智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于动态记忆增强的轻量化相位保真语音增强网络

沈学利, 卢呈祥, 崔益烽, 金海波

引用本文:
沈学利, 卢呈祥, 崔益烽, 等. 基于动态记忆增强的轻量化相位保真语音增强网络[J]. 智能系统学报, 2026, 21(3): 802-812.
SHEN Xueli, LU Chengxiang, CUI Yifeng, et al. Lightweight phase-preserving speech enhancement network with dynamic memory augmentation[J]. *CAAI Transactions on Intelligent Systems*, 2026, 21(3): 802-812.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202506018

DOI: 10.11992/tis.202506018
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20260311.1350.006

基于动态记忆增强的轻量化相位保真语音增强网络

沈学利, 卢呈祥, 崔益烽, 金海波
(辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125105)

摘　要: 针对复杂声学场景中低信噪比下的相位失真与噪声适应性难题, 该研究基于幅度−相位显式语音增强框架提出一种改进的语音增强网络。设计记忆增强时频转换器, 利用动态记忆矩阵与门控融合机制提升突发噪声建模能力; 通过稀疏检索机制, 减小模型参数交互规模, 显著降低参数量; 构建任务不确定性驱动的动态损失权重, 协同优化抗包裹相位恢复、复谱重建与感知质量。相较原模型, 改进模型在缩减 9.7% 参数量的同时, 在 VoiceBank+DEMAND 数据集上, 实现低信噪比环境 (−5 dB) 下宽频带语音质量感知质量 (WB-PESQ) 提升 2.3%; 在 DNS Challenge 数据集上, 获得 1.33% 的性能增益, 验证了其在相位保真度与噪声鲁棒性上的有效性。

关键词: 深度学习; 深度神经网络; 智能信息处理; 自然语言处理; 相位优化; 参数优化; 记忆网络; 鲁棒性
中图分类号: TP183; TN912.34　　**文献标志码**: A　　**文章编号**: 1673−4785(2026)03−0802−11

中文引用格式: 沈学利, 卢呈祥, 崔益烽, 等. 基于动态记忆增强的轻量化相位保真语音增强网络 [J]. 智能系统学报, 2026, 21(3): 802–812.
英文引用格式: SHEN Xueli, LU Chengxiang, CUI Yifeng, et al. Lightweight phase-preserving speech enhancement network with dynamic memory augmentation[J]. CAAI transactions on intelligent systems, 2026, 21(3): 802–812.

Lightweight phase-preserving speech enhancement network with dynamic memory augmentation

SHEN Xueli, LU Chengxiang, CUI Yifeng, JIN Haibo
(School of Software, Liaoning Technical University, Huludao 125105, China)

Abstract: To address phase distortion under low-signal-to-noise ratio (SNR) conditions and inadequate noise adaptability in complex acoustic scenes, this study proposes an enhanced speech enhancement network based on an explicit magnitude-phase framework. First, a memory-enhanced time-frequency transformer is designed. It utilizes a dynamic memory matrix and a gated fusion mechanism to improve modeling of impulsive noise. Second, a sparse retrieval mechanism reduces the scale of parameter interaction, thereby significantly reducing model parameters. Finally, a task-uncertainty-driven dynamic loss weighting strategy is developed to jointly optimize anti-wrapping phase restoration, complex spectral reconstruction, and perceptual quality. Compared with the baseline model, the proposed model achieves a 9.7% reduction in parameters while delivering a 2.3% higher wideband perceptual evaluation of speech quality (WB-PESQ) at −5 dB SNR on the VoiceBank+DEMAND dataset and a 1.33% performance gain on the Domain Name System (DNS) Challenge dataset, demonstrating its effectiveness in phase fidelity and noise robustness.

Keywords: deep learning; deep neural networks; intelligent information processing; natural language processing; phase optimization; parameter optimization; memory-augmented networks; robustness

在现实场景中, 语音信号常因环境噪声干扰和混响效应导致失真。语音增强的目标是从带噪信号中恢复纯净语音, 其性能受相位失真与噪声鲁棒性影响显著。传统时频域方法遵循幅度优先原则, 通常将相位视为次要因素, 依赖含噪相位或通过复数谱隐式优化相位。这种处理方式会导致语音谐波结构断裂、听觉质量下降。尽管深度学习技术显著推动了语音增强领域的发展, 但现有方法[1-2]受限于静态参数配置, 难以实现动态噪声抑制, 在语音保真度与噪声消除的均衡控制、相位信息的有效利用等方面仍存在局限性[3], 尤其是对动态噪声抑制能力不足。近期研究虽尝试通过多阶段优化[4]或解耦式架构[5]改善相位恢复

收稿日期: 2025−06−18.　**网络出版日期**: 2026−03−11.
基金项目: 国家自然科学基金项目 (62173171).
通信作者: 沈学利. E-mail: shenxueli@lntu.edu.cn.

效果, 但其固定权重分配策略难以实现幅度−相位动态协同建模, 且在突发性噪声场景下仍存在泛化性能衰减问题。

当前语音增强方法的核心挑战源于相位建模精度与噪声鲁棒性间的固有矛盾。一方面, 相位信息的包裹特性使直接建模面临困难, 传统隐式优化策略往往引发幅度与相位的相互干扰, 加剧高频细节的丢失; 另一方面, 现有经典的显式相位优化架构 (如 speech enhancement model with parallel denoising of magnitude and phase spectra, MP-SEnet[6]) 虽然提高了相位保真度, 但对非平稳噪声的时变特性适应性不足, 且模型冗余度较高, 限制了模型在复杂场景下的泛化能力。与此同时, 多任务优化目标间的动态平衡机制缺失, 导致固定损失权重策略难以兼顾频谱精度与听觉感知需求, 成为制约性能提升的一大瓶颈[7]。为此, 本文提出一种改进的记忆融合相位语音增强网络 (memory fusion phase speech enhancement network, MFP-SEnet), 通过动态记忆引导与多目标自适应优化机制协同提升噪声鲁棒性与相位保真度。该方案在继承显式相位建模优势的基础上, 突破静态架构限制: 一方面利用历史声学场景特征增强瞬态噪声建模能力; 另一方面建立任务驱动的动态平衡策略, 解决频谱精度与感知质量的优化冲突。本文核心贡献包括: 1) 提升模型鲁棒性。通过设计可训练的动态记忆矩阵与门控自适应融合机制, 降低模型对原始信号质量的依赖, 有效建模突发性噪声。2) 压缩模型冗余。引入稀疏检索机制约束记忆交互规模, 在保证性能的前提下显著减少参数量。3) 优化多目标权衡。基于任务不确定性理论构建动态加权网络, 自适应平衡幅度重建、相位连续性与感知质量等优化目标。

1　相关工作

1.1　语音去噪

语音去噪技术旨在从受噪声污染的语音信号中恢复出清晰的语音[8]。传统的去噪方法包括谱减法[9]、维纳滤波[10]、统计建模[11] 以及子空间方法[12]等。这些方法通常依赖于一定的先验知识, 但在应对动态变化的噪声时, 其效果往往有限。

近年来, 随着深度学习技术的迅速发展, 基于深度神经网络 (deep neural network,DNN) 的语音去噪方法[12] 在非平稳噪声抑制方面展现出了显著的优势, 并在语音增强 (speech enhancement, SE) 领域取得了显著进展, 其方法大致分为时域去噪方法和时频域去噪方法两大类。

