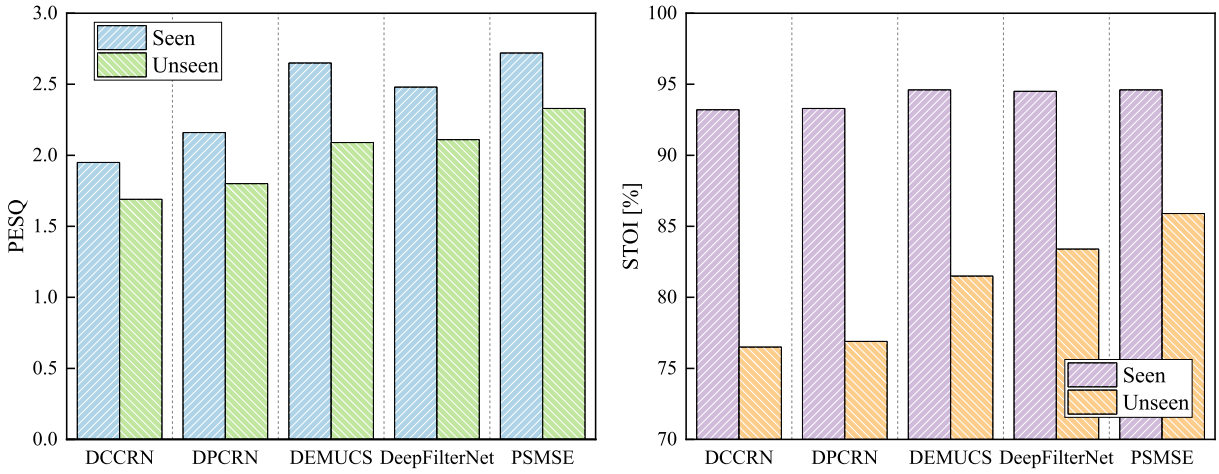

Fig. 11. PESQ and STOI comparisons between seen and unseen data.

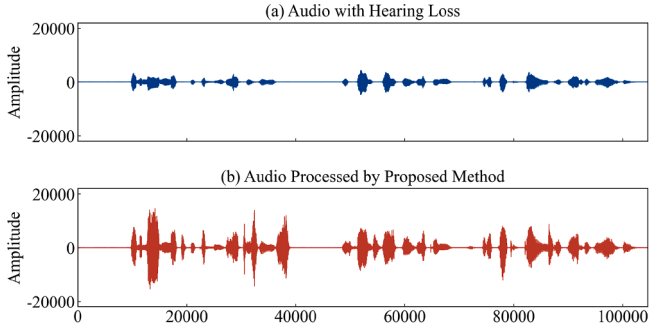

Fig. 12. Comparison between the input audio with hearing loss (a) and the output audio enhanced by our method (b).

6.6. Hearing aid performance

In these experiments, we first simulate hearing loss for all audio signals using the MBGS method [70,71] proposed by Cambridge University, based on audiograms provided by Clarity [46], followed by enhancing and testing these data using different models.

Based on the experimental results in Table 7, the proposed PSMSE achieves the best overall performance, with a HASPI score of 0.90 and a HASQI score of 0.60. Compared with classical hearing-loss-compensation-only methods, including NAL-R, NAL-NL2, and WDRC, PSMSE provides substantial improvements in both intelligibility and quality. This indicates that compensation-only strategies are insufficient when the input speech is also corrupted by noise, as they primarily adjust spectral gain based on the audiogram rather than explicitly suppressing background noise.

Compared with two-stage systems that first perform speech enhancement and then apply hearing loss compensation, PSMSE also achieves better results than the strongest baseline, DeepFilterNet + HLC, which obtains 0.89 HASPI and 0.56 HASQI. This suggests that jointly optimising speech enhancement and hearing compensation in an end-to-end framework can better preserve perceptually important speech cues while reducing accumulated distortion introduced by cascaded processing. In addition, the end-to-end design simplifies the processing pipeline and avoids redundant computation across separate modules, making it more suitable for hearing aids with strict real-time and resource constraints.

