基于 Transformer-LSTM 架构的语音去噪方法研究

胡必波，刘红英，王传传，甄雅迪
（广州工商学院工学院 广东 广州 510850）

【摘要】在语音识别系统中，语音信号与噪声的叠加使得传统方法难以有效进行语音识别。针对这一问题，本研究提出一种基于 Transformer 的长短期记忆（long short-term memory，LSTM）架构的语音去噪方法，该方法结合 Transformer 的多头注意力机制和 LSTM 的时序建模能力，能够有效地从混合语音信号中去除背景噪声；语音去噪模型采用自适应学习率进行训练优化，基于 WSJ0-Mix 数据集的评估结果表明：基于 Transformer-LSTM 架构的语音去噪方法在信噪比、信号失真比和感知语音质量评估等指标上均优于传统的梯度下降方法，表明 Transformer-LSTM 架构能够在语音去噪任务中提供更为精确的信号恢复和噪声抑制能力，也充分验证了该方法在语音去噪任务中的有效性和优越性。

【关键词】Transformer；长短期记忆（LSTM）；多头注意力机制；语音去噪；自适应学习率
【中图分类号】TN912.34 【文献标识码】A 【文章编号】1009-5624（2025）04-0049-03

0 引言

语音识别技术作为人工智能领域的一个重要分支，近年来在通信、智能家居及人机交互等领域得到广泛应用[1]。然而，实际环境中的语音信号往往受到噪声干扰，严重影响着语音识别系统的性能。因此，语音信号的噪声抑制问题已成为影响语音识别技术发展的关键瓶颈之一。现阶段，主流的语音信号噪声抑制方法主要包括传统的信号处理方法[2]和基于深度学习的方法[3]。传统的信号处理方法如频谱减法和维纳滤波等依赖于先验噪声模型，在处理复杂非平稳噪声环境时表现出较大的局限性。基于深度学习的方法利用神经网络的建模能力，已逐步成为研究热点。其中，卷积神经网络[4]、循环神经网络[5]及近年来兴起的 Transformer 模型[6]在信号去噪任务中表现出优异的性能。然而，这些方法在建模长时间依赖关系或捕捉全局特征时仍存在一定的局限性，例如难以平衡模型复杂性与实时性。

针对现有方法的不足，本研究提出一种结合 Transformer 与长短期记忆（long short-term memory，LSTM）[7]网络的混合架构，用于语音信号的噪声抑制，该方法充分利用 Transformer 在捕捉全局依赖关系方面的优势与 LSTM 在处理时间序列数据时的高效性，有效提高噪声抑制的性能与鲁棒性。此外，本研究通过自适应学习率对语音去噪模型进行训练优化，以确保其实际应用效果。

1 Transformer-LSTM 架构设计

Transformer-LSTM 架构通过捕捉语音信号中的全局和局部特征提升去噪性能。总体架构如图 1 所示，主要包括输入层、位置编码、编码器、解码器、全连接层和输出层模块。

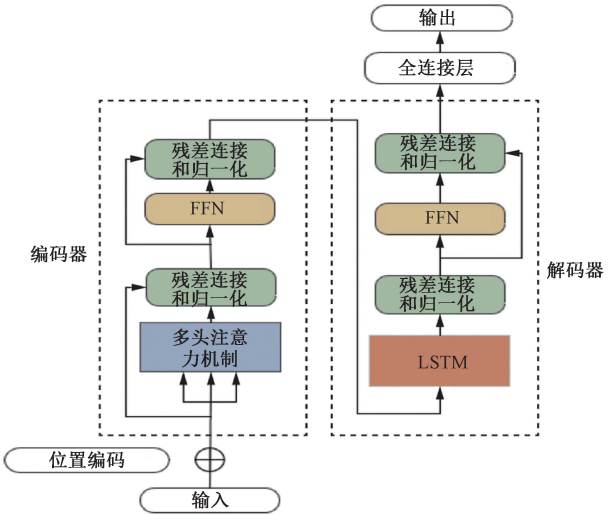

图 1 基于 Transformer-LSTM 架构的语音去噪方法基本原理

输入层的任务是接收语音信号并将其转换为特征表示，由于 Transformer 无法捕捉序列数据的位置信息，因此，需要位置编码模块通过添加固定或可学习的位置向量使模型能够识别特征的时序关系。

编码器模块是 Transformer 结构的核心部分，其中的多头注意力机制通过对输入信号的不同子空间进行独立的特征提取，捕捉语音信号的全局相关性

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{*}}{\sqrt{d_k}}\right)\boldsymbol{V} \quad (1)$$

