2024 届硕士专业学位研究生学位论文

分类号：_____ 学校代码：10269

密　级：_____ 学　　号：71215902102

華東師範大學

East China Normal University

硕士专业学位论文

Master's Degree Thesis (Professional)

基于轻量级神经网络和小波变换的语音降噪方法研究

院　　系： 软件工程学院

专业学位类别： 电子信息

专业学位领域： 软件工程

学位申请人： 李东阳

指导教师： 王江涛　高级工程师

2024 年 4 月 28 日

Thesis for Master's Degree (Professional) in 2024

University Code: 10269

Student ID: 71215902102

EAST CHINA NORMAL UNIVERSITY

Research on Speech Noise Reduction Method Based on Lightweight Neural Network and Wavelet Transform

Department:	Software Engineering Institute
Category:	Electronic Information
Field:	Software Engineering
Candidate:	Li Dongyang
Supervisor:	Wang Jiangtao Senior Engineer

April, 2024

李东阳 硕士学位论文答辩委员会成员名单

姓名	职称	单位	备注
刘晓强	教授	东华大学	主席
王高丽	教授	华东师范大学	
王丽苹	副教授	华东师范大学	

摘　要

近年来，随着语音通信和语音识别技术的快速发展，语音信号的降噪问题越来越受到关注。语音降噪技术是指通过一系列的处理方法将包含噪声的语音信号转换为干净的语音信号，以提高语音通信的质量和语音识别的准确性。在语音降噪领域，深度学习和神经网络技术已经成为一种流行的方法。其中，基于轻量级神经网络的语音降噪方法由于其高效性和可靠性而备受关注。

U-Net 是一种轻量级卷积神经网络，其优点是能够同时处理局部信息和全局信息，并且可以跨层传递信息以提高模型的泛化能力。小波变换则是一种常用的信号分析技术，可以将时域信号转换为时频域信号，并提取出其不同频率成分的特征，在语音降噪任务中有很好的表现。尽管已经有很多基于 U-Net 的语音降噪方法被提出，但是这些方法在实际应用中仍面临着一些挑战。例如，模型的训练和使用的稳定性存在一定问题，同时降噪效果方面也还有一定的提升空间。

本论文旨在提出一种新的基于 U-Net 和小波变换的语音降噪方法，其主要研究内容和创新点如下：

（1）本论文提出了一个多级嵌套的改进 U-Net 网络模型，具体为创新性地在编码器阶段和解码器阶段引入双稳定器，通过阶段性地计算损失，使得每一层的编码里都包含完整的特征信息，实现了更充分的特征提取，并提高了模型的训练效率和稳定性。

（2）为了提高降噪效果，本论文在 U-Net 结构中引入了多尺度特征融合机制。通过在编码器和解码器之间引入多个跨层连接，并将不同尺度的小波分解特征注入连接层参与卷积过程，能够更好地恢复局部高频细节，提高模型的抗噪声能力和降噪效果。

通过上述两大改进，本论文提出的基于 U-Net 和小波变换的语音降噪方法，不仅在理论上进行了创新，而且在与当前一些主流语音降噪技术的对比实验中也表

现出了较好的降噪效果和一定的应用潜力，为语音信号处理领域提供了一种新的解决方案。

关键词：语音降噪、轻量级神经网络、U-Net、小波变换

ABSTRACT

In recent years, with the rapid development of speech communication and speech recognition technology, the issue of noise reduction in speech signals has received increasing attention. Speech denoising technology refers to the process of converting noisy speech signals into clean speech signals through a series of processing methods, in order to improve the quality of speech communication and the accuracy of speech recognition. In the field of speech denoising, deep learning and neural network techniques have become popular methods. Among them, speech denoising methods based on lightweight neural networks have attracted much attention due to their efficiency and reliability.

U-Net is a lightweight convolutional neural network that has the advantage of processing both local and global information simultaneously, and can transmit information across layers to improve the model's generalization ability. Wavelet transform is a commonly used signal analysis technique that can convert time-domain signals into time-frequency domain signals and extract features of their different frequency components, which performs well in speech denoising tasks. Although many U-Net based speech denoising methods have been proposed, these methods still face some challenges in practical applications. For example, there are certain issues with the stability of model training and usage, and there is also room for improvement in noise reduction effectiveness.

This paper aims to propose a new speech denoising method based on U-Net and wavelet transform. The main research content and innovative points are as follows:

(1) This paper proposes a multi-level nested improved U-Net network model, which innovatively introduces dual stabilizers in the encoder and decoder stages. By calculating losses in stages, each layer of encoding contains complete feature information, achieving more comprehensive feature extraction and improving the training efficiency and stability of the model.

(2) In order to improve the noise reduction effect, this paper introduces a multi-scale

feature fusion mechanism in the U-Net structure. By introducing multiple cross layer connections between the encoder and decoder, and injecting wavelet decomposition features of different scales into the connection layer to participate in the convolution process, local high-frequency details can be better restored, and the model's noise resistance and denoising effect can be improved.

Through the above two major improvements, the speech denoising method based on U-Net and wavelet transform proposed in this paper not only innovates in theory, but also demonstrates good denoising effects and certain application potential in comparative experiments with some mainstream speech denoising technologies, providing a new solution for the field of speech signal processing.