Besides, Fig. 12 visually illustrates the compensation effect with a comparison of the time-domain waveforms. In the input audio after simulating hearing loss, its amplitude is severely attenuated due to the increased auditory threshold, making many details difficult to perceive. After processing by our model, the amplitude of the output audio is significantly and reasonably restored, and the waveform envelope is fuller, indicating that the speech energy is effectively enhanced. Analysis of the frequency response and gain curves (Fig. 13) further reveals the model's working mechanism. The experiment simulated a typical high-frequency hearing-loss scenario, in which the hearing threshold gradually decreases with increasing frequency, reaching approximately 80 dB HL at 6000 Hz. Correspondingly, the model's generated gain curve exhibits precise frequency-dependent characteristics, maintaining low or even negative gain at low frequencies to avoid over-amplification, while increasing gain amplitude at high frequencies, providing nearly 30 dB of gain compensation near 6000 Hz. This gain distribution complements the hearing loss curve well, indicating that the model has successfully learned to adaptively compensate the spectrum based on a specific audiogram.

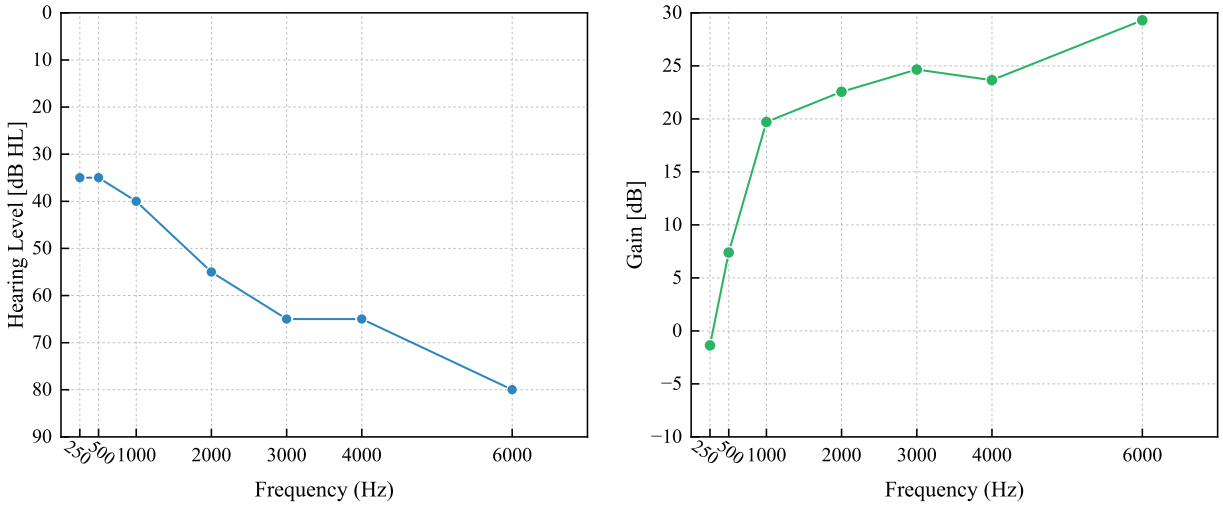

Fig. 13. The input hearing loss level and the gain predicted by the proposed model at different frequency points.

Table 8
Results of ablation studies on our model's architecture.

Item	PESQ	STOI	#Params	MACS [G/s]
Original Settings	2.85	95.1	0.26	0.11
Cascade Structure	2.81	95.0	0.26	0.11
w/o SMC	2.76	94.5	0.19	0.07
w/o SMR	2.70	93.2	0.17	0.06
w/o SNNs	2.75	94.5	0.26	0.61
w/o Mamba	2.80	94.9	0.22	0.10
Replace Mamba with GRU	2.83	94.6	0.25	0.10
Replace PLIF with LIF	2.81	94.7	0.26	0.10
w/o sparsity loss	2.82	94.7	0.26	0.10
w/o binary convolution	2.78	94.6	0.26	0.31
w/o Algorithm 1	2.79	94.8	0.26	0.11

6.7. Ablation study

To enhance explainability, we first conduct an ablation study of the model design. The results are summarised in Table 8.

First, we verified the positive effects of the proposed parallel branch design. Compared to the cascaded structure that connects the SMC and SMR modules, the original parallel setup achieved better performance. This indicates that the parallel structure can fuse local features extracted by SMC with global contextual information captured by SMR more effectively, thereby avoiding information bottlenecks or feature loss that may occur in the cascaded structure.

