研究与开发

基于深度可分离卷积的MambaUNet的语音增强方法

孟祥彩[1]，庄云朋[1]，钟强[2]，欧世峰[3]
（1. 烟台职业学院智能控制系，山东 烟台 264035；
2. 烟台弘武机电科技有限公司，山东 烟台 264035；
3. 烟台大学物理与电子信息学院，山东 烟台 264005）

摘　要：致力于提升语音增强技术在复杂噪声环境中的鲁棒性与建模效率，提出了一种基于深度可分离卷积与结构化状态空间模型融合的语音增强网络（DW-MambaUNet）。该网络以U-Net结构为基础，引入TF-Mamba模块在时序与频率双路径上建模全局依赖，同时结合深度可分离卷积增强局部特征提取能力，从而实现对语音信号的多尺度特征恢复与精细增强。模型在频谱重建过程中通过可学习Sigmoid与Arctan2函数分别优化幅度与相位输出，在保持参数量较小的前提下大幅提升了语音质量。此外，引入动态权重调节策略，结合损失历史的平滑趋势与语音质量感知评估（perceptual evaluation of speech quality，PESQ）感知反馈机制，自适应平衡多任务损失函数的重要性，有效缓解固定加权方式导致的训练收敛瓶颈。在VoiceBank+DEMAND与TIMIT数据集上的实验结果表明，所提DW-MambaUNet在PESQ、STOI、MOS等多个指标上均优于现有多种主流语音增强模型，尤其在低信噪比条件下表现出良好的增强效果与泛化能力。消融实验进一步验证了TF-Mamba模块与DWConv结构对模型性能的贡献。该研究为低复杂度、高性能的语音增强模型设计提供了新思路，具有良好的理论意义与应用价值。

关键词：语音增强；深度可分离卷积；结构化状态空间；动态权重调节；低信噪比
中图分类号：TN912.3
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025272

A speech enhancement method based on depthwise separable convolution and MambaUNet

MENG Xiangcai[1], ZHUANG Yunpeng[1], ZHONG Qiang[2], OU Shifeng[3]
1. Intelligent Control Department of Yantai Vocational College, Yantai 264035, China
2. Yantai Hongwu Electromechanical Technology Co., Ltd., Yantai 264035, China
3. College of Physics and Electronic Information, Yantai University, Yantai 264005, China

Abstract: Aiming to enhance the robustness and modeling efficiency of speech enhancement technology in complex

收稿日期：2025−05−30；**修回日期：**2025−07−16
通信作者：庄云朋，1042249346@qq.com
基金项目：山东省自然科学基金青年项目（No.ZR2024QB388）；烟台职业学院校本科研项目（No.2024XBYB005）
Foundation Items: The Natural Science Foundation Youth Project of Shandong Province (No.ZR2024QB388), Yantai Vocational College School based Research Project (No.2024XBYB005)

noisy environments, a novel speech enhancement network, DW-MambaUNet was proposed, which integrated depthwise separable convolution and a structured state space model. Built upon a U-Net architecture, the TF-Mamba module was incorporated to model global dependencies along both temporal and frequency paths, while the depthwise separable convolution enhanced local feature extraction. Effective multi-scale feature restoration and fine-grained enhancement of speech signals were enabled by this design. During the spectrogram reconstruction process, learnable Sigmoid and Arctan2 functions were used to separately optimize magnitude and phase outputs, significantly improving speech quality while maintaining a lightweight parameter count. Additionally, a dynamic weight adjustment strategy was introduced that adaptively balanced the importance of multi-task loss functions by leveraging smoothed loss history and PESQ-aware feedback, effectively alleviating convergence bottlenecks caused by fixed-weight schemes. Experimental results on the VoiceBank+DEMAND and TIMIT datasets demonstrate that the proposed DW-MambaUNet outperforms various mainstream speech enhancement models in terms of PESQ, STOI, and MOS metrics, particularly under low signal-to-noise ratio conditions, showing strong enhancement performance and generalization ability. Ablation studies further confirm the effectiveness of the TF-Mamba module and DWConv structure in improving model performance. This study provides a novel perspective for the design of low-complexity and high-performance speech enhancement models, with both theoretical significance and practical value.