时域方法[13] 直接处理语音波形, 避免了相位恢复难题, 但在精细刻画频域结构上存在局限, 易引入伪影, 故其性能常逊于时频域方法。时频域方法[14] 通过在频域进行分析与处理, 主要分为映射与掩蔽两大技术路线。映射方法通过建立含噪与纯净语音谱间的复杂函数关系进行增强, 涵盖幅度谱映射、复谱映射及相位映射等[15]; 掩蔽方法则通过构造理想二进制掩蔽 (ideal binary mask, IBM)[16]、理想比率掩蔽 (ideal ratio mask, IRM)[17] 或谱幅度掩蔽 (spectral magnitude mask, SMM)[18] 来抑制噪声。为在掩蔽框架中更有效地利用相位信息, 相位敏感掩蔽 (phase-sensitive mask,PSM)[19] 与复数理想比率掩蔽 (complex ideal ratio mask, cIRM)[20] 等方法被提出以协同优化幅度与相位。此外, 解耦技术[21]、双流幅度[22] 等策略也进一步丰富了时频域的技术框架。

1.2　MP-SEnet[6]

传统时频域语音增强方法通常依赖复谱映射隐式优化相位信息, 难以解决幅度与相位间的补偿问题。MP-SEnet 提出了一种并行幅度−相位显式增强框架, 通过联合优化相位及其时频连续性, 以统一架构支持去噪、解混响以及带宽拓展任务。尽管该模型在显式相位建模与并行解码架构上具有开创性, 但由于静态优化策略与泛化能力不足, 其在复杂场景下的性能上限仍受制约, 该架构在实际应用中仍需要进一步优化。

1.3　基于记忆增强的语音处理

采用静态参数进行建模的传统语音增强模型在应对训练集内分布一致的噪声时表现良好, 但缺乏根据输入信号动态调整内部表示的机制, 导致其泛化能力在面对突发性、瞬态性的未知噪声时, 难以快速适配在训练阶段未曾充分见到的、非平稳的噪声模式。为应对此问题, 能够模仿人类利用过往经验应对新情境的记忆增强网络被引入[23]。通过为模型配备一个外部可寻址的记忆模块, 用于显式地存储和检索丰富的声学原型, 有效提升了模型对非平稳噪声的抑制能力以及在复杂场景下的泛化能力。

1.4　稀疏检索与模型效率优化

记忆增强网络常面临全连接记忆交互导致的计算瓶颈, 制约了其在轻量化场景中的应用。为解决该问题, 稀疏检索机制应运而生。其核心思想是通过仅激活与输入最相关的少量记忆项, 将交互计算复杂度降至常数级别, 从而在保持模型泛化能力的同时实现高效计算。该思路与高效深度学习中的稀疏化策略一脉相承[24], 为构建高性能轻量化记忆网络提供了重要技术路径。

2　本文模型

2.1　模型结构

本文以 MP-SEnet 为基线模型，提出记忆融合相位语音增强网络 (memory fusion phase speech enhancement network, MFP-SMEnet)，旨在克服基线模型静态建模的局限性，提升其在复杂声学场景下的噪声鲁棒性与相位保真度。设计了记忆增强时频域转换器以替代标准 Transformer，通过引入动态记忆矩阵与稀疏检索机制，增强模型对训练集外突发噪声的泛化能力。在特征处理层面，深化了时频双路径的协同机制，利用记忆加权对双路径输出进行特征校正，从而更有效地保障谱结构的连续性，缓解相位失真引发的谐波断裂问题。在优化策略上，构建了任务不确定性驱动的动态加权损失函数，使模型能自适应调整优化目标，在复杂信噪比条件下获得更鲁棒的增强性能。模型整体结构如图 1 所示。

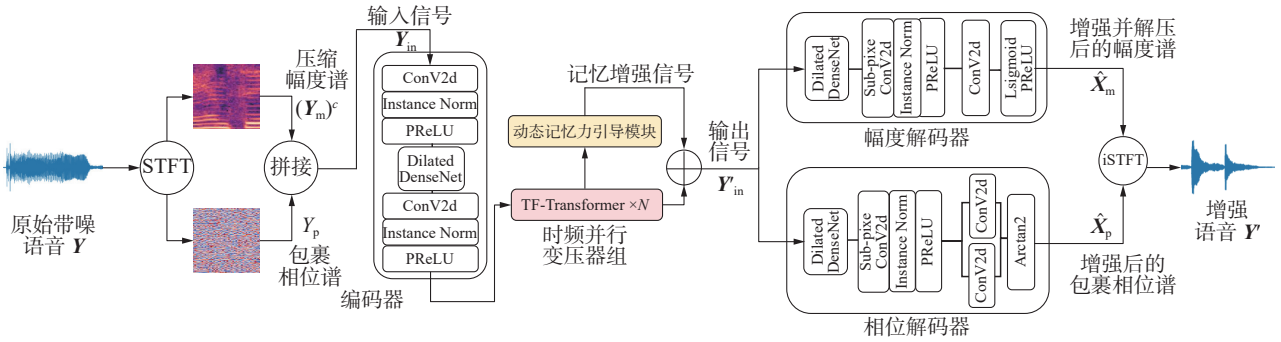


图 1　MFP-SEnet 模型整体结构

Fig. 1　Overall structure of MFP-SEnet

2.2　编码器

本模型编码器沿用基线模型编码器，采用前置卷积层、扩展密集连接网络 (dilated DenseNet) 与后置卷积层三级级联结构，将输入特征 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{T\times F\times 2}$ 转换为通道数为 C、时间维度为 T、频率维度为 $F'=F/2$ 的时频域表示。每个卷积模块由二维卷积层、实例归一化 (instance normalization) 和参数化修正线性单元 (parametric rectified linear unit, PReLU) 激活函数依次构成。其中，首个卷积模块将输入特征的通道数扩展至 C，第 2 个卷积模块则将频率维度 F 压缩至 F'，以减少后续时频转换器模块的计算复杂度。扩展密集网络采用 1、2、4、8 递增扩张率的四层卷积，通过逐步扩大时间轴上的感受野，实现多分辨率上下文信息聚合。每层卷积操作中引入密集连接 (dense connections)，将当前层的输入与所有前置层的输出特征图沿通道维度拼接，从而在缓解梯度消失问题的同时提高了参数效率。后续层卷积的输入由先前层及原始输入的特征拼接而成，确保浅层声学特征 (如基频轮廓) 能够直接传递至深层网络。

2.3　记忆增强时频域转换器

为提升模型对于未知噪声的泛化能力，本文以基于注意力的双路径结构[25] 为基础，提出记忆增强时频域转换器模块，该模块由时频域并行转换器组与动态记忆引导模块组成。

2.3.1　时频域并行转换器组

为克服单一序列建模在捕获语音长时上下文依赖与频域局部相关性上的固有局限，本文采用基于注意力机制的双路径结构。该架构通过将输入特征在时间与频率维度上分别进行重塑与建模，实现互补且增强的时频域特征表示。

该转换器组以经过压缩编码的时频特征张量 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{B\times C\times T\times F'}$ 作为输入，其中 B 为批大小，C 为输入语音的通道数。随后，将该输入同时进行重塑并送入时间转换器与频率转换器中，分别捕获其时间与频率特征。具体而言，将输入的时频特征向量 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{B\times C\times T\times F'}$ 分别重塑为 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{(BF'\times T)\times C}$ 与 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{(BT\times F')\times C}$，并分别送入时间转换器和频率转换器以捕获其时间依赖性和频率依赖性，最终将经过处理的时频域表示重塑为 $\mathbf{R}^{B\times C\times T\times F'}$ 的形式，传递给下一个时频转换器，如此重复进行 N 次。经过处理后的信号将被送往动态记忆引导模块进行增强，所得到的语音信号 $\boldsymbol{Y}'_{\text{in}} \in \mathbf{R}^{B\times C\times T\times F}$ 将作为下一个模块的输入，其中 F 为频率维度。