Keywords: *Speech Denoising, Lightweight neural networks, U-Net, Wavelet Transform*

目　录

图目录

表目录

第一章 绪 论

1.1 研究背景及意义

语音信号作为一种非常重要的媒介，被广泛应用于人类的日常生活和工作中，例如手机通话、视频会议、语音识别等。但是，在现实应用中，语音信号常常受到背景噪声及录音设备本身的噪声干扰，导致语音质量下降，甚至无法得到有效识别和处理。因此，需要对语音进行降噪处理，将噪声信号从语音信号中去除，以提高语音的质量和可靠性。

1.1.1 研究背景

语音降噪的研究背景可以追溯到二十世纪七十年代，当时降噪技术还处于发展初期。随着科技的不断进步和计算机算力的提高，语音降噪技术也在不断发展和完善。现在，语音降噪已经成为了人们所关注的研究领域之一。在特定行业应用方面，如矿业领域，矿下环境的挑战显著增加了语音降噪的复杂度。矿下环境中的巨大机械噪声，如采煤机和液压支架的运作声，严重影响了矿工使用矿下话机的沟通效率。这种环境下的噪声通常是非平稳的，变化快速且强度大，对语音降噪技术有着更高的要求。

早期的语音降噪方法主要基于统计信号处理和滤波技术，例如均值滤波 [1]、谐波平滑等。这些方法主要通过统计建模和频域分析来减小噪声对语音信号的影响。近年来，随着深度学习技术的快速发展，基于深度神经网络的语音降噪方法也得到了广泛研究和应用。深度神经网络可以学习到语音信号的复杂特征和噪声的统计特性，进一步提高降噪效果。此外，还有许多其他的语音降噪方法被提出和研究，包括基于稀疏表示 [2] 的降噪方法、基于小波变换 [3] 的降噪方法等。

总的来说，语音降噪技术的研究背景源于对语音通信质量的要求和对噪声干扰的认识。通过降低背景噪声对语音信号的影响，可以提高语音质量和可靠性，促进语音通信和语音处理技术的发展。

1.1.2 研究意义

语音降噪技术具有广泛的研究意义和应用价值，可以从以下几个方面来介绍：

（1）提高语音通信质量：在现代社会，人们在日常生活和工作中越来越依赖语音通信，其中电话和视频通话等是最常用的方式之一。然而，在各种环境中进行语音通信时，如开车、公共场合、咖啡厅等，存在着各种噪声干扰，这些噪声会严重影响语音的清晰度和可识别性。因此，通过语音降噪技术可以显著提高语音通信的质量和可靠性。

（2）改善人机交互体验：随着人工智能技术的发展，语音助手和语音识别等人机交互应用正越来越广泛地应用在智能家居、汽车电子和机器人等领域。在这些应用中，语音的清晰度和准确性对于交互效果具有至关重要的作用。通过语音降噪技术，可以提高语音识别准确度，改善人机交互的体验。

（3）促进音频处理技术的发展：除了语音通信和人机交互领域外，语音降噪技术还可以促进音频处理技术的发展，包括音频编解码、音乐合成等。在这些音频处理领域中，语音降噪技术可以改善音频质量，提高音乐的音质，以及减少环境噪声对音频处理的干扰。

（4）推进语音信号处理和机器学习算法的研究：语音降噪技术涉及信号处理、机器学习、深度学习等多个领域，在这些领域中的研究将会推动相关技术和算法的发展，带来更广泛的应用和更大的社会效益。

总之，语音降噪技术具有广泛的研究意义和应用价值。通过降低噪声干扰、提高语音质量和可辨识度，语音降噪技术可以改善语音通信、人机交互和音频处理等各个领域的性能和用户体验，同时推动语音信号处理和机器学习等领域技术的不断进步和发展。

1.2 国内外研究现状

语音降噪的目标是从语音信号中去除或减弱噪声信号，以提高语音的质量和可靠性。随着技术的不断进步，语音降噪的效果和性能也得到了极大的提升。

语音降噪技术经历了从传统信号处理到统计建模，再到深度学习的发展过程。

传统算法的限制主要包括不一致的噪声处理、手动调参和对非线性噪声的限制。而神经网络具有非线性建模能力、自适应性、泛化性能和可扩展性等优点，使其在降噪领域具有显著的优势。

传统的语音降噪算法主要有谱减法 [4]（Spectral Subtraction）、最小均方差 [5]（Minimum Mean Square Error, MMSE）、统计模型法、自适应滤波法 [6]、子空间方法 [7] 等算法。谱减法是一种基于频谱特性的传统语音降噪算法。它的工作原理是通过对观测语音信号的频谱进行处理，将估计的噪声谱从观测语音的频谱中减去，得到估计的干净语音谱，然后通过逆变换获得降噪后的语音信号。谱减法在处理固定噪声类型和噪声强度较低的情况下效果较好，但对于非平稳的噪声（如说话人噪声）和高噪声强度的情况，谱减法存在以下问题：

(1) 声音失真：减去噪声谱会导致语音失真，尤其在低信噪比情况下；

(2) 谱包络失真：对于连续谱变化的语音，谱减法会导致谱包络失真，影响语音质量；

(3) 估计噪声谱困难：准确估计噪声谱困难，会导致估计的干净语音谱中存在噪声残留。

最小均方差（MMSE）算法是一种基于频域的传统语音降噪算法。它通过在频域建模语音信号和噪声，并最小化信号与噪声之间的均方误差来估计和恢复干净语音信号。MMSE 算法利用观测语音信号的频谱特性，分别建模语音和噪声谱的统计特性，通过频域 Wiener 滤波器预估干净语音的谱，然后通过逆变换得到降噪后的语音信号。MMSE 算法在降低噪声和保留语音细节方面相对于谱减法有所改进，但仍存在一些问题：