Key words: speech enhancement, depthwise separable convolution, structured state space, dynamic weight adjustment, low signal-to-noise ratio

0 引言

本文聚焦于语音增强技术在真实环境下的噪声抑制问题[1]。传统方法如谱减法[2]、滤波法及子空间法[3]等虽广泛应用，但其依赖噪声或语音先验知识的特性，导致在非平稳噪声场景下面临显著挑战。近年来，基于深度神经网络（DNN）的语音增强技术凭借优异的性能表现取得突破性进展[4-5]。

随着循环神经网络（RNN）[6]、卷积神经网络（CNN）[7]和Transformer[8]等架构的演进与跨领域应用成功，语音增强系统性能得到显著提升。早期DNN模型主要采用RNN和CNN：前者通过时序建模捕获语音动态特征，但梯度消失问题制约了长序列依赖学习；后者借助局部感受野提取频谱特征，但全局上下文建模能力不足。为整合二者优势，卷积循环网络（convolutional recurrent network，CRN）应运而生[9]，其编码器–解码器架构采用CNN实现高效特征提取，并通过双向RNN建模帧间相关性。进一步优化的双路径CRN（DPCRN）[10]创新性引入时频域双路径结构，通过并行优化时频域特征表示显著提升去噪性能。当前，U-Net[11]多尺度特征融合、Transformer自注意力机制以及Conformer[12]混合架构等创新方法，正在持续推动语音增强技术向更高精度与更强鲁棒性方向发展。

在损失函数设计层面，传统均方误差（MSE）和$\ell 1$范数损失虽能直接约束频谱重构精度，却未能充分建模人耳听觉感知特性。为突破这一局限，学术界相继提出基于感知评价指标（PESQ、STOI）的监督损失[13]，以及融合生成对抗网络（GAN）的对抗训练范式[14]。其中，GAN框架通过可微分判别器模拟主观听觉评价，驱动生成器产生符合人类感知偏好的增强语音。然而，现有GAN语音增强系统仍存在显著瓶颈：基于Transformer的生成器因自注意力机制导致计算复杂度呈序列长度平方级增长，而标准卷积生成器则受限于局部感受野，难以有效建模长期声学依赖。

针对上述挑战，结构化状态空间模型（structured state space model，SSM）——尤其是Mamba

架构[15]——凭借线性计算复杂度的全局建模能力引发广泛关注。Mamba 通过选择性状态传递机制，在保持序列长度线性计算复杂度的同时，实现对长距离时序依赖的高效捕捉，其性能已在多项长序列任务中超越 Transformer 模型。与此同时，深度可分离卷积（depthwise separable convolution，DWConv）[16]通过解耦空间与通道维度卷积操作，在降低参数量的同时强化局部特征表征能力。然而，如何将 SSM 的全局建模优势与 DWConv 的局部特征提取能力有机融合，构建兼具高效性与鲁棒性的语音增强网络，仍是亟待解决的关键问题。

随着深度学习的广泛应用，深度可分离卷积（DWConv）和结构化状态空间模型（SSM）在图像识别、序列建模等任务中表现出强大的建模能力。将这两种结构引入语音增强任务，不仅仅是简单地借鉴新方法，更重要的是其技术特性在语音建模中具有高度适应性。一方面，DWConv 能够以极低的参数开销提取局部特征，对于语音频谱中微弱细节的恢复尤其有效。另一方面，Mamba 作为一种改进的结构化状态空间模型，在处理长序列语音信号时展现出优异的全局建模能力，可有效替代传统的 RNN 和 Transformer 模型。因此，将 DWConv 和 Mamba 有机结合，能同时兼顾语音增强中的局部细节建模与全局上下文感知。

受这些研究的启发，本文提出了基于 DW-Conv 的 MambaUNet 模型（DW-MambaUNet），这是一个将 Mamba 与 U-Net 集成的模型。通过在 U-Net 框架内利用 TF-Mamba 学习不同分辨率的粗粒度和细粒度信息，并执行多尺度特征融合，该网络展现出增强的远程依赖建模和细节恢复能力。在 VCTK+DEMAND 数据集上进行的广泛实验[17]表明，Mamba SEUNet 实现了最先进的（SOTA）性能，在 FLOP 中测量的计算复杂度显著降低。

1　DW-MambaUNet 模型

1.1　Mamba：选择性状态空间模型

作为一种结构化状态空间模型，Mamba 通过引入选择性状态传递机制，有效扩展了 S4 模型在语音建模中的适应能力。与 Transformer 相比，Mamba 具备线性时间复杂度，在保持长距离建模能力的同时显著降低计算资源消耗。尤其在语音增强任务中，语音信号呈现为高维、长时序的频谱特征，Mamba 通过状态空间映射能够捕捉语音长期动态依赖关系，增强模型对不同时间尺度下语音特征的理解能力。此外，Mamba 具有良好的并行性和硬件感知设计，使其在训练和推理中均具备较高效率。该机制允许通过高维隐藏状态 $h(t)$ 将输入 $x(t)$ 选择性映射到输出 $y(t)$，其可以表示如下：

$$h'(t)=Ah(t)+Bx(t)y(t)=Ch(t) \quad (1)$$

为了应用于离散语音信号，有必要对式（1）中的 A 和 B 进行离散化。具体来说，给定时间尺度参数 Δ，离散参数可以使用零阶保持近似，如下所示：

$$\bar{A}=e^{(\Delta A)},\bar{B}=(\Delta A)^{-1}\left(e^{(\Delta A)}-I\right)\cdot(\Delta B) \quad (2)$$