Subsequently, we further explored the contributions of the two core modules, SMC and SMR. Results show that the model's performance declined most significantly when the SMR module was removed, with PESQ dropping from 2.85 to 2.70 and STOI decreasing to 93.2%, demonstrating the core role of global contextual information in speech enhancement tasks. Similarly, removing the SMC module also causes the PESQ to drop to 2.76, indicating that relying solely on global information while ignoring local spectral details can compromise the precision of speech reconstruction. Although removing either of these modules could significantly reduce the number of parameters and computational complexity, the substantial performance loss makes this simplification counterproductive.

To verify the energy efficiency advantages of SNNs, we replace the spiking neurons in the model with standard artificial neurons. The convolutional and recurrent layers based on spikes are also replaced by their common counterparts. The results are impressive; removing the SNN feature increases the model's computational cost, with MACS rising from 0.11 G/s to 0.61 G/s. This means that the energy consumption of a regular neural network is several times that of the spiking structure presented in this paper. Furthermore, the non-spiking version of the model shows a performance decline, with a PESQ of only 2.75, further illustrating the mismatch between the ANN's structure and the original framework.

We also evaluate the impact of the state-space model and the Algorithm 1. Removing the mamba module slightly reduced the number of parameters and computational cost, but the PESQ score dropped to 2.80, validating that introducing a state-space model further improves feature representation capabilities. Meanwhile, the lack of Algorithm 1 support also leads to a performance decline, demonstrating the necessity of this algorithm in cleaning the training corpus.

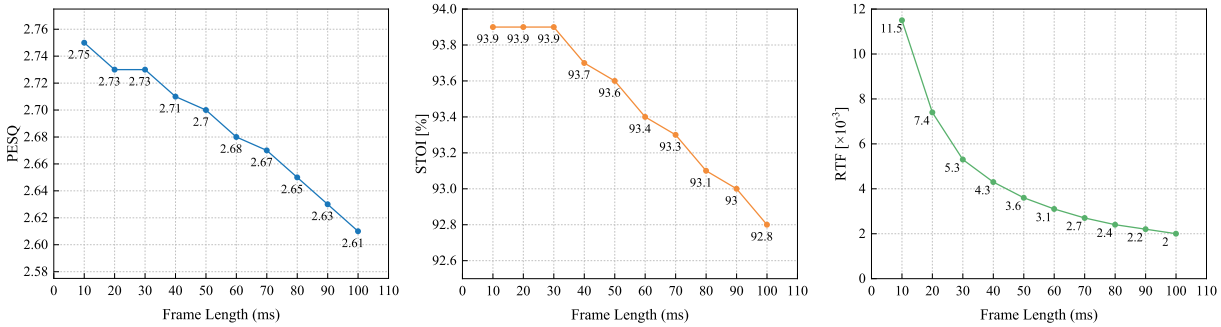

Fig. 14. Relationship between the frame length and the performance. The metrics include PESQ (left), STOI (middle), and real-time factor (RTF, right).

In addition, replacing Mamba with a GRU causes a slight degradation, with PESQ decreasing from 2.85 to 2.81 and STOI decreasing from 95.1% to 94.6%. This suggests that Mamba provides a more effective long-range sequence-modelling capability within a comparable computational budget. Replacing PLIF neurons with standard LIF neurons also results in a slight performance drop, indicating that the learnable membrane time constant in PLIF improves the adaptability of spiking dynamics to speech signals. Moreover, removing the sparsity loss weakens the event-driven regularisation and slightly reduces enhancement performance. Finally, removing the binary convolution results in a noticeable increase in MACS from 0.11 G/s to 0.31 G/s, while PESQ decreases to 2.78, demonstrating that binary convolution is important for maintaining both computational efficiency and an effective sparse feature representation.

To explore the trade-off between computational efficiency and improved quality in a streaming inference framework, we investigate the impact of frame length variations on system performance. Experiments gradually increased the frame length from 10 ms to 100 ms, recording the trends in corresponding speech quality metrics and the real-time rate factor. The experimental environment is the same as the training phase (as mentioned in Section 5.1). The results are put in Fig. 14. From a computational efficiency perspective, increasing the frame length significantly reduces the system's computational load. As shown in the figure, as the frame length increases from 10 ms to 30 ms, the real-time factor declines sharply. This significant acceleration is mainly due to the inverse relationship between the number of inference steps and the frame length: a longer frame length drastically reduces the number of forward propagations the network needs to perform per unit time. However, as the frame length increases to 100 ms, the speed gain plateaus gradually. This is because, although the number of executions of the core network continues to decrease, the system's inherent fixed overhead and the computational cost of a single-layer encoder, which increases linearly with frame length, gradually become dominant, diminishing the marginal benefit of computational efficiency. As for speech enhancement quality, both PESQ and STOI metrics show a gradual decrease with increasing frame length. Experimental data show that over a short frame length range of 10 to 30 ms, the model performance remains high and relatively stable. This is because the short frame length provides high temporal resolution, enabling the model to capture phase and transient changes in the waveform with a complete feature representation. However, when the frame length exceeds 30 ms, performance begins to decline significantly. The reason is that a longer frame length results in a significantly reduced update frequency for spiking neurons. This low refresh rate weakens the network's ability to track the rapidly changing characteristics of the speech signal, leading to the loss of details in the reconstructed waveform.