(1) 误差估计困难：准确估计语音信号和噪声的统计特性常常较为困难；

(2) 算法复杂度较高：MMSE 算法在计算上相对复杂，实时性较差；

(3) 对非平稳噪声的适应性较差：MMSE 算法对于非平稳的噪声环境的适应能力有限。

统计模型法是一类基于语音和噪声的统计模型进行降噪的方法，包括高斯混合模型 [8]（GMM）、隐马尔可夫模型 [9]（HMM）等。统计模型法通过对语音和噪声的统计特性进行建模，通过利用模型参数对观测语音信号进行估计和恢复。其

中，GMM 常用于建模语音和噪声的概率分布，HMM 则用于建模语音和噪声的时序特性。统计模型法在一些特定的噪声环境下，如白噪声、定点噪声等，具有一定的降噪效果。但在复杂的噪声环境下，模型参数的准确估计成为挑战。此外，统计模型法往往需要大量的训练数据。

自适应滤波法是一种利用自适应滤波器对观测信号进行处理的传统语音降噪算法。它的工作原理是通过对输入信号进行滤波，根据观测信号和预测误差的关系来估计和抑制噪声。自适应滤波法利用自适应滤波器根据观测信号和预测误差的相关性来更新滤波器系数，从而在滤波过程中适应并抑制噪声。其中，时域自适应滤波常使用最小均方差（LMS）算法或其变种，频域自适应滤波包括迭代自适应滤波和块自适应滤波等。自适应滤波法在处理非平稳和强噪声环境方面相对较好。然而，自适应滤波算法需要根据具体噪声环境进行模型训练和滤波器系数估计，这导致算法的实时性较差。此外，自适应滤波器的性能高度依赖于噪声相关性和预测误差的准确估计。

子空间方法是一种基于信号和噪声的子空间特性进行降噪的传统算法。这些方法通过对信号和噪声的子空间进行抽取和投影，实现对噪声的抑制和语音信号的恢复。子空间方法通常使用主成分分析（PCA）或独立分量分析（ICA）等技术，将输入信号分解为语音子空间和噪声子空间，然后对噪声子空间进行估计和投影处理，以实现噪声的抑制。子空间方法在一些特定噪声环境下（如定点噪声）表现较好。但对于非平稳噪声的处理和子空间抽取的准确性，子空间方法存在一定的限制。此外，子空间方法在实际应用中可能受到信号和噪声子空间不完全分离等问题的影响。

总体来说，传统语音降噪算法在一定程度上能够抑制噪声，提高语音质量。但这些算法在复杂噪声环境下的降噪效果有限，且对于非平稳噪声的处理能力较弱。随着深度学习等新技术的发展，越来越多的研究关注利用神经网络等方法进行语音降噪，并获得了更好的降噪效果。

神经网络在语音降噪领域的发展经历了多个阶段和技术突破。首先是基于深度神经网络 (DNN) 的降噪方法的出现给语音降噪领域带来了重大突破。研究人员开始使用深度神经网络来学习语音信号的噪声模型，并通过训练来估计和减少噪

声成分，从而实现语音降噪。Xia 等人 [10] 在 2015 年提出了一种新的基于噪声约束最小二乘估计的递归神经网络卡尔曼滤波器用于语音降噪。首先利用所提出的递归神经网络对自回归过程建模的语音信号的参数进行估计，然后通过卡尔曼滤波对语音信号进行恢复。所提出的递归神经网络对噪声约束估计是全局渐近稳定的。由于噪声约束估计对非高斯噪声具有鲁棒性，因此上述学者所提出的基于递归神经网络的语音增强算法可以最小化卡尔曼滤波器参数在非高斯噪声中的估计误差。此外，上述学者所提出的基于神经网络的语音增强算法具有低维模型特征，其速度比现有的基于递归神经网络的语言增强算法更快。中国科学技术大学语音与语言信息处理国家工程实验室 [11] 在 2015 年提出了提出一种基于深度神经网络的有监督的语音增强方法，与传统的基于最小均方误差（MMSE）的降噪技术相比，该方法通过寻找噪声和干净语音信号之间的映射函数来增强语音。为了能够在现实世界中处理广泛的附加噪声，研究首先设计了一个大型训练集，该训练集包括语音和噪声类型的许多可能组合。然后采用 DNN 架构作为非线性回归函数，以确保强大的建模能力。还提出了几种技术来改进基于 DNN 的语音增强系统，包括全局方差均衡以缓解回归模型的过平滑问题，以及丢弃和噪声感知训练策略以进一步提高 DNN 对未知噪声条件的泛化能力。在 2016 年，由以色列的 Romat-Gan[12] 将 Gaussians-MoG 模型的生成混合与判别深度神经网络 DNN 相结合，提出了一种混合方法, 应用于单通道语音增强。所提出的算法分两个阶段执行，即不重复的训练阶段和测试阶段。首先，将无噪声语音对数功率谱密度建模为 MoG，表示语音信号中基于音素的分集。然后使用干净语音信号的音素标记数据库来训练 DNN，用于以梅尔频率倒谱系数作为输入特征的音素分类。在测试阶段，对未经训练的语音的嘈杂话语进行处理。给定噪声语音话语的音素分类结果，使用生成模型和判别模型两者来获得语音存在概率 SPP。然后将 SPP 控制的衰减应用于噪声语音，同时更新噪声估计，最终呈现出较好的降噪效果。同年，Wei HAN 等人 [13] 提出了一种基于 DNN 的感知激励单通道语音增强方法。该研究考虑到人类听觉系统良好的掩蔽特性，提出了一种新的 DNN 结构来减少残余噪声的感知效应。这种新的 DNN 架构被直接训练来学习增益函数，该增益函数用于估计干净语音的功率谱，同时对残余噪声的频谱进行整形。该研究也通过实验表明，无论噪声条件是否包括在