因此，式（1）可以改写为：

$$h(t)=\bar{A}h(t-1)+\bar{B}x(t)y(t)=\bar{C}h(t) \quad (3)$$

沿着序列进一步扩展式（3）中的计算，可以看出输出 y 是通过核 $\bar{K}$ 的全局卷积从输入 x 计算出来的：

$$\bar{K}=\left(C\overline{B},\overline{AB},\cdots,C\overline{A}^{L-1}\overline{B}\right)y=x\times\bar{K} \quad (4)$$

其中，L 是输入序列的大小。

此外，Mamba 的一个重要贡献是其硬件感知算法，该算法提高了现代硬件上模型执行的效率。核心思想是利用图形处理单元（graphics processing unit，GPU）等现代硬件的内存层次结构，最大限度地减少不同内存级别之间的 I/O 访问。

通过在较快的静态随机存储器（static random access memory，SRAM）中执行离散和递归操作，并将最终结果返回给较慢的高带宽内存（high bandwidth memory，HBM），该算法显著减少了与（B，L，D，N）相关的计算，其中B、L、D和N分别表示批大小、序列长度、通道数量和状态维度，从而提高了计算速度。

1.2 DW-MambaUNet网络结构

DW-MambaUNet架构如图1所示，图1（a）中的逐元素相乘实际运算应为将特征图与相应的掩码或权重矩阵逐元素相乘，该操作的两个输入含义：一是待增强特征（如频谱幅值），二是对应掩码（或注意力权重）。图1中“R”圈表示对张量进行reshape操作，用于调整特征维度匹配。给定含噪语音，首先应用短时傅里叶变换（STFT）来获取幅度谱和相位谱。然后，这些特征被组合并通过特征编码器转换为中间特征。随后，这些中间特征通过分块嵌入和上下采样进入一系列TS-Mamba Block进行特征处理，分块嵌入的目的是使得目标特征在不同分辨率下进行分层处理，而下采样的目的是减少输入到TS-Mamba Block中的信息通道数，从而在降低模型参数量的同时提高训练效率，而上采样的目的是恢复原始通道信息。其中间通过DWConv layer进行连接，能有效提高对局部信息的提取。最后，通过幅度和相位解码器恢复原始的幅度和相位，两者在进行组合和短时逆傅里叶变换（ISTFT）后生成估计语音。接下来，将详细分析网络结构中每个块的结构和作用。

1.3 TS-Mamba Block

在特征学习架构设计层面，本文研究创新性地构建了时序–频率级联的Mamba模块（TS-Mamba），如图1（b）所示。图1中的TF-Mamba模块由“时间Mamba”和“频率Mamba”串联构成，前者建模时序依赖，后者建模频域结构，联合提升时频特征表达能力。为了有效地捕获全局和局部信息，每个Mamba块都采用了文献[18]中提出的双向SSM机制，通过前向与反向扫描路径的并行处理，实现对全局时序动态与局部频域特征的协同捕获。具体而言，输入特征x首先经由前向Mamba分支执行因果状态传递，同步通过反向Mamba分支进行非因果信息聚合。这种双向信息融合机制有效克服传统单向建模的时序信息截断问题，其功能等效于双向长短期记忆（BLSTM）