式中：$\boldsymbol{Q}$ 表示查询矩阵；$\boldsymbol{K}$ 表示键矩阵；$\boldsymbol{V}$ 表示值矩阵；d_k 表示键向量的维度；$\boldsymbol{Q}\boldsymbol{K}^{*}$ 表示计算查询和键之间的相关性。

后续的残差连接通过引入原始输入信号，缓解深层网络中的梯度消失问题。

$$\text{Output} = \text{LayerNorm}(\boldsymbol{X} + \text{SubLayer}(\boldsymbol{X})) \quad (2)$$

式中：$\boldsymbol{X}$ 表示子层输入信号；SubLayer（$\boldsymbol{X}$）表示多头注意力机制或前馈神经网络（feedforward neural network，FNN）的输出；LayerNorm 表示归一化操作。

FNN 由两个线性变换和一个非线性激活函数组成，

基金项目：2023 年度校级教材建设项目（2023JC-01）；2023 年度校级质量工程（大学生校外实践教学基地）项目（XWSJJXJD2023006）。
作者简介：胡必波（1979—），男，湖北天门，本科，副教授，研究方向：计算机应用。

用于对每个时间步的特征进行独立处理。

$$\mathrm{FFN}(x)=\mathrm{ReLU}(x\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \tag{3}$$

式中：x 表示输入特征；$\boldsymbol{W}_1$ 和 $\boldsymbol{W}_2$ 表示权重矩阵；$\boldsymbol{b}_1$ 和 $\boldsymbol{b}_2$ 表示偏置向量；ReLU 表示修正线性单元激活函数。

解码器模块则以 LSTM 为核心，其主要任务是结合编码器的输出进一步捕捉语音信号的时间序列特性。LSTM 通过引入门控机制，有效缓解传统循环神经网络中的长期依赖问题，其输入门为

$$i_t=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},x_t]+\boldsymbol{b}_i) \tag{4}$$

式中：i_t 表示当前时间步的输入门激活值；σ 表示 Sigmoid 激活函数；$\boldsymbol{W}_i$ 表示输入门权重矩阵；$[\boldsymbol{h}_{t-1},x_t]$ 表示前一时间步的隐藏状态和当前时间步输入信号的拼接向量；$\boldsymbol{b}_i$ 表示输入门的偏置向量。

LSTM 的遗忘门为

$$f_t=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},x_t]+\boldsymbol{b}_f) \tag{5}$$

式中：f_t 表示遗忘门激活值；$\boldsymbol{W}_f$ 表示遗忘门权重矩阵；$\boldsymbol{b}_f$ 表示遗忘门的偏置向量。

该模型的输出门工作机制为

$$o_t=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},x_t]+\boldsymbol{b}_o) \tag{6}$$

式中：o_t 表示输出门激活值；$\boldsymbol{W}_o$ 表示输出门权重矩阵；$\boldsymbol{b}_o$ 表示输出门的偏置项。

模型的细胞状态更新方法为

$$C_t=f_t\odot C_{t-1}+i_t\odot\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},x_t]+\boldsymbol{b}_C) \tag{7}$$

式中：C_t 表示当前时间步的细胞状态；tanh 表示双曲正切函数；$\boldsymbol{W}_C$ 表示细胞状态更新的权重矩阵；$\boldsymbol{b}_C$ 表示偏置向量。

最后，模型对隐藏状态进行更新。

$$\boldsymbol{h}_t=o_t\odot\tanh(C_t) \tag{8}$$

式中，$\boldsymbol{h}_t$ 表示当前时间步的隐藏状态。

此外，解码器还包括残差连接、归一化和 FNN，与编码器中的功能类似，用于进一步提升特征提取能力。该部分经过全连接层进行输出，全连接层用于将解码器生成的高维特征映射到目标空间，输出层则通过适当的激活函数生成最终的语音信号估计结果，完成噪声抑制任务。

2　语音去噪模型训练优化

在深度学习模型构建过程中，训练优化是至关重要的环节。训练阶段的核心目标是通过调整网络参数使模型能够有效地从输入数据中提取特征，进而在特定任务上实现预期性能。在噪声抑制任务中，训练优化的有效性直接决定模型对语音信号的去噪效果和泛化能力。