训练集中，当用被各种类型的噪声破坏的 TIMIT（英语语音库）进行测试时，所提出的感知激励语音增强方法都可以获得更好的客观语音质量。

循环神经网络（RNN）的引入进一步改进了语音降噪的效果。RNN 可以更好地处理时序数据，并在模型中引入长期依赖信息。这些性质使得 RNN 成为处理语音信号的有力工具。在 2019 年，一些研究人员 [14][15][16] 开始使用 RNN 进行语音降噪处理，并取得了不错的效果。综合来讲，RNN 应用于语音降噪有以下几个显著的优点：

（1）RNN 被设计用于处理时序数据，对于语音信号这样的时域数据非常适用。RNN 通过使用循环连接来传递信息，并可以捕捉到语音信号的时序关系，例如音频中的短时相干性和语音片段的持续性，从而更好地建模和恢复语音信号。

（2）RNN 中的隐含状态可以记忆先前的信息，并在后续的时间步骤中自动更新和传递。这种记忆能力使得 RNN 能够建模长期依赖关系，这对于语音降噪任务非常重要。在复杂的噪声环境中，语音信号的短时特征可能无法提供足够的信息，长期依赖关系可以帮助 RNN 更好地还原真实的语音信号。

（3）RNN 具有灵活的架构，可以通过添加或修改循环单元的数量和类型来适应不同的任务和数据要求。这使得研究人员可以根据特定的语音降噪问题进行模型设计和优化，以获得更好的性能。

尽管 RNN 在语音降噪中具有许多优势，但也存在一些限制。比如：

（1）RNN 的训练通常需要处理梯度消失或梯度爆炸的问题，这是由于长期依赖关系的存在。这导致了训练的困难，尤其是在处理更长的语音片段时。

（2）RNN 的计算复杂性受其递归性质的影响。在每个时间步上，RNN 的计算都涉及到先前时间步的计算结果作为输入，这种递归关系导致计算在时间上的依赖性，使得并行计算变得困难。与其他前馈神经网络相比，RNN 需要按照时间顺序进行计算，因此在处理大量数据时，计算时间会显著增加。另外，由于 RNN 在时间上的依赖性，反向传播算法可能面临梯度消失或梯度爆炸的问题。在反向传播过程中，梯度信息需要沿着时间步骤反向传播，而在训练过程中，梯度可以在递归传播过程中指数级地增大或减小。梯度消失或梯度爆炸都会导致训练的困难，因为梯度无法有效传播或权重更新过大，模型无法收敛。而且，在每个时间步上，

RNN 都需要进行参数更新和状态更新，同时需要处理实际输入数据。随着时间步数的增加，计算量会呈线性增长。当处理长序列或大量数据时，计算复杂度将迅速增加，这可能导致训练和推理过程变得非常耗时。

（3）传统的 RNN 只能考虑到当前时间步之前的信息，无法处理超过其上下文窗口的长期依赖关系。这限制了其对深层次语境的建模能力。

为了解决 RNN 的前两点问题，一些改进的 RNN 结构被提出，如长短期记忆网络 [17]（LSTM）和门控循环单元 [18]（GRU）。例如，Zhenqing 等人 [17] 在 2024 年提出了一种沙漏形 LSTM，它能够在不丢失数据的情况下通过降低特征分辨率来捕获长期时间相关性。该研究在非相邻层中使用了跳跃连接，以避免梯度衰减。此外，在跳跃连接中加入了注意力过程，以强调基本的光谱特征和光谱区域。研究所提出的 LSTM 模型不使用未来的信息，从而产生了适合实时处理的因果系统，对于神经网络降噪的实时性应用也提供了参考价值。而门控循环单元（GRU）是一种改进的循环神经网络（RNN）结构，它通过引入门控机制，可以帮助缓解了 RNN 在语音降噪上的一些劣势。具体表现为 GRU 通过使用门控机制，特别是更新门和重置门，有效地调整梯度的流动。更新门可以控制传递前一时间步的隐含状态信息，以平衡新信息的进入，从而减轻梯度消失和梯度爆炸的问题。其次，GRU 通过引入重置门，可以选择性地忘记或保留先前的信息。这使得 GRU 能够更好地处理时序数据中的长期依赖性，尤其在复杂的噪声环境中，可以更好地恢复真实的语音信号。而且，GRU 具有更少的门控单元和参数。这使得 GRU 在计算复杂性上相对较低，对于语音降噪任务而言更具实用性。GRU 在减少计算开销的同时，仍然能够保持较好的性能。

为了解决 RNN 的第三点问题，一些改进的结构被引入，如双向 RNN 和 Transformer[19] 等。双向 RNN 通过引入第二个反向的 RNN 来处理时间上的依赖性。除了前向传播的隐藏状态，反向传播的隐藏状态也被捕捉到，这样就可以同时考虑过去和未来的信息。这种双向传播的方式使得模型能够更全面地理解序列数据，从而提高了对上下文信息的捕捉能力。然而，双向 RNN 仍然受到 RNN 的一些限制，如计算复杂性和长期依赖关系的捕捉问题。这就引出了 Transformer 的概念。Transformer 通过引入自注意力机制，不再依赖递归的方式进行序列处理。它可以同时考

虑输入序列中所有位置之间的相关性，从而克服了 RNN 在处理长期依赖关系时的一些问题。Transformer 不仅可以自由地访问上下文信息，而且还具有更好的并行计算性能，能够更快地处理大规模的序列数据。这种全局特征的考虑使得 Transformer 能够更全面地理解序列的语义和关系。Yu 等人 [19] 在 2021 年提出了一种基于认知计算的语音增强模型，称为 SETtransformer，它可以在未知的噪声环境中提高语音质量。所提出的 SETtransformer 利用了 LSTM 和多头注意力机制，这两种机制都受到了人类听觉感知原理的启发。具体而言，SET 转换器具有表征语音频谱中涉及的局部结构的能力，并且由于其独特的并行化性能而具有更低的计算复杂度。