每个时间与频率转换器均基于门控循环单元 (gated recurrent unit, GRU) 构建，包含多头自注意力机制 (multi-head self-attention, MHSA) 和一个基于 GRU 的位置敏感前馈神经网络 (position-sensitive feed-forward network, FFN)。MHSA 使用 M 个独立注意力头进行协同工作，使模型能够从不同位置的特征表示子空间中整合动态信息。基于 GRU 的位置敏感 FFN 则由双向 GRU 层 (Bi-GRU)、线性整流激活函数 (ReLU) 和线性变换层 (Linear)

构成三层处理结构, 通过时序建模能力对对应的局部特征进行捕获。

2.3.2 动态记忆引导模块

为了提升模型面对突发噪声时的泛化能力, 同时使模型在处理复杂声学场景下的噪声多样性上有更好的表现, 本文提出了基于记忆矩阵[26]的动态记忆引导模块。该机制通过将输入信号与记忆矩阵内信号进行加权融合, 赋予模型动态调用历史声学经验的能力, 从而实现对未知及瞬态噪声的辨识与抑制。该模块的整体结构如图 2 所示。

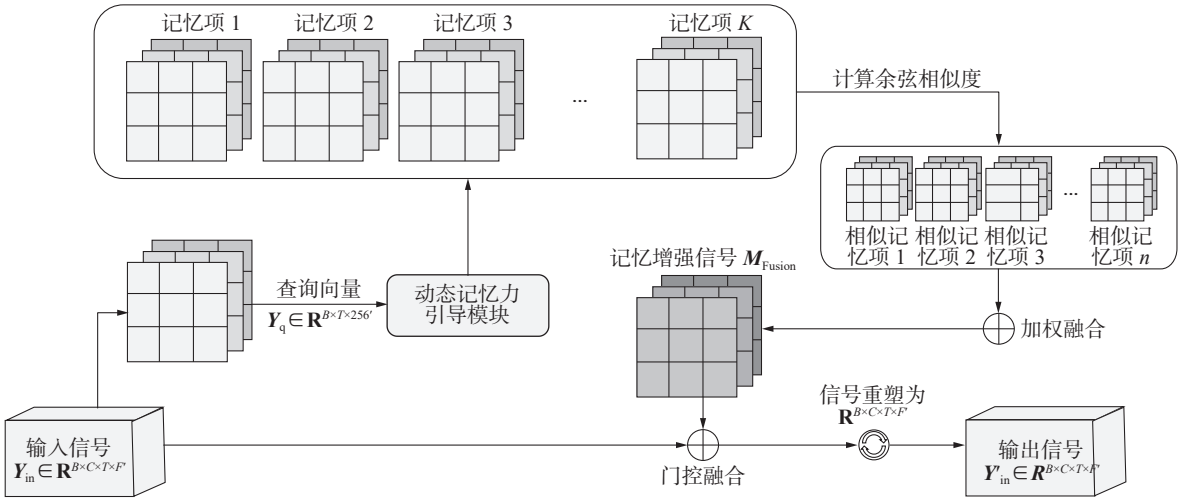


图 2　动态记忆引导模块结构

Fig. 2　Overall structure of dynamic memory guidance module

具体而言, 首先对输入信号 $\boldsymbol{Y}_{\text{in}} \in \mathbf{R}^{B\times C\times T\times F'}$进行全局平均池化和卷积处理, 得到查询向量 $\boldsymbol{Y}_{\text{q}} \in \mathbf{R}^{B\times T\times 256}$, 其表达式为

$$\boldsymbol{Y}_{\text{q}} = \text{Conv1D}_{D=256}(\text{GAP}(\boldsymbol{Y}_{\text{in}})) \tag{1}$$

式中: $\text{Conv1D}_{D=256}$ 表示一维卷积操作, 其输出通道数为 256, 用于进行特征变换; GAP 表示全局平均池化操作, 在频率维度 F'上进行, 将输入 $\boldsymbol{Y}_{\text{in}}$ 从 $\mathbf{R}^{B\times C\times T\times F'}$压缩为 $\mathbf{R}^{B\times C\times 1}$。随后将该查询向量传递给记忆矩阵 $\boldsymbol{M} \in \mathbf{R}^{K\times 256}$($K$ 表示记忆容量), 计算查询向量与记忆项的余弦相似度 s_i:

$$s_i = \frac{\boldsymbol{Y}_{\text{q}} \cdot \boldsymbol{M}_i}{\|\boldsymbol{Y}_{\text{q}}\|\|\boldsymbol{M}_i\|} (i \in [1, K]) \tag{2}$$

为平衡模型性能与计算效率, 本文引入稀疏检索机制。该机制通过 Top-n 选择策略, 仅对相似度最高的 n 个记忆项进行后续处理, 显著降低了计算复杂度。具体而言, 传统的全连接记忆交互需要对所有的记忆项进行完整的相似度计算和权重分配, 计算复杂度为 $O(K\times d)$, 其中 d 为特征维度。而稀疏检索机制首先计算所有记忆项的余弦相似度, 然后仅保留相似度最高的 n 个记忆项进行后续的加权融合, 将计算复杂度降低至 $O(n\times d)$。这一设计使得记忆交互的参数量从与记忆容量 K 成正比的规模缩减至固定常数规模, 在维持模型性能的同时实现了参数量降低。

随后, 通过对最高相似度的 n 个记忆项使用带系数的归一化指数函数生成加权融合权重ω_i:

$$\omega_i = \frac{\exp(s_i/\tau)}{\sum_{k=1}^{n}\exp(s_k/\tau)} \tag{3}$$

再根据该权重对高相似记忆项进行加权融合, 得到记忆增强信号 $\boldsymbol{M}_{\text{fusion}} \in \mathbf{R}^{B\times T\times 256}$:

$$\boldsymbol{M}_{\text{fusion}} = \sum_{i=1}^{n}\omega_i \boldsymbol{M}_i \tag{4}$$

最后将该记忆增强信号重塑为四维形式 $\boldsymbol{M}'_{\text{fusion}} \in \mathbf{R}^{B\times C\times T\times F'}$, 与输入信号 $\boldsymbol{Y}_{\text{in}}$ 进行门控融合得到最终增强信号 $\boldsymbol{Y}'_{\text{in}}$:

$$\gamma = \text{Sigmoid}\left(\boldsymbol{W}_{\text{g}}\left[\boldsymbol{Y}_{\text{in}}; M'_{\text{fusion}}\right]\right) \in [0,1]$$
$$\boldsymbol{Y}'_{\text{in}} = \boldsymbol{Y}_{\text{in}} + \gamma \cdot \boldsymbol{M}'_{\text{fusion}} \tag{5}$$

式中: $\boldsymbol{W}_{\text{g}}$ 为可学习的参数矩阵, 用于将拼接后的特征映射为标量的门控系数, 该参数能够通过主损失函数的反向传播自主更新, 不需要额外的监督。

该模块通过动态记忆引导、时频域双重建模及门控自适应融合的串联特征学习框架, 有效应对噪声多样性挑战与上下文依赖性需求, 实现了对复杂声学模式的高效建模。

2.4　幅度−相位并行解码器

幅度掩码解码器负责从时频转换器模块输出的时频域表示中预测压缩幅度掩码 $\hat{\boldsymbol{M}}_{\text{c}} \in \mathbf{R}^{T\times F}$, 并通过逐点相乘修正输入的压缩失真幅度谱 $(\boldsymbol{Y}_{\text{m}})^c$, 生成增强后的压缩幅度谱 $(\hat{\boldsymbol{Y}}_{\text{m}})^c$。