设模型的参数为 $\theta=\{\theta_1,\theta_2,\cdots,\theta_n\}$，其中 θ_i 表示模型的第 i 个参数，n 表示参数的总数。模型训练目标是通过最小化损失函数优化模型参数，损失函数通常采用均方误差度量模型输出与真实语音信号之间的差异，具体为

$$L(\theta)=\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i(\theta))^2 \tag{9}$$

式中：$L(\theta)$ 表示损失函数；y_i 表示第 i 个样本的真实输出值；$\hat{y}_i(\theta)$ 表示模型预测的输出值；N 表示样本数量。

为最小化损失函数，本方法采用梯度下降法更新模型参数，同时，为避免使用固定学习率所带来的训练不稳定或收敛速度慢的问题，本方法采用自适应学习率的方法对每个参数 θ_i 进行更新。

$$v_i(t)=\beta_1 v_i(t-1)+(1-\beta_1)\nabla_{\theta_i}L(\theta_t) \tag{10}$$

$$s_i(t)=\beta_2 s_i(t-1)+(1-\beta_2)\nabla_{\theta_i}L(\theta_t)^2 \tag{11}$$

$$\hat{v}_i(t)=\frac{v_i(t)}{1-\beta_1^t} \tag{12}$$

$$\hat{s}_i(t)=\frac{s_i(t)}{1-\beta_2^t} \tag{13}$$

$$\theta_{i,t+1}=\theta_{i,t}-\frac{\eta}{\sqrt{\hat{s}_i(t)+\epsilon}}\hat{v}_i(t) \tag{14}$$

式中：$v_i(t)$ 和 $s_i(t)$ 分别表示第 i 个参数在第 t 次迭代中梯度的加权平均和梯度的平方加权平均；β_1 和 β_2 分别表示梯度平均和平方梯度平均的衰减系数；ϵ 表示防止除零的平滑项，通常取小值（如 10^{-8}）；$\hat{v}_i(t)$ 和 $\hat{s}_i(t)$ 分别表示对梯度平均和平方梯度平均的偏差进行修正后的值；η 表示全局学习率。

通过自适应调整学习率，模型可以在训练过程中对不同参数赋予不同的学习步长，使得较大梯度的参数更新较快，而较小梯度的参数更新较慢，从而有效避免训练中的梯度爆炸和梯度消失问题。

3　实验结果与分析

WSJ0-Mix 数据集作为广泛应用于语音增强与分离领域的基准数据集，由多个发言人的混合语音数据组成。在基于 Transformer-LSTM 架构的语音去噪方法（本文方法）的去噪性能评估过程中，首先对语音信号进行归一化处理，应用梅尔频谱倒谱系数进行特征提取，为模型输入准备数据。在模型设置方面，Transformer-LSTM 架构的编码器层数设置为 6，注意力头数为 8，LSTM 层的隐藏维度为 512。在训练优化时，设置初始学习率为 1×10^{-3}，每 5 个训练轮次应用 0.9 的衰减因子（共 50 个训练轮次），训练和测试的批量大小均设置为 32。以传统梯度下降法（传统方法）作为对照，以信噪比（signal-to-noise ratio，SNR）、信号失真比（signal distortion ratio，SDR）及感知语音质量评估（perceptual speech quality evaluation，PESQ）作为评估指标量化模型的去噪性能，结果如表 1 所示。

表 1　实验结果表

方法	SNR/dB	SDR/dB	PESQ
本文方法	15.2	12.5	3.7
传统方法	10.4	8.1	2.5

分析表 1 数据可知，本文方法的去噪效果在各项评估指标上均显著优于传统梯度下降法。具体在 SNR 方面，本文方法为 15.2 dB，而传统方法仅为 10.4 dB，表明基于

Transformer-LSTM 架构的语音去噪模型能够在较大的噪声环境下有效地提高语音信号的清晰度,减少噪声对语音质量的干扰。在 SDR 方面,本文方法同样表现出明显的优势,SDR 为 12.5 dB,而传统方法仅为 8.1 dB。SDR 是衡量语音信号去噪效果的重要指标,较高的 SDR 表明去噪后的语音信号在失真程度上得到了更好的控制。PESQ 是一种常用的客观评估指标,旨在模拟人类听觉的感知效果。本文方法的 PESQ 为 3.7,而传统方法的 PESQ 仅为 2.5,表明基于 Transformer-LSTM 架构的语音去噪模型在主观音质上也表现出更强的去噪能力,能够使语音听感更加自然和流畅。

综上所述,基于 Transformer-LSTM 架构的语音去噪方法在 SNR、SDR 和 PESQ 等指标上均优于传统梯度下降法,充分验证了该方法在语音去噪任务中的有效性和优越性。