卷积神经网络（CNN）在图像处理领域取得了巨大的成功，而研究人员发现 CNN 也可以应用于语音降噪任务。CNN 可以有效地学习语音信号的空间特征，对于降噪任务而言具有很大的潜力。在 2019 年，一些研究人员 [20][21][22] 已将 CNN 广泛应用于语音信号的处理。CNN 应用于语音降噪主要有以下三点优势：

（1）CNN 的卷积层通过滑动窗口的方式，在局部区域内进行特征提取。这使得 CNN 能够捕捉到输入数据中的局部结构和纹理信息，从而对噪声进行局部的建模和消除。

（2）CNN 的卷积层使用共享参数的方式进行特征提取。这意味着网络可以通过学习共享参数来识别不同位置的相似特征。在降噪任务中，这种参数共享机制可以提高模型的泛化能力，从而更好地处理具有相似特征的噪声。

（3）通过深层网络的堆叠，CNN 可以逐渐学习高级抽象特征，从而更好地区分噪声和信号。这使得 CNN 在处理复杂的噪声分布时具有优势，能够逐渐从噪声数据中提取出真实信号。

当然，CNN 应用于语音降噪也有其明显的劣势，比如：

（1）CNN 在输入数据的局部区域内进行特征提取，因此对于输入的平移或位置变化非常敏感。这意味着对于输入数据的微小位置变化，CNN 的输出可能会产生不同的结果。在一些对位置敏感的降噪任务中，这可能会导致性能下降。

（2）相比于其他模型，如 RNN 或 Transformer，CNN 在处理序列数据时通常需要更多的预处理和后处理步骤。例如，在语音信号降噪中，使用 CNN 需要将语音数据转换成音频频谱图，而后续的逆变换（如 Mel 频谱逆转换）也需要被应用到

CNN 的输出上。这些预处理和后处理步骤的复杂性可能增加部署和维护的难度。

（3）CNN 主要关注局部感受野内的特征提取，因此在某些任务中可能无法很好地建模长期的依赖关系。尤其是在处理时间序列数据等具有长距离相关性的任务时，CNN 可能无法捕捉序列中远距离的依赖结构。

综上所述，CNN 适用于处理局部特征提取和对噪声进行局部建模的降噪任务。然而，它在处理位置敏感性、训练复杂性以及长期依赖性建模方面存在一些劣势。因此，研究人员也在不断探索结合多种神经网络模型的方法来进一步提高语音降噪的性能。例如，Andong Li 等人 [23] 在 2019 年提出了将 CNN 和 RNN 结合的 CRN 应用于单耳道语音降噪领域，并取得了不错的效果。CRN 相较于单独使用 CNN 和 RNN 主要有以下优势：

（1）CRN 结合了 CNN 和 RNN 的优势，能够同时捕捉输入序列的局部空间特征和时序信息。在语音降噪任务中，局部特征能够帮助识别噪声和语音信号之间的差异，而时序信息有助于建模语音信号在时间上的动态变化，从而更好地还原干净的语音信号。

（2）CRN 可以通过卷积层对输入语音进行特征提取，有助于抑制噪声的影响。同时，循环层可以通过建模时序关系，对噪声进行建模并更好地恢复干净的语音信号。这种综合的方法可以提供更准确和鲁棒的语音降噪效果。

（3）卷积层在 CRN 中可以实现参数共享，减少模型的参数量。这样的设计使得 CRN 在计算效率方面有一定优势，可以更快地处理语音数据。总之，CRN 通过同时利用 CNN 和 RNN 的特性，能够更准确地抑制噪声并增强语音信号。其能够同时捕捉局部特征和时序信息，处理长时依赖关系，并具有较高的计算效率，从而在实际的语音降噪应用中取得更好的效果。

1.3 主要问题

当前的语音降噪技术经历了多年的发展，已经能够有效地从包含噪声的语音中消除已知的噪声，并在去除未知噪声方面也取得了一定的成果。然而，目前的语音降噪算法仍然存在一些需要改进的空间。比如在应对复杂噪声环境、增强抗干扰能力和语音质量、提高非线性和时变噪声的鲁棒性等方面，语音降噪算法还有

待进一步的研究，具体表现如下：

1.3.1 应对复杂噪声环境

解决复杂环境下的语音降噪问题是一项不断发展的研究领域。经典噪声抑制技术如谱减法、维纳滤波、最小均方误差估计等，这些技术通过在频域中处理信号来减少噪声 [4][5][6]。它们的优点是计算效率高，适用于实时处理。但是这些方法在非平稳噪声环境，特别是在混响等复杂环境中性能有限，可能会导致语音失真。而基于统计模型的方法如隐马尔可夫模型等，旨在模拟语音生成过程和噪声环境。隐马尔可夫模型假设系统可以用一组隐含的状态来表示，且状态的变换过程遵循马尔科夫性质（即下一个状态的概率只依赖于当前状态）。在语音识别和语音处理中，隐马尔可夫模型通常用来建模语音信号的时间序列特性，每一个状态可能代表着某个特定的音素或音节 [9]。隐马尔可夫模型能够自然地对语音的时序特性进行建模，因为它可以表示不同语音单元的序列以及这些单元之间转换的概率，这对于处理连续语音信号非常有效。并且，通过适当的训练，隐马尔可夫模型能学习到噪声环境下语音特征与干净语音特征之间的关系，能在一定程度上抵抗噪声干扰。隐马尔可夫模型还引入了统计方法，能够处理由于噪声带来的不确定性和变异性，它通过求解最优状态序列为每个观测序列（实际接收到的信号）分配一个概率，因此，隐马尔可夫模型具有统计稳健性。