相位解码器直接预测增强后的包裹相位谱

$(\hat{\boldsymbol{Y}}_{\mathrm{p}})^{c}$, 其主体架构与幅度解码器共享扩张密集网络与反卷积模块, 但在反卷积模块后部署并行估计架构以应对相位包裹问题。具体而言, 通过两条独立的二维卷积层分别输出伪实部 $\hat{\boldsymbol{R}} \in \mathbf{R}^{T\times F}$ 与伪虚部 $\hat{\boldsymbol{I}} \in \mathbf{R}^{T\times F}$, 随后利用双参数反正切函数 (Arctan2) 计算相位值:

$$\hat{\boldsymbol{Y}}_{\mathrm{p}} = \mathrm{Arctan2}\left(\hat{\boldsymbol{I}}, \hat{\boldsymbol{R}}\right)$$

根据现有研究[27] 表明, 该策略通过隐式约束伪实虚部的正交性, 在规避直接回归相位时因 2π 周期性导致的梯度跳变问题上有明显作用, 因而本模块选择沿用基线模型模块。

2.5　损失函数

考虑到幅度−相位并行解码器的结构与各组件间的功能差异, 构建独立的损失函数以改善训练效果。

对于幅度谱的训练, 选择以初始压缩幅度谱 $(\boldsymbol{Y}_{\mathrm{m}})^{c} \in \mathbf{R}^{T\times F}$ 与增强后的压缩幅度谱 $\left(\hat{\boldsymbol{Y}}_{\mathrm{m}}\right)^{c} \in \mathbf{R}^{T\times F}$ 之间的均方误差 (mean square error, MSE) 损失对幅度谱进行显式优化, 其损失函数的表达式为

$$\mathcal{L}_{\mathrm{Mag}} = \mathrm{E}_{(\boldsymbol{Y}_{\mathrm{m}}, \hat{\boldsymbol{Y}}_{\mathrm{m}})}\left[\left\| (\boldsymbol{Y}_{\mathrm{m}})^{c} - \left(\hat{\boldsymbol{Y}}_{\mathrm{m}}\right)^{c} \right\|_{\mathrm{F}}^{2}\right] \tag{6}$$

对于相位谱的训练, 由于存在相位包裹特性, 直接计算原始相位谱 $\boldsymbol{Y}_{\mathrm{p}}$ 与增强后的相位谱 $\hat{\boldsymbol{Y}}_{\mathrm{p}}$ 之间的绝对距离可能无法精准描述其实际距离, 对模型的训练造成干扰。为此, 本模型选择基于反包裹函数的瞬时相位损失函数来替代通过计算绝对距离得出的常规相位损失, 以解决相位包裹特性引起的误差问题。反包裹函数 $f_{\mathrm{AW}}(t)$ 的定义为

$$f_{\mathrm{AW}}(t) = \left| t - 2\pi \cdot \mathrm{round}\left(\frac{t}{2\pi}\right)\right|, t \in \mathrm{R}$$

其中 t 为包裹相位值。该函数通过将包裹相位值映射至连续区间 $[0,\pi]$, 以消除因相位周期性跃变导致的梯度不连续问题。基于反包裹函数, 定义相位谱训练的损失函数为

$$\mathcal{L}_{\mathrm{IP}} = \mathrm{E}_{(\boldsymbol{Y}_{\mathrm{p}}, \hat{\boldsymbol{Y}}_{\mathrm{p}})}\left[\left\| f_{\mathrm{AW}}\left(\boldsymbol{X}_{\mathrm{p}} - \hat{\boldsymbol{X}}_{\mathrm{p}}\right)\right\|_{1}\right] \tag{7}$$

同时, 为了在复数域内对相位谱进一步优化, 本模型在增强相位谱的编码阶段通过双参数正切函数对伪实部 $\hat{\boldsymbol{R}} \in \mathbf{R}^{T\times F}$ 与伪虚部 $\hat{\boldsymbol{I}} \in \mathbf{R}^{T\times F}$ 进行计算, 得出了相位值。相较于原始相位谱的实部 $\boldsymbol{R} \in \mathbf{R}^{T\times F}$ 与虚部 $\boldsymbol{I} \in \mathbf{R}^{T\times F}$, 定义复谱损失函数为

$$\mathcal{L}_{\mathrm{Com}} = \mathrm{E}_{(Y,\hat{Y})}\left[\left\| \boldsymbol{R} - \hat{\boldsymbol{R}}\right\|_{\mathrm{F}}^{2} + \left\| \boldsymbol{I} - \hat{\boldsymbol{I}}\right\|_{\mathrm{F}}^{2}\right] \tag{8}$$

除此之外, 由于语音增强任务的各项评估指标与人类的感知具有较强的关联性, 如语音质量的感知评价体系 (perceptual evaluation of speech quality, PESQ) 和短时客观可理解性 (short-time objective intelligibility, STOI) 等参考指标存在不可微分的特点, 难以直接用作损失函数的构建; 除此之外, 为实现感知质量与频谱精度优化目标的动态平衡, 本文引入并改进了 MetricGAN[1] 框架。

该度量判别器通过迭代式逼近策略, 构建与目标评价指标高度关联的代理损失曲面, 使语音增强模型能够依据该判别器提供的梯度方向进行参数更新, 从而实现对相关质量评价体系的定向优化。基于该度量判别器, 现定义判别器的损失函数为

$$\begin{aligned}\mathcal{L}_{\mathrm{D}} = {} & \mathrm{E}_{X}\left[\left\| D(\boldsymbol{X},\boldsymbol{X}) - 1\right\|_{\mathrm{F}}^{2}\right] + \\ & \mathrm{E}_{(X,\hat{X})}\left[\left\| D\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right) - Q_{\mathrm{PSEQ}}\right\|_{\mathrm{F}}^{2}\right]\end{aligned} \tag{9}$$

式中: $\boldsymbol{X} = \boldsymbol{Y}_{\mathrm{m}} \cdot \mathrm{e}^{\mathrm{j}\boldsymbol{Y}_{\mathrm{p}}} \in \mathbf{C}^{T\times F}$ 表示原始复谱, $\hat{\boldsymbol{X}} = \hat{\boldsymbol{Y}}_{\mathrm{m}} \cdot \mathrm{e}^{\mathrm{j}\hat{\boldsymbol{Y}}_{\mathrm{p}}} \in \mathbf{C}^{T\times F}$ 表示增强后的复谱, D 表示判别器, $Q_{\mathrm{PESQ}} \in [0,1]$ 表示 $\boldsymbol{X}$ 与 $\hat{\boldsymbol{X}}$ 之间的实比例 PESQ 分数。值得注意的是, 判别器的输出分数 $D(\boldsymbol{X})$ 不仅代表了语音的绝对质量, 其方差与不确定性更隐含了当前样本感知质量任务的优化难度。在贝叶斯视角下, 该分数被建模为一个概率分布 $P(Q|\boldsymbol{X})$, 其均值 μ 代表预测的感知质量, 其方差 σ^{2} 则代表判别器对该评价的认知不确定性。

在构建动态损失平衡机制时, 参考多任务学习中不确定性加权方法[28], 假设该感知质量任务的损失服从高斯分布, 其噪声尺度由判别器的不确定性 σ 决定。通过最大化对数似然, 可推导出感知质量任务的损失函数应为

$$\begin{aligned}\mathcal{L}_{\mathrm{Metric}} &= \mathrm{E}_{(X,\hat{X})}\left[\left\| D\left(\boldsymbol{X},\hat{\boldsymbol{X}}\right) - 1\right\|_{\mathrm{F}}^{2}\right] \\ \tilde{\mathcal{L}}_{\mathrm{Metric}} &= \log\sigma\left(D\left(\hat{\boldsymbol{X}}\right)\right) + \frac{1}{2\sigma\left(D\left(\hat{\boldsymbol{X}}\right)\right)^{2}}\mathcal{L}_{\mathrm{Metric}}\end{aligned} \tag{10}$$