4 结语

基于 Transformer-LSTM 架构的语音去噪方法能够有效地提高语音信号的质量,相较于传统方法表现出显著的优势。基于 WSJ0-Mix 数据集的实验结果表明,该方法在 SNR、SDR 和 PESQ 等多个方面均取得了优于传统梯度下降法的结果。这表明 Transformer-LSTM 架构能够在语音去噪任务中提供更为精确的信号恢复和噪声抑制能力,具有较好的应用前景。在未来的研究中,可以进一步优化该模型,探索更加高效的训练方法,以及结合其他深度学习技术提升其在复杂环境下的去噪效果。

【参考文献】

[1] 范向民, 范俊君, 田丰, 等. 人机交互与人工智能:从交替浮沉到协同共进[J]. 中国科学(信息科学), 2019, 49(3): 361-368.

[2] 李文志, 屈晓旭. 基于 EEMD 和共振峰的自适应语音去噪[J]. 现代电子技术, 2021, 44(23): 52-56.

[3] 李蕊. 基于深度学习的语音去噪方法研究[D]. 西安: 陕西师范大学, 2021.

[4] 杨帆, 李祎男, 乔涵, 等. 基于深度卷积神经网络的语音信号去噪关键技术研究[J]. 计算机与数字工程, 2022, 50(2): 344-349.

[5] 韩盈, 安志国, 底青云, 等. 基于循环神经网络的大地电磁信号噪声压制研究[J]. 地球物理学报, 2023, 66(10): 4317-4331.

[6] 高志强, 戴琳琳, 景辉, 等. 面向铁路客运站场景的语音降噪模型研究[J]. 铁路计算机应用, 2023, 32(2): 7-12.

[7] SHI J W, WANG S Q, QU P F, et al. Time series prediction model using LSTM-Transformer neural network for mine water inflow [J]. Scientific Reports, 2024, 14(1): 18284.

(上接第 22 页)

4 结语

综上所述,本研究基于 IWGO 算法,通过引入非线性收敛因子,优化了热荷载作用下 FG-GPLRC 层合圆锥壳的半锥角、层数和静载参数,并对优化后结构的动力稳定性进行评估。结果表明,IWGO 算法可显著增强全局搜索能力,提升优化结果的精确性;边界条件为两端固支且 W_G 取 1%、3%、5%条件下,圆锥壳动力稳定性达到最优时,半锥角为 39.67°~56.18°、壳层数为 4~6 层,静载参数为 0;在层合壳内外面掺入越多 GPLs 时,其激励频率越大,即 F-X 壳体的动力稳定性最佳,其次是 F-U、F-O。

【参考文献】

[1] WANG Z Z, WANG T, DING Y M, et al. A simple refined plate theory for the analysis of bending, buckling and free vibration of functionally graded porous plates reinforced by graphene platelets [J]. Mech Adv Mater Struct, 2024, 31(8): 1699-1716.

[2] GOLDFELD Y, ARBOCZ J, ROTHWELL A. Design and optimization of laminated conical shells for buckling [J]. Thin Walled Struct, 2005, 43(1): 107-133.

[3] KARIMI MAHABADI R, BAKHTIARI-NEJAD F. Optimization of joined conical shells based on free vibration [C]//Volume 4B: Dynamics, Vibration, and Control. Phoenix, Arizona, USA: American Society of Mechanical Engineers, 2016: 50558.

[4] HU H T, CHEN H C. Buckling optimization of laminated truncated conical shells subjected to external hydrostatic compression [J]. Compos Part B Eng, 2018, 135: 95-109.

[5] QATU M S, SULLIVAN R W, WANG W C. Recent research advances on the dynamic analysis of composite shells: 2000-2009 [J]. Compos Struct, 2010, 93(1): 14-31.

[6] YANG S W, HAO Y X, ZHANG W, et al. NoNinear dynamic behavior of functionally graded truncated conical shell under complex loads [J]. Int J Bifurcation Chaos, 2015, 25(2): 1550025.

[7] SHU C. Differential quadrature and its application in engineering [M]. London: Springer, 2000.

[8] MIRJALILI S, MIRJALILI S, LEWIS A. Grey wolf optimizer [J]. Adv Eng Softw, 2014, 69: 46-61.

[9] 欧云, 周恺卿, 尹鹏飞, 等. 双收敛因子策略下的改进灰狼优化算法 [J]. 计算机应用, 2023, 43(9): 2679-2685.