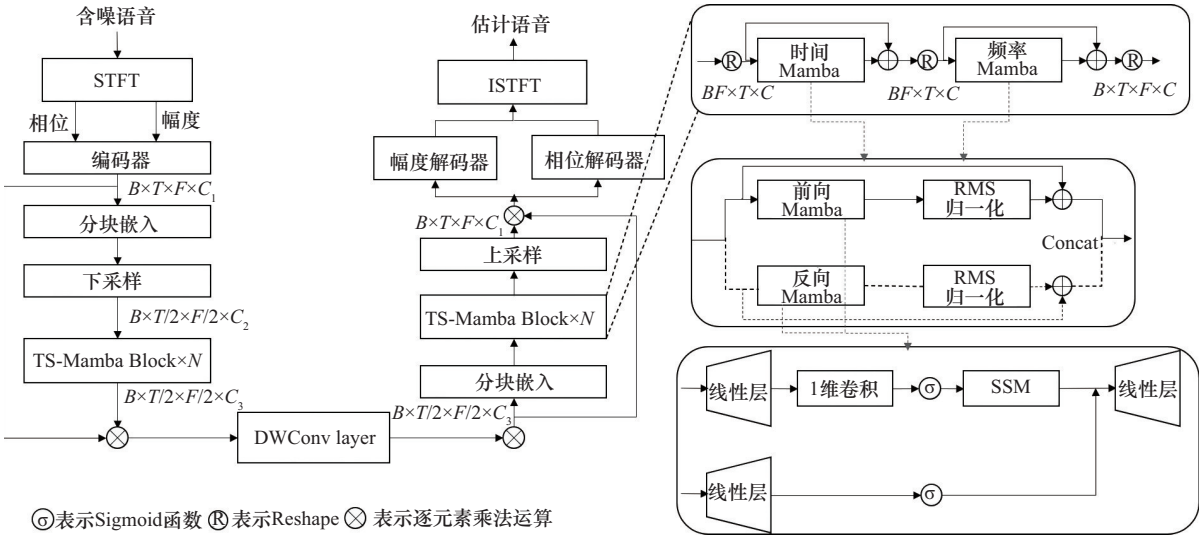

图1 DW-MambaUNet架构

网络[19]的正反向扫描策略，但通过状态空间模型的线性计算复杂度实现更高效的长程依赖建模。在特征处理流程中，前向与反向路径的输出首先分别经过均方根归一化（RMSNorm）[20]实现特征分布稳定性控制，随后通过通道拼接与线性投影层进行跨维度特征融合，最终形成增强后的特征表示x'。该处理流程可形式化表示为：

$$x_f=\mathrm{RMS}\left(\mathrm{FMamba}\left(x\right)\right)+x$$
$$x_b=\mathrm{RMS}\left(\mathrm{BMamba}\left(\mathrm{Flip}\left(x\right)\right)\right)+x$$
$$x'=\mathrm{Linear}\left(\mathrm{Concat}\left(x_f,x_b\right)\right) \tag{5}$$

其中，x_f表示正向Mamba分支输出，x_b表示反向Mamba分支输出，RMS表示均方根归一化，FMamba和BMamba分别表示前向和后向Mamba模块，Flip表示翻转操作，Concat表示通道拼接操作。

前向和后向的Mamba具有相同的结构。具体来说，对于输入序列$x_{\mathrm{in}}\in\mathbb{R}^{L\times C}$，首先应用线性层将其映射到$x'_{\mathrm{in}}\in\mathbb{R}^{L\times 2C}$中。随后是一个核大小为4的卷积运算，再加上一个Sigmoid线性单元（SiLU）激活函数。然后使用式（4）中描述的SSM获得一侧$x'_1\in\mathbb{R}^{L\times 2C}$的输出。为了补偿SSM的顺序约束而导致的任何信息丢失，添加了一个没有卷积和SSM的对称门控分支，由一个额外的线性层和SiLU激活组成。最后，将两个分支的输出连接起来，并通过最后一个线性层进行投影，以获得输出$x_{\mathrm{out}}\in\mathbb{R}^{L\times C}$，如式（6）所示：

$$x_1=\mathrm{SSM}\left(\delta\left(\mathrm{Conv}\left(\mathrm{Linear}\left(x_{\mathrm{in}}\right)\right)\right)\right)$$
$$x_2=\delta\left(\mathrm{Linear}\left(x_{\mathrm{in}}\right)\right)$$
$$x_{\mathrm{out}}=\mathrm{Linear}\left(\mathrm{Concat}\left(x_1,x_2\right)\right) \tag{6}$$

其中，Conv表示1-D卷积运算，δ表示SiLU激活函数，Linear表示线性投影操作，Concat表示级联运算，x_{in}表示输入序列，x_1、x_2表示两分支输出，x_{out}表示最终输出。

1.4　深度可分离卷积线性层（DWConv layer）

DWConv layer结构如图2所示。它以含噪语音的幅度谱和相位谱作为输入，输入最初由1-D卷积层处理，该卷积层由逐帧层归一化（LN）[21]预激活，然后是ReLU激活。其次经过深度可分离卷积，该卷积包括深度卷积（depthwise convolution）：在通道维度上独立进行空间卷积，每个输入通道对应一个独立的卷积核，提取局部频域特征。以及逐点卷积（pointwise convolution）：采用1×1卷积核融合跨通道信息，增强特征的全局表达能力。