式中: $\sigma(D(\hat{\boldsymbol{X}}))$ 表示判别器对当前输入 $\hat{\boldsymbol{X}}$ 的预测不确定性, 而 σ 充当了一个可学习的自适应权重; 当判别器对评估结果不确定时, 权重项 $1/(2\sigma^{2})$ 会减小, 从而自动降低不可靠的感知质量损失 $\mathcal{L}_{\mathrm{Metric}}$ 对整体优化的影响。为简化优化过程并维持稳定性, 将式 (10) 中的常数项与缩放因子融入基础权重 α, 得到感知质量任务损失函数的动态权重 λ_{4}:

$$\lambda_{4} = \frac{\alpha}{2\sigma\left(D\left(\hat{\boldsymbol{X}}\right)\right)^{2}}$$

为确保各项损失在训练初期具有相近的量级, 基础权重 α 的初值设置为 1, 以确保损失函数各个组成部分之间的平衡。后续训练中, 该权重将根据前 N 个训练批次中各损失项的平均比值自动确定, 从而避免任何单一损失主导优化过程。

综上所述, 最终生成器总损失由频谱重建损失与度量损失加权构成:

$$\mathcal{L}_G = \lambda_1\mathcal{L}_{Mag} + \lambda_2\mathcal{L}_{IP} + \lambda_3\mathcal{L}_{Com} + \lambda_4\mathcal{L}_{Metric} \quad (11)$$

模型的最终目的是通过联合优化判别器与生成器的损失函数, 定向提高语音的感知质量, 在训练中实现对语音的主观听感和客观指标的协同优化。

3 实验结果与分析

3.1 数据集

本模型语音去噪任务采用公开数据集 VoiceBank+DEMAND 与 DNS Challenge 2020 进行验证。

VoiceBank+DEMAND 数据集的训练集包含来自 28 名说话人的 11 572 条纯净语音片段, 采样率统一降为 16 kHz, 并与 10 类噪声 (8 类来自 DEMAND 数据库, 2 类人工合成) 在 0、5、10 和 15 dB 信噪比 (signal-to-noise ratio, SNR) 下混合生成含噪样本。测试集则由 2 名未见说话人的 824 条语音构成, 与 5 类未知环境噪声 (如公交车、开放广场等) 在 2.5、7.5、12.5 和 17.5 dB SNR 下混合, 用于评估模型在未知噪声场景下的泛化能力。

DNS Challenge 2020 数据集则包含 500 h 纯净语音 (来自 2 150 名说话人) 与 180 h 噪声数据, 通过官方脚本生成 3 000 h 含噪训练样本 (SNR 范围: −5 ~15 dB), 测试集选用无混响场景下的 4 270 条语音 (20 名说话人), 其 SNR 均匀分布于 0~ 25 dB。

语音去混响任务基于 Reverb Challenge[29] 数据集的单通道子集展开。训练集由 WSJCAM0 纯净语音与多条件合成数据组成, 后者通过 24 组实测房间冲激响应 (room impulse response, RIR), 混响时间 T_{60} 范围: 0.2~0.8 s) 与背景噪声 (固定 SNR=20 dB) 卷积生成。测试集包含两部分数据: 其一为人工合成的模拟混响, 通过在 3 种不同房间 (T_{60} 分别为 0.25、0.5、0.7 s) 内模拟近场 (0.5 m) 与远场 (2 m) 麦克风采集的混响语音, 并添加固定 SNR=20 dB 的背景噪声; 其二为真实会议室录音, 采集于 T_{60}≈0.7s 的会议室环境 (麦克风距离为 1 m 与 2.5 m), 用于评估真实场景性能。

3.2 实验设置

本文的实验基于如下环境: Windows 10, 64 位操作系统, NVIDIA GeForce RTX 3080(10 GB), 软件环境 Python3.11、torch2.2.1、CUDA12.4、numpy1.26。

在进行实验前, 需要对数据进行预处理, 将语音数据采样率统一为 16 kHz。

训练阶段, 模型参数设置为: STFT 窗长 25 ms, 帧移 10 ms, 频率点数 512。批大小设置为 4(初始学习率为 3×10^{-4}), 动态损失权重参数 σ_i 初始化为 0.5。评估指标涵盖语音质量感知评估 (PESQ)、短时客观可懂度 (STOI)、尺度不变信号失真比 (scale-invariant signal-to-noise ratio, SI-SNR) 及主观平均意见分[30](mean opinion score, MOS)。在信号输入时, 所有语音片段均被裁剪为 2 s 长度, 并通过短时傅里叶变换提取时频特征。幅度谱幂律压缩因子 c 初始值设为 0.3。

模型架构方面, 编码器通道数 C 设为 64, 堆叠 4 个记忆增强时频域转换器模块 (N=4), 每个模块的多头自注意力 (MHSA) 层包含 4 个头 (M=4), 训练批大小 B 设置为 4。对于动态记忆力模块, 其稀疏度 n 的设置经对比实验得出: 随着 n 的增加, 模型参数量和推理时间呈现明显增长趋势, 而性能提升在 n>8 后趋于饱和。综合考虑效率与性能的平衡, 最终选择 n=8 作为最优稀疏度配置对输入信号进行记忆增强; 记忆容量 K 设置为 1 024、可学习参数矩阵 $\boldsymbol{W}_g$ 设置为均值为 0、标准差为 0.02 的正态分布矩阵。

在损失函数权重配置方面, 本文通过系统的权重敏感性分析确定了最优参数。具体而言, 固定其他超参数, 分别调整各损失项权重并观察在验证集上的性能变化。幅度损失权重 λ_1 在 0.7~1.0 范围内能有效平衡频谱精度与感知质量, 过低导致 SI-SDR 显著下降, 过高则抑制感知质量优化; 相位损失权重 λ_2 在 0.2~0.4 范围内效果最佳, 能有效缓解相位失真而不引入训练不稳定; 复谱损失权重 λ_3 在 0.05~0.15 内可为模型提供适度的复数域约束, 作为相位恢复的辅助梯度信号, 权重过高容易导致主损失项冗余, 过低则约束不足; 感知损失权重 λ_4 在 0.03~0.08 的小范围内存在明确最优值, 过大会与频谱重建目标冲突。基于上述分析, 模型的总损失函数权重最终确定为 λ_1=0.9、λ_2=0.3、λ_3=0.1、λ_4=0.05, 该配置在验证集上实现了频谱精度与感知质量的最佳平衡。

优化器采用 AdamW[31], 其超参数配置为 β_1=0.8、β_2=0.99, 权重衰减系数 0.01, 初始学习率 0.000 5 并按每训练周期 0.99 的比例衰减。模型总计训练 50 万步以确保充分收敛。

3.3 性能评估

3.3.1 评估指标

在 VoiceBank+DEMAND 数据集上, 采用 5 项客观指标衡量提高语音质量: 宽带语音感知质量[32](wideband perceptual evaluation of speech quality,WB-PESQ, 范围−0.5~4.5)、短时客观可理解性[33](STOI, 范围 0~1) 以及 3 个来自 P.863[34] 的复合指标 (CSIG、CBAK、COVL, 范围 1~5), 分别表征信

号失真 (signal distortion)、背景噪声干扰 (background noise interference) 与整体效果 (overall quality)。

针对 DNS Challenge 数据集，额外引入窄带语音感知质量[35](narrowband perceptual evaluation of speech quality, NB-PESQ) 与尺度不变信噪比[36](scale-invariant signal-to-distortion ratio, SI-SDR)，以量化增强语音的失真程度。所有指标数值越高，表示性能越优。