The proposed streaming system uses a frame length of 30 ms with a 50% overlap-add strategy. The frame length can be adjusted according to different application scenarios. Since the proposed model is strictly causal and does not use future frames, the convolutional lookahead is 0 ms. In addition, no extra future-frame buffering is introduced during streaming inference. Therefore, the end-to-end algorithmic latency is determined by the frame length, i.e., 30 ms in the default setting. For a 1-second speech input during inference, the observed peak memory usage is 266 MB. It is worth mentioning that this reported peak memory was measured using the FP32 implementation and therefore does not directly represent the memory requirement of a final hearing-aid deployment. Reduced-precision inference, such as mixed-precision or INT8 quantization, may reduce the storage cost. However, quantizing the spiking and Mamba-based operators while preserving stable enhancement performance remains nontrivial. Accordingly, the present hardware evaluation should be interpreted as an assessment of algorithmic efficiency rather than a complete validation on hearing-aid-class embedded hardware. Implementing a quantized version and benchmarking it on mobile, DSP, MCU, or hearing-aid-class processors will be investigated in future work.

Table 9 reports the layer-wise spike firing rates of the event-driven modules in PSMSE, providing additional evidence for its suitability for lightweight streaming speech enhancement. Since spike firing rate is only meaningful for event-driven computation, we calculate it only for spiking layers. As shown in the table, all spiking modules maintain relatively low firing rates. The binary convolution layer has a firing rate of 0.2427, while the spike convolution and recurrent layer further reduce the firing rates to 0.1876 and 0.1539, respectively. This indicates that most neurons remain inactive during inference, confirming the proposed architecture's sparse event-driven behaviour. Such sparsity helps reduce unnecessary operations and supports the low-complexity design of PSMSE. The readout layer shows a relatively higher firing rate of 0.3466, which is reasonable because it aggregates the outputs from the two parallel branches and converts sparse spike representations back into features for waveform reconstruction.

Table 9
Layer-wise spike firing rates of the event-driven modules in PSMSE.

Layer	Firing Rate
BinaryConv	0.2427
SpikeConv	0.1876
Recurrent Layer	0.1539
Readout (average of 2 branches)	0.3466

7. Conclusion and future work

In this paper, we propose a lightweight and end-to-end streaming speech enhancement framework based on a parallel spiking Mamba architecture for AIoT-enabled wearable hearing aids. By integrating the energy-efficient, event-driven properties of spiking neural networks with the long-range sequence-modelling capabilities of state-space models, the proposed method effectively addresses the challenges of real-time speech enhancement under strict edge constraints.

Experimental results demonstrate that the proposed model achieves competitive performance in speech quality and intelligibility while significantly reducing parameter size and computational complexity. The streaming design further enables low-latency processing, making the framework suitable for real-time deployment on resource-constrained devices. In addition, the incorporation of personalised auditory information highlights the potential of the proposed approach for human-centred hearing assistance systems.

From an AIoT perspective, this work provides a practical solution for bridging efficient edge intelligence with wearable healthcare applications, demonstrating that neuromorphic-inspired architectures offer a promising direction for low-cost, real-time signal processing in cyber-physical-human systems.

In future work, we plan to further validate the proposed framework on real hardware platforms, including neuromorphic devices, to evaluate its power efficiency and latency in real-world scenarios. We will also extend the current model to multi-channel settings and more complex acoustic environments, aiming to improve robustness and applicability in practical AIoT systems.

CRediT authorship contribution statement

Junkang Yang: Writing – original draft, Visualization, Validation, Software, Methodology, Conceptualization; **Hiromitsu Nishizaki:** Writing – review & editing, Supervision, Funding acquisition; **Chee Siang Leow:** Writing – review & editing; **Hongqing Liu:** Writing – review & editing; **Lu Gan:** Writing – review & editing.