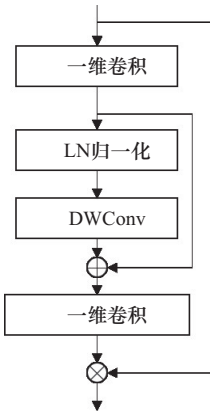

图2　DWConv layer结构

给定输入频谱$X\in\mathbb{R}^{T\times F\times C}$（$T$为时间帧数，$F$为频率点数，$C$为通道数），深度可分离卷积的输出计算如下：

$$Y_{\mathrm{depth}}=\mathrm{DepthwiseConv}\left(X,K_{\mathrm{depth}}\right) \tag{7}$$
$$Y_{\mathrm{point}}=\mathrm{PointwiseConv}\left(Y_{\mathrm{depth}},K_{\mathrm{point}}\right) \tag{8}$$

其中，$K_{\mathrm{depth}}\in\mathbb{R}^{k\times 1\times C}$为深度卷积核（$k$为卷积核大小），$K_{\mathrm{point}}\in\mathbb{R}^{1\times 1\times C\times D}$为逐点卷积核。

D为输出通道数。

1.5　编码器和解码器

DW-MambaUNet编解码器框架的灵感来自MP-SENet中采用的方法[22]。它的特征编码器由两个卷积层和一个扩展的DenseNet组成[23]。第一卷积层将输入通道增加到C1，产生中间特征图，而第二卷积层将频率维度减半以优化计算效率。在这项研究中，扩展DenseNet的深度为4，扩展因子为2，用于捕获语音的基本频谱特征。幅度

和相位解码器都结合了编码器的扩展DenseNet结构，然后是二维转置卷积和1×1卷积。两者的主要区别在于激活函数：幅度解码器使用可学习的Sigmoid函数（LSigmoid）[24]，而相位解码器使用双参数反正切函数（Arctan2）。

1.6 损失函数

DW-MambaUNet旨在在时频（time-frequency，TF）域中优化幅度与相位的重建，以确保增强语音具备较高的感知质量。在训练过程中，模型引入了多组成分的损失函数，并采用动态加权策略，根据训练进度和感知评估分数（如PESQ）自适应地调整各个损失项的贡献。将总损失函数定义为多个损失项的加权求和形式：

$$L_{\text{total}}=\lambda_{\text{metric}}L_{\text{metric}}+\lambda_{\text{mag}}L_{\text{mag}}+\lambda_{\text{phase}}L_{\text{phase}}+\lambda_{\text{com}}L_{\text{com}}+\lambda_{\text{time}}L_{\text{time}} \quad (9)$$

其中，L_{metric}度量判别器损失（metric discriminator loss）：引入判别器网络，通过学习区分真实频谱与模型生成频谱，引导生成结果在感知层面更接近于人耳可接受的真实语音。L_{mag}幅度损失（magnitude loss）：计算估计幅度谱与真实干净幅度谱之间的均方误差（MSE），用于提升幅度重建的准确性。L_{phase}相位损失（phase loss）：包括瞬时相位（instantaneous phase，IP）、群延迟（group delay，GD）和瞬时振幅频率（instantaneous amplitude frequency，IAF）3项指标的组合损失，旨在提高相位恢复的精度。L_{com}复频谱损失（complex spectrogram loss）：计算估计复频谱与干净复频谱之间的MSE，有助于同时优化幅度与相位的协同重建。L_{time}时域损失（time-domain loss）：采用L1损失衡量估计波形与真实波形之间的差异，确保增强语音在时域内的结构一致性。

为了在整个训练过程中有效平衡不同的损失项，采用了一种基于滑动窗口统计和基于PESQ反馈的动态损失加权机制。每个损失项都使用具有窗口大小的移动平均滤波器进行跟踪，以确保稳定的损失变化。采用四分位数范围（IQR）归一化来减轻极端值的影响，并保持稳健的加权。在整个训练过程中，都会对语音质量感知评估（PESQ）分数进行监测。当PESQ较低时，相位损失和复合损失的权重会增加，以强调相位重建。当PESQ提高时，重点就会转向度量损失和幅度损失，以进行微调增强。

2 实验与结果分析

2.1 实验数据与设置

在实验验证部分，本文研究采用VoiceBank+DEMAND[25]与TIMIT双基准数据集构建系统性评估体系，分别承担算法对比与消融分析的核心功能。作为语音增强领域的黄金基准，VoiceBank+DEMAND数据集包含30位说话者的纯净语音（源自VoiceBank语料库）与多环境噪声（选自DEMAND多通道声学噪声数据库）的合成语音对。其标准划分方案如下：训练集集成28位说话者的语音数据与10类环境噪声在{0 dB, 5 dB, 10 dB, 15 dB}信噪比范围内混合，形成11 572组训练样本；测试集则包含2位说话者语音与5类噪声在{2.5 dB, 7.5 dB, 12.5 dB, 17.5 dB}中间信噪比条件下的824组测试样本。这种渐进式信噪比设计有效避免了训练–测试场景的过拟合风险。