尽管隐马尔可夫模型能某种程度上处理噪声，它主要适合处理比较平稳的噪声。对于非平稳噪声（即噪声特性随时间变化，如突发噪声，多个说话者背景等），隐马尔可夫模型的表现通常不会很好。虽然它能够处理线性的信号和噪声模型，但对于复杂的声学环境，如房间混响等非线性问题，隐马尔可夫模型就没有深度学习模型那么有效了。而且，为了让它能够有效地在噪声环境下工作，通常需要大量在特定噪声条件下录制的训练数据来训练模型 [24][25]。

在过去几年中，深度学习在语音降噪领域取得了显著的进展。诸如卷积神经网络、循环神经网络，特别是长短期记忆网络和门控循环单元等，都被广泛用于从噪声信号中恢复干净的语音。深度学习模型能够学习到更复杂的信号特征，处理非平稳噪声效果好，鲁棒性强。但是这些模型也需要大量标注数据进行训练，计算

复杂度通常较高，可能不适合低延迟的实时应用 [16][17][18]。同时，模型可能过拟合到训练数据的特定噪声类型上。

除了新的语音降噪模型的提出，也诞生了一些新的研究思路，比如引用多通道语音来帮助解决复杂噪声环境的问题，多通道语音降噪是在复杂噪声环境中提升语音质量的非常有效的方法。多通道系统通过使用多个麦克风采集语音信号，可以利用空间信息来区分目标语音和干扰噪声。Jingxian Tu 等人 [26] 在 2018 年将多通道语音与卡尔曼滤波技术想结合。相较于以往的单通道语音与卡尔曼滤波技术结合，该算法具有较高的降噪效果、较小的信号失真和较高的语音可懂度，能更好的应对复杂噪声环境。多通道系统可以实现空间滤波，比如波束形成 [27] 技术，它可以增强来自特定方向的信号，同时抑制来自其他方向的噪声和干扰。这对于鸡尾酒会效应（即在多说话者环境中分离目标说话者的语音）是特别有用的。此外，通过使用多通道系统，可以更准确地估计噪声和干扰源的位置，从而有助于更好地设计滤波器来抑制干扰。在一个通道受到干扰或无法使用时，多通道系统可以通过其他通道继续进行语音降噪，从而提供更鲁棒的性能 [28]。但是，多通道处理通常涉及复杂的算法，比如多麦克风阵列处理，这可能导致计算资源的要求提高。多通道系统中麦克风之间的同步必须精确，任何时间偏移都可能会影响降噪性能。而且，算法必须考虑到不同的噪声类型和声学环境，而这样的算法设计通常较为复杂。麦克风的位置和数量对系统的性能也有很大的影响 [29][30][31]，而在实际环境中这些因素可能是受限的。

1.3.2 非线性和时变噪声的鲁棒性

非线性和时变噪声的鲁棒性仍然是语音降噪领域的重点关注问题，因为这些类型的噪声在现实世界中是非常常见的，而它们对语音通信和自动语音识别系统的性能产生了巨大的负面影响。以下是几个具体的原因解释为什么这两类噪声是具有挑战性的，并成为研究的焦点：

（1）非线性噪声指的是那些不符合线性规律的噪声。比如，某些电子设备的故障声、汽车的引擎声、以及某些生物声音等。非线性噪声往往不容易通过简单的线性滤波来消除，因为它们的统计特性可能随时间变化，所以传统的线性降噪方法

在处理上述声音时效果有限。

（2）时变噪声是随时间而变化的噪声，比如人群噪声、交通声等。这类噪声的特征在较短的时间内就可能变化很快，使得降噪算法很难准确估计噪声，因此难以有效从语音信号中移除这种噪声。

在实时通信以及语音识别系统中，以下挑战使得对于非线性和时变噪声的鲁棒性尤为重要：

（1）适应性：降噪算法需要实时地调整参数来适应噪声的变化，以保持降噪效果。

（2）干扰抑制：在多通道降噪系统中，时变多源噪声可能来自不同方向，需要先进的空间滤波技术来抑制干扰。

（3）声学模型的泛化：在机器学习和深度学习的情境下，训练得到的声学模型需要在多种不同和未见过的噪声条件下都保持有效。

（4）计算效率：对于移动和嵌入式设备，算法需要保持低延迟和低能耗的同时，还能实时地处理这些复杂噪声。

为了处理非线性和时变噪声，研究者们开发了多种方法，包括基于统计模型的噪声估计和降噪技术、基于机器学习的自适应滤波器，以及使用深度学习的端到端噪音消除解决方案。通过这些技术，可以在很大程度上提升系统在非线性和时变噪声环境下的鲁棒性。然而，由于在现实世界中噪声环境的多样性和复杂性，这些问题仍然是语音降噪领域的关键挑战。

1.3.3 实时性和低功耗

实时性和低功耗在语音降噪领域至关重要，尤其当神经网络被应用于语音降噪时，以下几个因素进一步强化了对这两个属性的关注：

（1）用户体验：在进行实时通信（如电话通话、会议系统）和交互（如语音控制的智能助手）时，用户期待拥有无可察觉延迟的即时响应。延迟或等待的增加可能导致通信中断和用户体验下降。