Funding

This work was supported by JSPS KAKENHI Grant Numbers 23H00493 and 25H00566.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] N.L. Westhausen, H. Kayser, T. Jansen, B.T. Meyer, Real-time multichannel deep speech enhancement in hearing aids: comparing monaural and binaural processing in complex acoustic scenarios, IEEE/ACM Trans. Audio Speech Lang. Process. 32 (2024) 4596–4606. https://doi.org/10.1109/TASLP.2024.3473315
[2] G. Zhao, Y. Zhang, J. Chu, A multimodal teacher speech emotion recognition method in the smart classroom, Internet Things 25 (2024) 101069. https://doi.org/10.1016/j.iot.2024.101069
[3] A.J. Soto-Vergel, P. Sankaran, J.C. Velez, R. Amaya-Mier, D. Ramirez-Rios, AtomicVAD: a tiny voice activity detection model for efficient inference in intelligent IoT systems, Internet Things 35 (2026) 101822. https://doi.org/10.1016/j.iot.2025.101822
[4] Y. Yang, A. Pandey, D. Wang, Towards decoupling frontend enhancement and backend recognition in monaural robust ASR, Comput. Speech Lang. 95 (2026) 101821. https://doi.org/10.1016/j.csl.2025.101821
[5] N. Saleem, T.S. Gunawan, S. Dhahbi, S. Bourouis, Time domain speech enhancement with CNN and time-attention transformer, Digit. Signal Process. 147 (2024) 104408. https://doi.org/10.1016/j.dsp.2024.104408
[6] Y. Luo, N. Mesgarani, Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation, IEEE/ACM Trans. Audio Speech Lang. Process. 27 (8) (2019) 1256–1266. https://doi.org/10.1109/TASLP.2019.2915167
[7] S. Ghosh-Dastidar, H. Adeli, Spiking neural networks, Int. J. Neural Syst. 19 (04) (2009) 295–308. https://doi.org/10.1142/S0129065709002002
[8] A. Gu, K. Goel, C. Re, Efficiently modeling long sequences with structured state spaces, in: International Conference on Learning Representations, 2022. https://openreview.net/forum?id=uYLFoz1vlAC.
[9] A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, in: First Conference on Language Modeling, 2024. https://openreview.net/forum?id=tEYskw1VY2.