3.3.2　对比实验

将现有的几种具有代表性的模型与本文提出的 MFP-SEnet 模型在 VoiceBank+DEMEND 数据集中进行性能对比。其中包含代表频域方法的度量生成对抗网络 (metric generative adversarial network, MetricGAN)、双路径 Transformer 融合子带网络 (dual-path Transformer fusion subband network, DPT-FSNet)[37]、三叉戟 Transformer 语音增强网络 (trident Transformer for speech enhancement, TridentSE)[38] 模型；代表混合域方法的相位与谐波感知语音增强网络 (phase-and-harmonics-aware speech enhancement network, PHASEN)、双分支注意力内注意力变换器 (dual-branch attention-in-attention Transformer, DB-AIAT)[39]、基于 Conformer 的度量生成对抗网络 (Conformer-based metric generative adversarial network, CMGAN)[40]、幅度与相位谱并行估计网络 (magnitude and phase spectra estimation network, MP-SEnet) 模型。实验结果如表 1 所示。

表 1　VoiceBank+DEMAND 数据集上对比实验结果表

Table 1　Comparison table of experimental results in VoiceBank+DEMEND dataset

方法	处理域	参数量/10^6	WB-PESQ↑	CSIG↑	CBAK↑	COVL↑	STOI↑
原始噪声	—	—	1.97	3.35	2.44	2.63	0.91
MetricGAN	频域(幅度谱)	—	2.86	3.99	3.18	3.42	—
MetricGAN+	频域(幅度谱)	—	3.15	4.14	3.16	3.64	—
DPT-FSNet	频域(复数谱)	0.88	3.33	4.58	3.72	4.00	**0.96**
TridentSE	频域(复数谱)	3.03	3.47	4.70	3.81	4.10	**0.96**
DB-AIAT	混合域(幅度+复数谱)	2.81	3.31	4.61	3.75	3.96	—
CMGAN	混合域(幅度+复数谱)	1.83	3.41	4.63	3.94	4.12	**0.96**
PHASEN	混合域(幅度+相位)	20.90	2.99	4.21	3.55	3.62	—
MP-SEnet	混合域(幅度+相位)	2.26	3.60	4.81	3.99	4.34	**0.96**
MFP-SEnet	混合域(幅度+相位)	**2.04**	**3.62**	**4.84**	**4.05**	**4.38**	**0.96**

注：加粗数字表示最优结果。

实验结果表明，MFP-SEnet 通过创新的动态记忆增强结构，在参数量显著缩减的前提下，实现了多项指标的提升。

此外，考虑到 WB-PESQ 是一个与幅度相关的数值指标，难以直观地反映人耳听到的实际语音质量，因此引入虚拟语音质量客观听感评估系统[41](virtual speech quality objective listener, ViSQOL) 作为参考。该指标通过计算参考语音与增强语音的频谱−时域联合相似性，生成基于听觉感知的客观平均意见分。该分数越高，表示听者收听到的语音信号质量越好。基于该系统，对不同信噪比环境下 MFP-SEnet 的表现与同样基于混合域处理的 DB-AIAT、CMGAN 模型以及 MP-SEnet 进行对比，其结果如图 3、4 所示。

从图 3、4 得知，改进的显式相位优化结构使模型对原始语音质量的依赖性降低，在低信噪比环境下可以借助动态记忆增强模块实现更高效的信号增强。值得注意的是，通过横向对比，采用频域方法进行处理的语音增强模型，效果整体优于时域方法模型，进一步验证了时频特征建模对去噪任务的重要性。

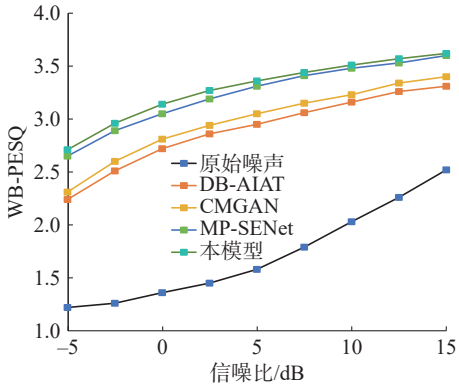


图 3　不同信噪比下各模型的 WB-PESQ 值对比

Fig. 3　Comparison of WB-PESQ values under different signal-to-noise ratios

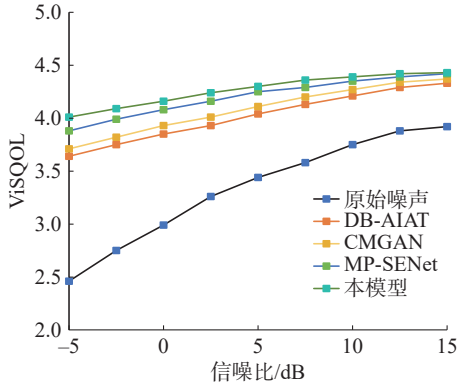


图 4　不同信噪比下 ViSQOL 值对比

Fig. 4　Comparison of ViSQOL values under different signal-to-noise ratios

在 DNS Challenge 数据集中, 与已有的频域基线方法全频带与子带融合网络 (full-band and sub-band fusion network, FullSubNet)[42]、频率循环卷积循环网络 (frequency recurrent convolutional recurrent network, FRCRN)[43]、协同时空网络 (collaborative temporal-spatial network, CTSNet)[44] 等对比, 结果如表 2 所示。MFP-SEnet 在 WB-PESQ、NB-PESQ 与 SI-SDR 上均取得显著优势, 且参数量最小。

表 2　DNS Challenge 数据集上对比实验结果表

Table 2　Comparison table of experimental results in DNS Challenge dataset

方法	参数量/10^6	WB-PESQ	NB-PESQ	STOI/%	SI-SDR/dB
原始噪声	—	1.58	2.45	91.52	9.07
FullSubNet	5.64	2.78	3.31	96.11	17.29
CTSNet	4.35	2.94	3.42	96.21	16.69
TaylorSENet	5.40	3.22	3.59	97.36	19.15
FRCRN	6.90	3.23	3.60	97.69	19.78
MFNet	—	3.43	3.74	97.78	20.31
MP-SEnet	2.26	3.62	3.92	98.16	21.03
MFP-SEnet	**2.06**	**3.68**	**4.01**	**98.32**	**21.31**

注: 加粗数字表示最优结果。

对于语音去混响任务, 使用 Reverb 数据集提供的指标对增强效果进行评估, 包括 WB-PESQ、倒谱距离 (cepstral distance, CD)、对数似然比 (log-Likelihood ratio, LLR)、频率加权分段信噪比 (frequency-weighted segmental signal-to-noise ratio, FWSegSNR) 和信号−混响调制能量比 (signal-to-reverberation modulation energy ratio, SRMR)。其中, 更低的 CD、LLR 值与更高的 FWSegSNR、SRMR 值表示更好的性能。实验的结果如表 3 所示。

表 3　Reverb 数据集上对比实验结果表

Table 3　Comparison table of experimental results in Reverb dataset

方法	PESQ	CD	LLR	FWSeg-SNR	模拟 SRMR	真实 SRMR
混响	1.50	3.97	0.58	3.62	3.69	3.18
WPE	1.72	3.75	0.51	4.90	4.22	3.98
UNet	—	2.50	0.40	10.70	4.88	5.58
SkipConv-GAN	2.91	2.32	**0.23**	11.90	**5.89**	6.36
CMGAN	—	2.25	0.31	11.74	5.47	6.55
DCN	2.94	2.00	**0.23**	13.33	5.27	6.48
MP-SEnet	2.97	**1.97**	0.24	14.07	5.51	6.67
MFP-SEnet	3.04	**1.97**	0.25	**14.22**	5.82	**6.81**