（2）嵌入式设备和移动性：许多语音降噪系统被设计用于嵌入式设备，如智能手机、听筒、助听器和物联网设备等，这些设备的处理能力和电池容量有限。因此，

需要降噪算法在不过度消耗电池的情况下运行。

（3）算法复杂性：神经网络尤其是深度学习模型，通常计算量大且参数众多，这意味着它们往往需要高计算资源和功率来实时处理信号。设计低功耗且高效的神经网络模型至关重要，尤其是对于希望实现长期运行的便携设备。

（4）边缘计算：为了减少传输数据到云服务器并等待回传结果的时间，很多处理流程都转移到设备上（即边缘计算）。在设备上进行实时声音处理要求低延迟和低功耗算法，保证设备能够快速响应并长时间运行。

（5）可扩展性与部署：低功耗和实时性能的算法在没有专用硬件加速的设备上更容易实现部署。这些算法可以更广泛地应用到低端硬件平台上，使更多的用户受益。

（6）绿色计算：在全球范围内减少能源消耗和提高能效已经成为趋势，这要求所有技术和服务，包括语音降噪，都要采用低功耗解决方案。

针对实时性和低功耗的需求，研究人员和工程师们正在探索和开发各种方法，比如模型剪枝、量化、低秩分解和知识蒸馏，以简化深度神经网络模型，以及利用专门的硬件加速器来提高模型的效率。这些手段旨在平衡计算负载与能耗，实现快速有效地降噪，同时维持设备的电池寿命和处理能力。

1.4 论文主要研究内容与组织架构

1.4.1 论文主要研究内容

本项研究提出了一种融合了 U-Net 架构和小波变换的创新语音降噪技术。研究的核心内容和创新之处概括如下：

首先，我们针对训练稳定性的关键问题，特别在 U-Net 的编码器阶段探索了稳定器的使用。这一设计采用了一种新的输入标准化策略，目的是为了加快和稳定梯度的传播，从而大幅提升训练过程的效率以及模型整体的性能。

其次，为了进一步优化降噪质量，本研究在 U-Net 架构中加入了多尺度特征融合机制。这一机制通过建立起编码器与解码器之间的多个跨层连接，可以更加有效地融合不同尺度的特征信息，从而更精准地恢复信号中的高频局部细节，增强模型对噪声的鉴别能力以及提升降噪性能。

本研究的重要贡献在于引入和融合了两大创新点：一是稳定器的应用，提升了模型的稳定性和训练效益；二是多尺度特征融合策略，显著提升了降噪质量。通过严谨的实验验证，本研究证实了这一新方法的有效性和可行性，为语音降噪技术的发展贡献了新思路和参考价值。

1.4.2 论文组织架构

1. 论文的第一章为绪论，本章节作为引言，旨在为读者提供研究的背景和论文的基本框架。本章主要内容包括：

（1）研究背景：首先介绍语音降噪技术的重要性以及深度学习在此领域的应用背景，说明研究的必要性和意义。

（2）研究现状：概述当前语音降噪技术的发展现状，特别是 U-Net 网络在处理相关问题时的应用情况，以及存在的主要问题和挑战。

（3）主要研究问题：明确本研究旨在解决的关键问题，并阐述预期的研究目标与贡献。

2. 论文的第二章为语音降噪技术发展的调研。本章节为文献综述，调研历史和当前的语音降噪技术，具体内容包括：

（1）传统语音降噪方法的发展历程，包括它们的优点和局限性。

（2）现代深度学习方法在语音降噪中的应用，突出显示 U-Net 等网络结构的优势。

（3）现有研究的不足之处和潜在改进点，为后续章节提出的创新点奠定理论基础。

3. 论文的第三章为基于多级嵌套 U-Net 的神经网络模型。本章节系统地介绍改进 U-Net 网络的设计理念、结构和实施方法。重点内容包括：

（1）改进 U-Net 网络的动机和设计原则。

（2）网络的详细架构描述，包括编码器和解码器阶段的修改。

（3）基于 MSE 重构误差的权重更新策略。

4. 论文的第四章为引入多尺度融合的 U-Net 神经网络模型。继续在第三章的基础上，深入研究通过多级小波变换对 U-Net 进行优化。主要内容如下：

（1）多级小波变换在特征提取中的作用和重要性。

（2）小波变换特征与改进 U-Net 模型的融合策略。

（3）融合小波变换特征的 U-Net 网络性能和有效性的验证。

5. 论文的第五章为研究总结与未来展望。本章总结了论文在语音降噪任务中实现的主要技术突破和贡献，以及讨论了模型的局限性和可能的改进意见，并对未来可能的研究方向进行了展望。

第二章　语音降噪的发展及调研

2.1　语音降噪的基础理论

2.1.1　语音信号的声学特征

语音信号的声学特征是指那些可以用来描述语音声波属性的参数，这些特征通常被用于语音识别、语音合成、说话人识别、语音情感分析等多种语音处理任务。以下是一些基本和常见的声学特征：

（1）基频：基频是语音信号的震动频率，决定了声音的音高。男性语音的基频通常较低，而女性和儿童的语音基频较高。

（2）共振峰：共振峰是指在发音管（从声带至唇部的空间）中特定频段的声音增强。它们主要由口腔和咽喉的形状及大小决定，并与元音的辨识息息相关。第一个共振峰 F1 和第二个共振峰 F2 最为重要，分别对应了元音的高度和前后位置。