[10] World Report on Hearing, World Health Organization, 2021. https://www.who.int/publications/i/item/9789240020481.
[11] Y. Luo, Z. Chen, T. Yoshioka, Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 46–50. https://doi.org/10.1109/ICASSP40776.2020.9054266
[12] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, L. Xie, DCCRN: deep complex convolution recurrent network for phase-Aware speech enhancement, in: Interspeech 2020, 2020, pp. 2472–2476. https://doi.org/10.21437/Interspeech.2020-2537
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, 30, Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[14] Z. Kong, W. Ping, A. Dantrey, B. Catanzaro, CleanUNet 2: a hybrid speech denoising model on waveform and spectrogram, in: Interspeech 2023, 2023, pp. 790–794. https://doi.org/10.21437/Interspeech.2023-1287
[15] L. Yang, W. Liu, R. Meng, G. Lee, S. Baek, H.-G. Moon, FSPEN: an ultra-lightweight network for real time speech enahncment, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 10671–10675.
[16] X. Rong, T. Sun, X. Zhang, Y. Hu, C. Zhu, J. Lu, GTCRN: a speech enhancement model requiring ultralow computational resources, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 971–975. https://doi.org/10.1109/ICASSP48485.2024.10448310
[17] H. Yan, J. Zhang, C. Fan, Y. Zhou, P. Liu, LiSenNet: lightweight sub-band and dual-path modeling for real-time speech enhancement, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10888272
[18] X. Rong, D. Wang, Y. Hu, C. Zhu, K. Chen, J. Lu, UL-UNAS: ultra-lightweight U-nets for real-time speech enhancement via network architecture search, 2025. https://arxiv.org/abs/2503.00340. arXiv:2503.00340[eess.AS]
[19] M. Abdullah Alohali, N. Saleem, D. Rhouma, M. Medani, H. Elmannai, S. Bourouis, Temporally dynamic spiking transformer network for speech enhancement, IEEE Access 12 (2024) 146513–146526. https://doi.org/10.1109/ACCESS.2024.3444596
[20] X. Hao, C. Ma, Q. Yang, J. Wu, K.C. Tan, Toward ultralow-power neuromorphic speech enhancement with spiking-FullSubNet, IEEE Trans. Neural Netw. Learn. Syst. (2025) 1–15. https://doi.org/10.1109/TNNLS.2025.3566021
[21] Y. Du, X. Liu, Y. Chua, Spiking structured state space model for monaural speech enhancement, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 766–770. https://doi.org/10.1109/ICASSP48485.2024.10448152
[22] T. Sun, S. Bohté, DPSNN: spiking neural network for low-latency streaming speech enhancement, Neuromorphic Comput. Eng. 4 (4) (2024) 044008. https://doi.org/10.1088/2634-4386/ad93f9
[23] J. Wang, Z. Lin, T. Wang, M. Ge, L. Wang, J. Dang, Mamba-SEUNet: mamba UNet for monaural speech enhancement, in: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025, pp. 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889525
[24] H. Schröter, N.E.-B. Alberto, T. Rosenkranz, A. Maier, DeepFilterNet: perceptually motivated real-time speech enhancement, in: Interspeech 2023, 2023, pp. 2008–2009.
[25] D. Byrne, H. Dillon, The national acoustic laboratories'(NAL) new procedure for selecting the gain and frequency response of a hearing aid, Ear Hear. 7 (4) (1986) 257–265. https://doi.org/10.1097/00003446-198608000-00007
[26] S. Ahmed, R.E. Zezario, H.-G. Yuan, A. Hussain, H.-M. Wang, W.-H. Chung, Y. Tsao, NeuroAMP: a novel end-to-end general purpose deep neural amplifier for personalized hearing aids, IEEE Trans. Artif. Intell. (2025) 1–16. https://doi.org/10.1109/TAI.2025.3603538
[27] Y. Ni, R. Liang, X. Hao, J. Cheng, Q. Wang, C. Huang, C. Zou, W. Zhou, W. Ding, B.W. Schuller, Affine modulation-based audiogram fusion network for joint noise reduction and hearing loss compensation, Inf. Fusion 127 (2026) 103726. https://doi.org/10.1016/j.inffus.2025.103726
[28] L.F. Abbott, T.B. Kepler, Model neurons: from Hodgkin-Huxley to Hopfield, in: L. Garrido (Ed.), Statistical Mechanics of Neural Networks, Springer Berlin Heidelberg, Berlin, Heidelberg, 1990, pp. 5–18.