注: 加粗数字表示最优结果。

实验结果表明, MFP-SEnet 在处理语音混响问题上同样表现出了优秀的性能。在模拟混响与真实混响两种测试中, 本文模型在多数指标上取得了最优的成果, 非最优指标与最优值的差距仅为 0.02 与 0.07, 验证了本模型在兼顾传统降噪的任务同时, 能够在一定程度上进行语音解混响处理。

3.3.3　消融实验

为验证模型中各模块的贡献, 本节基于 VoiceBank+DEMAND 数据集进行消融实验, 评估指标涵盖 WB-PESQ、相位距离 (phase distance, PD)、SI-SDR 与 ViSQOL。消融实验的结果如表 4 所示。

表 4　消融实验结果表

Table 4　Results of ablation experiment

消融变体	WB-PESQ	CSIG	CBAK	COVL	STOI	PD/(°)	ViSQOL	SI-SDR/dB	MSE/%
完整模型	3.62	4.84	4.05	4.38	0.96	23.1	4.01	12.3	—
仅幅度增强	2.63	3.48	2.94	2.80	0.89	65.4	2.80	9.8	+35
移除记忆增强模块	3.20	4.45	3.75	3.96	0.93	28.5	2.81	7.7	+15
使用固定权重记忆增强	3.30	4.58	3.72	4.00	0.89	41.7	2.96	11.1	+18
移除动态损失权重	3.54	4.71	3.89	4.15	0.94	26.8	3.72	11.4	+8
仅移除动态记忆矩阵	3.45	4.65	3.82	4.08	0.93	29.3	3.56	10.2	+12
仅移除门控融合	3.58	4.78	3.95	4.25	0.95	32.3	3.85	11.8	+5

当仅进行幅度增强，直接使用含噪相位重构语音时，WB-PESQ 下降 0.91(3.62→2.63)，PD 值飙升 42.3°(23.1°→65.4°)，能够验证相位误差恢复对幅度恢复的补偿效应。

移除记忆增强模块后，模型在低 SNR(−5 dB)场景下 ViSQOL 下降 1.20(4.01→2.81)，SI-SDR 降低 4.6 dB(12.3 dB→7.7 dB)，表明动态语音模式记忆对噪声泛化至关重要。

使用固定权重记忆增强后，观察到 PD 值增加 18.6°(23.1°→41.7°)，高频谐波区域 (>4 kHz) 的 STOI 下降 0.07，证明该设计为增强模型处理高频谐波及其相位优化的能力提供了有效帮助。

移除动态损失权重机制后，WB-PESQ 下降 0.08(3.62→3.54)，SI-SDR 降低 0.9 dB(12.3 dB→11.4 dB)，表明动态权重策略对平衡多目标优化、维持模型高性能的必要性。

仅移除动态记忆矩阵而保留门控融合时，ViSQOL 下降 0.45(4.01→3.56)，证明记忆矩阵在存储噪声原型方面对提升模型泛化能力的核心作用。

仅移除门控融合机制时，PD 值增加 9.2°(23.1°→32.3°)，凸显了门控机制在特征校正与相位保真中的重要性。

消融实验表明，本模型通过显式相位优化机制中复数域相位对齐与抗包裹残差预测的协同设计，显著降低相位失真，有效修复传统方法中高频谐波断裂与共振峰模糊问题；记忆增强模块凭借动态语音模式存储能力，强化了对突发噪声与未知干扰的泛化鲁棒性。而改进的记忆增强时频 Transformer 架构在缩小了参数量的同时维持了高语音质量评分。

4　结束语

本文提出了一种改进的混合域语音增强网络 (MFP-SEnet)，通过显式相位建模与动态记忆引导机制，实现了对语音信号幅度谱与相位谱的协同优化。该模型采用编码器−解码器架构，在数据处理过程中嵌入动态记忆矩阵，通过门控融合策略增强对复杂声学场景的建模能力。针对相位恢复难题，提出复数域正交约束的双路径相位解码方法，有效缓解传统相位回归中的梯度跳变问题。

在 VoiceBank+DEMAND 数据集上，MFP-SEnet 以 2.06×10^6 参数量取得 WB-PESQ 3.65，性能较混合域方法 CMGAN 提升 6.15%；在 DNS Challenge 数据集上实现 SI-SDR 21.31 dB，参数量减少 9.7%。消融实验表明，记忆模块显著改善低信噪比场景的语音质量，显式相位优化结构有效降低相位失真。此外，模型在解混响任务中 SRMR 达 6.81，验证了其多场景适应性。

尽管 MFP-SEnet 在语音增强任务中表现出色，但其计算复杂度与时频变换限制了某些实时场景下的应用，今后将在时域−频域联合建模、优化计算效率等方面进一步研究。

参考文献：

[1] FU S W, LIAO C F, TSAO Y, et al. MetricGAN: generative adversarial networks based black-box metric scores optimization for speech enhancement[C]//International Conference on Machine Learning. Graz: ISCA, 2019: 464−468.

[2] YIN Dacheng, LUO Chong, XIONG Zhiwei, et al. PHASEN: a phase-and-harmonics-aware speech enhancement network[J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(5): 9458−9465.

[3] ZHAO Xiaojia, WANG Yuxuan, WANG Deliang. Robust speaker identification in noisy and reverberant conditions[C]//2014 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2014: 3997−4001.

[4] TAN Ke, WANG Deliang. Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement[J]. IEEE/ACM transactions on audio, speech, and language processing, 2020, 28: 380−390.

[5] HASANNEZHAD M, YU Hongjiang, ZHU Weiping, et al. PACDNN: a phase-aware composite deep neural network for speech enhancement[J]. Speech communication, 2022, 136: 1−13.

[6] LU Yexin, AI Yang, LING Zhenhua. Explicit estimation of magnitude and phase spectra in parallel for high-quality speech enhancement[EB/OL]. (2023−08−17)[2024−06−20]. https://arxiv.org/abs/2308.08926.

[7] REDDY C K A, GOPAL V, CUTLER R, et al. The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results[C]//Interspeech 2020. Graz: ISCA, 2020: 2492−2496.

[8] VANAMBATHINA S, KUMAR T K. Speech enhancement by combining spectral subtraction and auditory masking effect[J]. Microelectronics and computer, 2014, 31(2): 123−128.

[9] 李哲, 王静. 基于门控记忆网络的突发噪声抑制方法[J]. 电子学报, 2023, 51(4): 1205−1214.
LI Zhe, WANG Jing. Impulsive noise suppression via gated memory network[J]. Chinese journal of electronics,

2023, 51(4): 1205−1214.

[10] GAZOR S, ZHANG Wei. Speech enhancement employing Laplacian-Gaussian mixture[J]. IEEE transactions on speech and audio processing, 2005, 13(5): 896−904.

[11] 周明, 张涛, 刘洋. 相位敏感损失函数对语音增强感知质量的影响分析[J]. 信号处理, 2022, 38(8): 1789−1800.
ZHOU Ming, ZHANG Tao, LIU Yang. Impact of phase-sensitive loss functions on perceptual quality in speech enhancement[J]. Journal of signal processing, 2022, 38(8): 1789−1800.

[12] 董娴, 邵玉斌, 杜庆治, 等. 谐波结构相位估计联合幅度补偿的语音增强方法[J]. 重庆邮电大学学报 (自然科学版), 2024, 36(5): 935−944.
DONG Xian, SHAO Yubin, DU Qingzhi, et al. Speech enhancement method combining phase estimation of harmonic structures and amplitude compensation[J]. Journal of Chongqing University of Posts and Telecommunications (natural science edition), 2024, 36(5): 935−944.

[13] 王鹏. 基于深度学习的语音增强方法研究[D]. 太原: 太原理工大学, 2024: 45−60.
WANG Peng. Research on speech enhancement methods based on deep learning[D]. Taiyuan: Taiyuan University of Technology, 2024: 45−60.