TIMIT语料库作为语音处理领域的经典基准，包含630位说话者（覆盖美国八大方言区）录制的6 300条纯净语音，按标准划分为4 620条训练样本与1 680条测试样本。为构建消融实验所需的噪声语音数据集，本文研究采用系统化噪声注入策略：从NOISEX-92噪声库中选取babble、f16、factory1/2、pink、white等6类典型噪声，在{−6 dB, −3 dB, 0 dB, 3 dB, 6 dB}信噪比范围内与纯净语音进行随机组合。每个纯净样本随机绑定单一噪声类型与特定信噪比，最终生成包含31 500组噪声语音的增强型数据集。

2.2 参数设置

本节介绍了实验部分的设置。在比较其他基

线的实验中，使用了 VoiceBank+DEMAND 数据集，并对数据集中的所有语音进行了 16 kHz 的重采样。噪声语音长度设置为 3 s。小于 3 s 的语音重复 3 s，大于 3 s 的音频随机截取 3 s。噪声语音在通过 DW-MambaUNet 模型之前经过短时傅里叶变换（STFT）。傅里叶变换被设置为 320 个采样点、160 个帧偏移点和窗口大小为 320 的汉宁窗口。模型训练优化器选择 Adam 优化器，批大小设置为 24，训练以 5×10^{-4} 的初始学习率开始。如果交叉验证损失值在训练过程中没有连续 3 次降低，则学习率设置为当前学习率的一半，直至训练 100 个迭代周期。

2.3　评估指标

在模型性能评估方面，本文研究构建多维度量化评估体系，覆盖感知质量、可懂度、主观听感及计算效率 4 个关键维度：（1）感知语音质量评估采用宽带 PESQ 指标，该指标通过心理声学模型模拟人类听觉系统特性，在−0.5~4.5 评分区间内量化语音信号保真度；（2）语音可懂度度量选用短时客观可懂度（STOI），其 0-1 标度值反映增强语音在噪声干扰下的语义可理解程度；（3）主观感知评价采用 ITU-T P.835 标准衍生的平均意见得分（MOS）体系，具体包含信号自然度（CSIG，1～5 分）、背景噪声抑制度（CBAK，1～5 分）和整体质量（COVL，1～5 分）3 个子维度；（4）计算效率通过浮点运算量（FLOPS）指标衡量，基于单块 GPU 处理 2 s 时长、16 kHz 采样率的音频样本进行计算。

2.4　结果分析

本文采用的对比模型包括 MetricGAN+[26]、TSTNN[27]、CMGAN[28]、DPCFCS-Net[29]、MUSE[30] 和 MP SENet[31]。不同模型在测试数据集上增强结果见表 1。从表 1 中可以看出，所提 DW-MambaUNet 模型取得了令人印象深刻的性能，PESQ 得分为 3.52，在大多数指标上都优于其他 SE 模型，证明了其强大的去噪能力。并且，与先进的 MP-SENet 相比，所提模型在 PESQ、CSIG 和 COVL 方面得分更高，绝对改善分别为 0.02、0.04 和 0.04 并且仅需要 10.28 GFLOPS 实现比 MP-SENet 更好的性能。

表 1　不同模型在测试数据集上增强结果

方法	GFLOPS	PESQ	STOI	CSIG	CBAK	COVL
带噪语音	—	1.97	0.91	3.35	2.44	2.63
MetricGAN+	—	3.15	—	4.14	3.16	3.64
TSTNN	—	2.96	0.95	4.33	3.53	3.67
CMGAN	63.15	3.41	0.96	4.63	3.94	4.12
DPCFCS-Net	130.65	3.42	0.96	4.71	3.88	4.15
MP SENet	74.29	3.50	0.96	4.73	3.95	4.22
MUSE	9.43	3.37	0.95	4.63	3.80	4.10
DW-MambaUNet	10.28	3.52	0.96	4.77	3.98	4.26

不同 TS-Mamba Block 块数 N 的测试结果见表 2。表 2 中的实验研究了改变 DW-MambaUNet 中 TS-Mamba Block N 的数量对 SE 性能的影响。随着 N 从 1 增加到 3，客观性能指标（包括 PESQ、STOI 和 3 个 MOS 分数）持续改善，证明了添加 TS-Mamba Block 在提高语音质量方面的有效性。