[29] D. Tal, E.L. Schwartz, Computing with the leaky integrate-and-fire neuron: logarithmic computation and multiplication, Neural Comput. 9 (2) (1997) 305–318. https://doi.org/10.1162/neco.1997.9.2.305
[30] A comprehensive survey of recent developments in neuronal communication and computational neuroscience, J. Ind. Inf. Integr. 13 (2019) 40–54. https://doi.org/10.1016/j.jii.2018.11.005
[31] W. Fang, Z. Yu, Y. Chen, T. Masquelier, T. Huang, Y. Tian, Incorporating learnable membrane time constant to enhance learning of spiking neural networks, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 2641–2651. https://doi.org/10.1109/ICCV48922.2021.00266
[32] A. Bittar, P.N. Garner, A surrogate gradient spiking baseline for speech command recognition, Front. Neurosci. 16 (2022) 865897. https://doi.org/10.3389/fnins.2022.865897
[33] A.C. Carl, J. Cornejo, Audiology pure tone evaluation (2022). https://europepmc.org/article/nbk/nbk580531.
[34] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh, R. Aichner, ICASSP 2023 deep noise suppression challenge, IEEE Open J. Signal Process. 5 (2024) 725–737. https://doi.org/10.1109/OJSP.2024.3378602
[35] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, G. Weber, Common voice: a massively-multilingual speech corpus, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 4218–4222. https://aclanthology.org/2020.lrec-1.520/.
[36] J. Yamagishi, C. Veaux, K. MacDonald, CSTR VCTK Corpus: english multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive: http://web.ku.edu/idea/readings/rainbow.htm. (2019). https://doi.org/10.7488/ds/2645
[37] G. Wichern, J. Antognini, M. Flynn, L.R. Zhu, E. McQuinn, D. Crow, E. Manilow, J.L. Roux, WHAM!: extending speech separation to noisy environments, in: Interspeech 2019, 2019, pp. 1368–1372. https://doi.org/10.21437/Interspeech.2019-2821
[38] A. Varga, H.J.M. Steeneken, Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems, Speech Commun. 12 (3) (1993) 247–251. https://doi.org/10.1016/0167-6393(93)90095-3
[39] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L.D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S.M. Mousavi, A. Nautsch, H. Nguyen, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperrière, M. Rouvier, R. De Mori, Y. Estève, Open-source conversational AI with SpeechBrain 1.0, J. Mach. Learn. Res. 25 (333) (2024) 1–11. http://jmlr.org/papers/v25/24-0991.html.
[40] C.K.A. Reddy, V. Gopal, R. Cutler, DNSMOS: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6493–6497. https://doi.org/10.1109/ICASSP39728.2021.9414878
[41] W. Zhang, R. Scheibler, K. Saijo, S. Cornell, C. Li, Z. Ni, J. Pirklbauer, M. Sach, S. Watanabe, T. Fingscheidt, Y. Qian, URGENT challenge: universality, robustness, and generalizability for speech enhancement, in: Interspeech 2024, 2024, pp. 4868–4872. https://doi.org/10.21437/Interspeech.2024-1239
[42] C. Valentini-Botinhao, X. Wang, S. Takaki, J. Yamagishi, Investigating RNN-based speech enhancement methods for noise-robust text-to-speech, in: 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), 2016, pp. 146–152. https://doi.org/10.21437/SSW.2016-24
[43] C.K.A. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, J. Gehrke, The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results, in: Interspeech 2020, 2020, pp. 2492–2496. https://doi.org/10.21437/Interspeech.2020-3038
[44] H. Bu, J. Du, X. Na, B. Wu, H. Zheng, AISHELL-1: an open-source mandarin speech corpus and a speech recognition baseline, in: 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA), 2017, pp. 1–5. https://doi.org/10.1109/ICSDA.2017.8384449
[45] J. Bai, M. Wang, H. Liu, H. Yin, Y. Jia, S. Huang, Y. Du, D. Zhang, D. Shi, W.-S. Gan, M.D. Plumbley, S. Rahardja, B. Xiang, J. Chen, Description on IEEE ICME 2024 Grand challenge: semi-supervised acoustic scene classification under domain shift, 2024. https://arxiv.org/abs/2402.02694. arXiv:2402.02694[eess.AS]