[14] 罗笑雪. 时频域单通道语音增强方法研究[D]. 北京: 中国科学院声学研究所, 2023: 30−48.
LUO Xiaoxue. Research on single-channel speech enhancement methods in time-frequency domain[D]. Beijing: Institute of Acoustics, Chinese Academy of Sciences, 2023: 30−48.

[15] 张天骐, 罗庆予, 张慧芝, 等. 复谱映射下融合高效 Transformer 的语音增强方法[J]. 信号处理, 2024, 40(2): 406−416.
ZHANG Tianqi, LUO Qingyu, ZHANG Huizhi, et al. Speech enhancement method based on complex spectrum mapping with efficient transformer[J]. Journal of signal processing, 2024, 40(2): 406−416.

[16] 王亚辉, 张伟, 刘强, 等. 基于理想二进制掩蔽的深度学习语音增强方法[J]. 声学学报, 2021, 46(3): 456−468.
WANG Yahui, ZHANG Wei, LIU Qiang, et al. Deep learning speech enhancement method based on ideal binary mask[J]. Acta acustica, 2021, 46(3): 456−468.

[17] 清华大学人工智能研究院. 轻量化 Transformer 的实时语音增强系统: 中国专利 CN115497128A[P]. 2023−06.
AI Institute of Tsinghua University. Real-time Speech Enhancement System with Lightweight Transformer: Chinese Patent CN115497128A[P]. 2023−06.

[18] LIU Y, CHEN Z. GTCRN: A speech enhancement model requiring ultralow computational resources[J]. IEEE signal processing letters, 2024, 31: 880−884.

[19] HUANG P S, KIM M, HASEGAWA-JOHNSON M, et al. Joint optimization of magnitude and phase for speech enhancement[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2018: 2461−2465.

[20] HU Y X, LIU Y, LV S B, et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement[C]//Proceedings of the 21st Annual Conference of the International Speech Communication Association. Shanghai: ISCA, 2020: 2472−2476.

[21] SHCHEKOTOV I, ANDREEV P K, IVANOV O, et al. FFC-SE: fast Fourier convolution for speech enhancement[C]//Proceedings of the 23rd Annual Conference of the International Speech Communication Association. Incheon: ISCA, 2022: 1188−1192.

[22] LI N, WANG L B, ZHANG Q Q, et al. Dual-stream noise and speech information perception for speech enhancement[J]. Expert systems with applications, 2024, 263: 125432.

[23] LI A, LIU W, LUO Z, et al. HPN: Hearing perception network for speech enhancement with memory mechanism[C]//Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 1−5.

[24] CHILD R, GRAY S, RADFORD A, et al. Generating Long Sequences with Sparse Transformers[EB/OL]. (2019−04−23)[2024−06−20]. https://arxiv.org/abs/1904.10509.

[25] LUO Y, MESGARANI N. 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. New York: IEEE, 2020: 46−50.

[26] WESTON J, CHOPRA S, BORDES A. Memory networks[C]//Proceedings of the 2015 International Conference on Learning Representations. San Diego: ICLR, 2015: 1−15.

[27] 吴迪. 复数域谐波结构重建的相位保真优化[D]. 上海: 上海交通大学, 2024: 55−70.
WU Di. Phase fidelity optimization via complex-domain harmonic structure reconstruction[D]. Shanghai: Shanghai Jiao Tong University, 2024: 55−70.

[28] KENDALL A, GAL Y, CIPOLIA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7482−7491.

[29] DOLBY L. Deep learning-based speech enhancement System: US Patent 11, 876, 321B2[P]. 2025−06−15.

[30] KINOSHITA K, DELCROIX M, GANNOT S, et al. A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research[J]. EURASIP journal on advances in signal processing, 2016, 2016(1): 7.

[31] 陈立, 王浩然, 赵静. 多目标动态加权语音增强算法[J]. 计算机研究与发展, 2024, 61(3): 688−701.
CHEN Li, WANG Haoran, ZHAO Jing. Multi-objective dynamic weighting algorithm for speech enhancement[J]. Journal of computer research and development, 2024, 61(3): 688−701.

[32] LIANG Kaizhao, CHEN Lizhang, LIU Bo, et al. Cautious optimizers: improving training with one line of code[EB/OL]. (2024−05−01)[2024−06−20]. https://arxiv.org/abs/2411.16085.

[33] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//2001 IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings. Piscataway: IEEE, 2002: 749−752.

[34] ITU-T. Perceptual objective listening quality assessment: P. 863[S]. Geneva: ITU, 2014.

[35] ITU-T. Perceptual evaluation of speech quality (PESQ): P. 862[S]. Geneva: ITU, 2001.

[36] 世邦通信股份有限公司. 基于深度学习的语音增强方法及系统: 中国专利 CN119170029A[P]. 2024−08−19.
SHIBANG Information Technology Co. , Ltd. Deep Learning-Based Speech Enhancement Method and System: Chinese Patent CN119170029A[P]. 2024−08−19.

[37] DANG Feng, CHEN Hangting, ZHANG Pengyuan. DPT-FSNet: dual-path transformer based full-band and sub-band fusion network for speech enhancement[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE, 2022: 6857−6861.

[38] YIN D, ZHAO Z, TANG C, et al. TridentSE: Guiding speech enhancement with 32 Global Tokens[C]//Proceedings of the 24th Annual Conference of the International Speech Communication Association. Dublin: ISCA, 2023: 1−5.

[39] ZHANG L, WANG H, LIU Y, et al. DB-AIAT: dual-branch attention-in-attention Transformer for speech enhancement[C]//Proceedings of the 2024 International Conference on Artificial Intelligence and Autonomous Traffic. Singapore: IEEE, 2024: 1234−1239.

[40] ABDULATIF S, CAO Ruizhe, YANG Bin. CMGAN: conformer-based metric-GAN for monaural speech enhancement[J]. IEEE/ACM transactions on audio, speech, and language processing, 2024, 32: 2477−2493.

[41] HU Y, LOIZOU P C. Subjective evaluation of speech enhancement algorithms[J]. IEEE transactions on audio, speech, and language processing, 2008, 16(5): 918−929.

[42] HAO Xiang, SU Xiangdong, HORAUD R, et al. Fullsubnet: a full-band and sub-band fusion model for real-time single-channel speech enhancement[C]//ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 6633−6637.

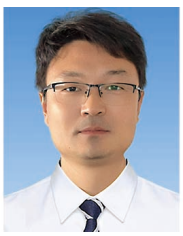
[43] HU M, WANG D L. FRCRN: Boosting feature representation using frequency recurrence for monaural speech enhancement[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 6912−6916.

[44] LI Y, WANG H, ZHANG P. CTSNet: A two-stage mapping mechanism for speech enhancement[C]//2021 IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto: IEEE, 2021: 1234−1238.

作者简介:

沈学利, 教授, 博士, 中国计算机学会杰出会员, 辽宁省人工智能学会副会长, 辽宁工程技术大学软件学院(人工智能学院) 院长, 主要研究方向为智能数据处理、网络信息安全。获省部级科研成果一等奖 1 项、二等奖 2 项、三等奖 4 项, 获省部级教学成果一等奖 1 项、二等奖 1 项, 三等奖 2 项。发表学术论文近百篇。E-mail: shenxueli@lntu.edu.cn。

卢呈祥, 硕士研究生, 主要研究方向为智能数据处理、语音增强技术。E-mail: 2553321250@qq.com。



金海波, 副教授, 博士, 主要研究方向为复杂系统可靠性分析、异常检测、优化维护维修策略制定。E-mail: jinhaibo@lntu.edu.cn。