然而，随着 N 进一步增加到 4，虽然 PESQ 和 CBAK 等一些指标略有改善，但其他指标保持不变，表明性能可能趋于平稳。此外，随着 N 的增加，模型参数的数量也会增加，凸显了性能增益和模型复杂性之间的权衡。

表 2　不同 TS-Mamba Block 块数 N 的测试结果

N	PESQ	STOI	CSIG	CBAK	COVL	参数量/10^6
1	3.44	0.96	4.71	3.92	4.21	1.15
2	3.46	0.96	4.75	3.95	4.24	2.10
3	3.50	0.96	4.77	3.97	4.26	2.92
4	3.52	0.96	4.77	3.98	4.26	3.89

为了验证 DW-MambaUNet 结构的有效性和合理性，DW-MambaUNet 通过双向 Mamba 路径建模时频域全局特征，结合 DWConv 加强局部结构感进行消融对比实验。其中，DW-

MambaUNet-1是去除了TF-Mamba的双路径结构改用单路径Mamba代替，DW-MambaUNet-2是去除了中间的DWConv layer，而DW-MambaUNet-3是完整的DW-MambaUNet结构。各消融网络在不同噪声下的PESQ结果见表3，在Babble、F16、Fctory1和Pink 4种噪声源下，DW-MambaUNet-1、DW-MambaUNet-2、DW-MambaUNet-3 3种网络结构的平均PESQ比MambaUNet-3分别下降了0.23、0.25、0.25和0.26。实验结果表明，DW-MambaUNet在PESQ、STOI等多个指标上均优于现有主流模型，且在消融实验中进一步验证了各模块的有效性。上述实验结果验证了所提网络结构的有效性。可知，本文提出的模型有效提升了在复杂噪声环境下的语音建模精度。同时，提出的动态损失加权策略基于滑动窗口平滑、方差调节与PESQ感知反馈，实现了多任务损失的自适应调控，显著改善了语音增强的主观感知质量。

表3 各消融网络在不同噪声下的PESQ结果

噪声类型	消融网络	γSNR/dB			
		−3	0	3	平均
Babble	DW-MambaUNet-1	2.76	2.91	3.09	2.92
	DW-MambaUNet-2	2.96	3.15	3.24	3.12
	DW-MambaUNet-3	3.21	3.35	3.50	3.35
F16	DW-MambaUNet-1	2.71	2.92	3.08	2.9
	DW-MambaUNet-2	2.86	3.14	3.27	3.09
	DW-MambaUNet-3	3.17	3.35	3.51	3.34
Fctory1	DW-MambaUNet-1	2.69	2.84	3.00	2.84
	DW-MambaUNet-2	2.81	3.10	3.20	3.04
	DW-MambaUNet-3	3.15	3.27	3.45	3.29
Pink	DW-MambaUNet-1	2.71	2.87	3.02	2.87
	DW-MambaUNet-2	2.83	3.12	3.21	3.05
	DW-MambaUNet-3	3.17	3.29	3.46	3.31

3 结束语

本文围绕语音增强中长程依赖建模与计算复杂度的平衡问题，提出并实现了DW-MambaUNet模型。该模型结合深度可分离卷积（DWConv）与TF-Mamba结构化状态空间模块，在提升全局建模能力的同时有效压缩参数规模，缓解了传统模型易过拟合与计算冗余的问题。

通过引入双向Mamba路径建模时频域特征，并配合DWConv加强局部结构建模能力，DW-MambaUNet在多个指标上超越现有方法。实验结果证实该模型在低信噪比下具备优异的增强性能和泛化能力。未来工作将聚焦于多通道语音增强、听觉感知建模及跨语言鲁棒性等方向，以进一步拓展该模型的实用性和适应性。

参考文献：

[1] LOIZOU P C. Speech enhancement: theory and practice[M]. Boca Raton: CRC Press, 2007 .

[2] BEROUTI M, SCHWARTZ R, MAKHOUL J. Enhancement of speech corrupted by acoustic noise[C]//Proceedings of the 2003. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '79). Piscataway: IEEE Press, 2003: 208-211.

[3] EPHRAIM Y. Statistical-model-based speech enhancement systems[J]. IEEE, 1992, 80(10): 1526-1555.

[4] ZHENG N J, ZHANG X L. Phase-aware speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(1): 63-76.

[5] TAN X, ZHANG X L. Speech enhancement aided end-to-end multi-task learning for voice activity detection[C]//Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 6823-6827.

[6] JIANG X L, HAN C, MESGARANI N. Dual-path mamba: short and long-term bidirectional selective structured state space models for speech separation[C]//Proceedings of the ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2025: 1-5.