[46] T.J. Cox, J. Barker, W. Bailey, S. Graetzer, M.A. Akeroyd, J.F. Culling, G. Naylor, Overview of the 2023 ICASSP SP clarity challenge: speech enhancement for hearing aids, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–2. https://doi.org/10.1109/ICASSP49357.2023.10433922
[47] A.W. Rix, J.G. Beerends, M.P. Hollier, 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), 2, 2001, pp. 749–752. https://doi.org/10.1109/ICASSP.2001.941023
[48] C.H. Taal, R.C. Hendriks, R. Heusdens, 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, 2010, pp. 4214–4217. https://doi.org/10.1109/ICASSP.2010.5495701
[49] Y. Hu, P.C. Loizou, Evaluation of objective quality measures for speech enhancement, IEEE Trans. Audio Speech Lang. Process. 16 (1) (2008) 229–238. https://doi.org/10.1109/TASL.2007.911054
[50] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, T. Ochiai, ESPnet: end-to-end speech processing toolkit, in: Proceedings of Interspeech, 2018, pp. 2207–2211. https://doi.org/10.21437/Interspeech.2018-1456
[51] J. Shi, H.-j. Shim, J. Tian, S. Arora, H. Wu, D. Petermann, J.Q. Yip, Y. Zhang, Y. Tang, W. Zhang, D.S. Alharthi, Y. Huang, K. Saito, J. Han, Y. Zhao, C. Donahue, S. Watanabe, VERSA: a versatile evaluation toolkit for speech, audio, and music, in: N. Dziri, S.X. Ren, S. Diao (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations), 2025, pp. 191–209. https://doi.org/10.18653/v1/2025.naacl-demo.19
[52] S. Wang, A. Maoliniyazi, X. Wu, X. Meng, Emo2Vec: learning emotional embeddings via multi-emotion category, ACM Trans. Internet Technol. 20 (2) (2020). https://doi.org/10.1145/3372152
[53] J.M. Kates, K.H. Arehart, An overview of the HASPI and HASQI metrics for predicting speech intelligibility and speech quality for normal hearing, hearing loss, and hearing aids, Hear. Res. 426 (2022) 108608. https://doi.org/10.1016/j.heares.2022.108608
[54] X. Le, H. Chen, K. Chen, J. Lu, DPCRN: dual-path convolution recurrent network for single channel speech enhancement, in: Proceedings Interspeech 2021, 2021, pp. 2811–2815. https://doi.org/10.21437/Interspeech.2021-296
[55] K. Wang, B. He, 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), 2021, pp. 7098–7102. https://doi.org/10.1109/ICASSP39728.2021.9413740
[56] I. Shchekotov, P.K. Andreev, O. Ivanov, A. Alanov, D. Vetrov, FFC-SE: fast fourier convolution for speech enhancement, in: Interspeech 2022, 2022, pp. 1188–1192. https://doi.org/{10.21437/Interspeech.2022-603}
[57] F. Dang, H. Chen, Q. Hu, P. Zhang, Y. Yan, First coarse, fine afterward: a lightweight two-stage complex approach for monaural speech enhancement, Speech Commun. 146 (2023) 32–44. https://doi.org/10.1016/j.specom.2022.11.004
[58] Y.R. Pei, R. Shrivastava, F. Sidharth, Optimized real-time speech enhancement with deep SSMs on raw audio, in: Interspeech 2025, 2025, pp. 51–55. https://doi.org/10.21437/Interspeech.2025-19
[59] G. FNU, H. Huang, A lightweight dual-path sub-band model for monaural speech enhancement, Available at SSRN (2025). https://doi.org/10.2139/ssrn.5340268
[60] E. Dmitrieva, M. Kaledin, HiFi-stream: streaming speech enhancement with generative adversarial networks, 2025. https://arxiv.org/abs/2503.17141. arXiv:2503.17141[cs.SD]
[61] A. Défossez, G. Synnaeve, Y. Adi, Real time speech enhancement in the waveform domain, in: Interspeech 2020, 2020, pp. 3291–3295. https://doi.org/10.21437/Interspeech.2020-2409
[62] P. Andreev, A. Alanov, O. Ivanov, D. Vetrov, HIFI++: a unified framework for bandwidth extension and speech enhancement, in: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5. https://doi.org/10.1109/ICASSP49357.2023.10097255
[63] Y. Xia, S. Braun, C.K.A. Reddy, H. Dubey, R. Cutler, I. Tashev, Weighted speech distortion losses for neural-network-based real-time speech enhancement, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 871–875. https://doi.org/10.1109/ICASSP40776.2020.9054254
[64] B. Tolooshams, R. Giri, A.H. Song, U. Isik, A. Krishnaswamy, Channel-attention dense U-net for multichannel speech enhancement, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 836–840. https://doi.org/10.1109/ICASSP40776.2020.9053989
[65] Y. Luo, Z. Chen, N. Mesgarani, T. Yoshioka, End-to-end microphone permutation and number invariant multi-channel speech separation, in: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6394–6398. https://doi.org/10.1109/ICASSP40776.2020.9054177
[66] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, D. Wang, TPARN: triple-path attentive recurrent network for time-domain multichannel speech enhancement, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022a, pp. 6497–6501. https://doi.org/10.1109/ICASSP43922.2022.9747373
[67] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, D. Wang, Multichannel speech enhancement without beamforming, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022b, pp. 6502–6506. https://doi.org/10.1109/ICASSP43922.2022.9746704
[68] J. Liu, X. Zhang, DRC-NET: densely connected recurrent convolutional neural network for speech dereverberation, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022, pp. 166–170. https://doi.org/10.1109/ICASSP43922.2022.9747111
[69] D. Lee, J.-W. Choi, DeFT-AN: dense frequency-time attentive network for multichannel speech enhancement, IEEE Signal Process. Lett. 30 (2023) 155–159. https://doi.org/10.1109/LSP.2023.3244428
[70] B.C.J. Moore, J.I. Alcantara, M.A. Stone, B.R. Glasberg, Use of a loudness model for hearing aid fitting: II. hearing aids with multi-channel compression, Br. J. Audiol. 33 (3) (1999) 157–170. https://doi.org/10.3109/03005369909090095
[71] Y. Nejime, B.C.J. Moore, Simulation of the effect of threshold elevation and loudness recruitment combined with reduced frequency selectivity on the intelligibility of speech in noise, J. Acoust. Soc. Am. 102 (1) (1997) 603–615. https://doi.org/10.1121/1.419733