[7] KONG Z F, PING W, DANTREY A, et al. Speech denoising in the waveform domain with self-attention[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7867-7871.

[8] DAO T, GU A. Transformers are SSMs: generalized models and efficient algorithms through structured state space duality[EB]. 2024.

[9] ZHAO H, ZARAR S, TASHEV I, et al. Convolutional-recurrent neural networks for speech enhancement[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2018: 2401-2405.

[10] LUO Y, CHEN Z, YOSHIOKA T. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation[C]//Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2020: 46-50.

[11] MACARTNEY C, WEYDE T. Improved speech enhancement with the Wave-U-Net[EB]. 2018.

[12] ABDULATIF S, CAO R Z, YANG B. CMGAN: conformer-based metric-GAN for monaural speech enhancement[EB]. 2022.

[13] FAN C H, LIU E R, LI A D, et al. BSDB-Net: band-split dual-branch network with selective state spaces mechanism for monaural speech enhancement[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2025, 39(22): 23850-23858.

[14] PASCUAL S, BONAFONTE A, SERRÀ J. SEGAN: speech enhancement generative adversarial network[EB]. 2017.

[15] KU P J, YANG C H, SINISCALCHI S, et al. A multi-dimensional deep structured state space approach to speech enhancement using small-footprint models[C]//Proceedings of the Interspeech 2023. Farmington Hills: Cengage Learning, 2023: 2453-2457.

[16] ZHANG P F, LO E, LU B T. High performance depthwise and pointwise convolutions on mobile devices[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Menlo Park: AAAI Press, 2020, 34(4): 6795-6802.

[17] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]//Proceedings of the 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2010: 4214-4217.

[18] ZHU L H, LIAO B C, ZHANG Q, et al. Vision mamba: efficient visual representation learning with bidirectional state space model[EB]. 2024.

[19] ZHANG S, ZHENG D, HU X, et al. Bidirectional long short-term memory networks for relation classification[C]//Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC). Stroudsburg: ACL, 2015: 73-78.

[20] ZHANG B, SENNRICH R. Root mean square layer normalization[C]//Proceedings of the 33th International Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, 2019: 32.

[21] ZHANG Q Q, SONG Q, NI Z H, et al. Time-frequency attention for monaural speech enhancement[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 7852-7856.

[22] LU Y X, AI Y, LING Z H. MP-SENet: a speech enhancement model with parallel denoising of magnitude and phase spectra[C]//Proceedings of the Interspeech 2023. Farmington Hills: Cengage Learning, 2023: 3834-3838.

[23] HUANG G, LIU Z, VAN DER MAATEN L, et al. Densely connected convolutional networks[C]//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2017: 2261-2269.

[24] DANG F, CHEN H T, ZHANG P Y. DPT-FSNet: dual-path transformer based full-band and sub-band fusion network for speech enhancement[C]//Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2022: 6857-6861.

[25] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]//Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9). Farmington Hills: Cengage Learning, 2016: 146-152.

[26] AI Y, LING Z H. Neural speech phase prediction based on parallel estimation architecture and anti-wrapping losses[C]//Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.

[27] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database (DEMAND): a database of multichannel environmental noise recordings[J]. Acoustics, 2013, 19: 035081.

[28] FU S W, YU C, HSIEH T A, et al. MetricGAN+: an improved version of MetricGAN for speech enhancement[C]//Proceedings of the Interspeech 2021. Farmington Hills: Cengage Learn-

ing, 2021: 201-205.

[29] WANG K, HE B B, ZHU W P. TSTNN: two-stage transformer based neural network for speech enhancement in the time domain[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2021: 7098-7102.

[30] WANG J Y. Efficient encoder-decoder and dual-path conformer for comprehensive feature learning in speech enhancement[C]// Proceedings of the Interspeech 2023. Farmington Hills: Cengage Learning, 2023: 2853-2857.

[31] LIN Z Z, CHEN X T, WANG J Y. MUSE: flexible voiceprint receptive fields and multi-path fusion enhanced Taylor transformer for U-Net-based speech enhancement[EB]. 2024.

[作者简介]

孟祥彩（1992－），女，烟台职业学院智能控制系讲师，主要研究方向为人工智能技术。

庄云朋（1994－），男，博士，烟台职业学院智能控制系讲师，主要研究方向为新能源技术。

钟强（1991－），男，烟台弘武机电科技有限公司工程师，主要研究方向为通信、电子对抗。

欧世峰（1979－），男，博士，烟台大学物理与电子信息学院教授，主要研究方向为语音信号处理。