（3）时域特征：时域特征包括语音信号的振幅、能量、零交叉率等。这些特征在进行语音活动检测和说话人识别时尤其重要。

（4）频率域特征：在频率域内，语音信号可以被表示成不同频率成分的组合。频率域特征包括短时傅里叶变换（STFT）的振幅和相位谱，以及谱质心、频谱熵等。

（5）倒谱参数：其中梅尔频率倒谱系数（MFCCs）是最常用的特征之一，它是一个在语音和音频处理领域广泛使用的特征，尤其是在获取与人类听觉感知相符的特征方面 [32]。它基于对语音信号在频率域的响应进行非线性梅尔尺度处理，然后计算对数谱和进行倒谱分析。

（6）频谱子带：语音信号可以分解成多个频段（子带），每个子带可以表征一定范围内的频率信息。特征如子带能量、子带倒谱等在某些语音识别任务中有其独特的应用。

（7）声学事件：语音中的爆破音、摩擦音、鼻音等都是有特定声学属性的事件，这些事件可以辅助语音识别和理解。

（8）语音质量：语音质量与个体的发音方式有关，如气声、沙哑声等。不同的发音质量会影响声波的频谱特征。

这些特性可用于区分不同的语音单位（如音素、字、词），并可以用于从语音信号中提取有用信息。在实际应用中，可根据任务的需求选择合适的声学特征进行分析和处理，如图2.1所示是常见的声学特征提取示意图：

图 2.1　声学特征提取示意图

2.1.2　语音信号的预处理流程

语音信号预处理是在进行任何高级语音分析之前，对原始语音数据进行的一系列处理步骤。目的是提升语音信号的质量，减少噪声和不必要的变量，使得后续的语音识别、说话人识别、情感分析等任务可以更准确地进行。以下是一个典型的语音信号预处理流程：

（1）采样与量化：将连续的语音信号通过麦克风等设备转换成离散的数字信号。这包括选择一个适当的采样率（例如，电话质量为 8 kHz，而 CD 音质为 44.1 kHz）和量化精度（如 16 位或 24 位）。

（2）预增强：通过一个高通滤波器对语音信号进行预增强来补偿发声系统中高频部分的能量下降，并改善高频成分的信噪比。这通常通过对语音信号进行差分处理来实现。

（3）框分与窗函数：由于语音是非平稳信号，需要对长的语音信号进行短时间分析。语音信号通常被切分成一系列短时间帧，每帧之间有一定的重叠。然后对每一帧应用窗函数以减少帧边界的不连续性，常用的窗函数包括汉明窗和汉宁窗。

（4）去噪声：语音信号常常包含各种背景噪声。通过使用噪声门限、谱减法、Wiener 滤波器、小波降噪等方法可以去除或减少噪声的影响。

（5）端点检测：检测语音信号中有效语音的起止点（即忽略前后的静默部分）。有效地提取语音端点有助于降低后续处理步骤的复杂度和提升分析精度。

（6）归一化：语音信号的振幅可能因为说话者距离麦克风的不同距离或声音的不同强度而变化。通过归一化，可以调整信号的振幅，使其落在某一确定的范围内，比如将信号幅值归一化到 [-1,1] 之间。

以上步骤完成后，语音信号预处理完成，可以进一步进行特征提取和分析。预处理的每一个步骤都对最终的语音识别或其他任务的性能有着直接的影响，因此需要根据具体的应用场景和要求仔细调整预处理流程中的各个参数。

2.1.3 语音信号的特征提取流程

语音信号的特征提取旨在提取出能够代表语音信号重要信息的数学表示。这些特征对于后续的语音识别、说话人验证或其他语音处理任务至关重要。以下是典型的语音信号特征提取步骤，其中最常用的特征之一是梅尔频率倒谱系数（MFCCs）：

（1）预加重

预加重步骤在预处理阶段执行，但对于特征提取也很重要。这里使用以下公式增强高频部分：

$$S'[n] = S[n] - \alpha s[n-1] \tag{2.1}$$

其中 $S[n]$ 是第 n 个采样点的信号，通常 α 被设定在 0.95 和 0.97 之间。

（2）分帧

将语音信号切割成短时间框架，每个框架长度一般为 20-40 毫秒。这样做是因为语音信号的特性在这么短的时间内被认为是平稳的。

（3）窗函数对每个框架应用窗函数（如汉明窗）以减少帧的边界效应。对于第

i 帧的第 j 个采样点（$Si[j]$）：

$$Wi[j] = Si[j] * hamming[j] \tag{2.2}$$

$$\text{hamming[j]} = 0.54 - 0.46\cos\left(\frac{2\pi j}{N-1}\right) \tag{2.3}$$

其中，N 是每帧的采样数，j 是窗口的索引，$hamming[j]$ 是对应索引处的窗口系数，汉明窗在主瓣区域边缘具有较好的平滑性，有利于减小频谱泄漏，同时具有较快的下降速度，且它在信号两端的值接近于零，这有助于减少频域分析时引入的频谱泄漏，汉明窗的峰值为 0.54，这可以保证信号的幅度不会被显著地抑制，并且汉明窗是实对称的，具有零相位响应。

（4）傅里叶变换

对每帧应用快速傅里叶变换（FFT）得到每一帧的频谱。对 j 采样点的 i 帧计算傅里叶变换：

$$X_i(k) = \sum_{j=0}^{N-1} w_i\left[j\right] e^{-\frac{2\pi kj}{N}} \tag{2.4}$$

其中，k 是频率索引，$w_i\left[j\right]$ 是输入信号序列，$X_i(k)$ 是变换后的频谱，N 是信号的长度。这个公式描述了从时域到频域的变换。它将信号分解成 N 个频率成分，每个频率成分的幅度和相位由复数表示。对于离散信号，这个公式将 N 个时域采样点映射到 N 个频域离散频率点。

快速傅里叶变换的思想是基于分治法 [33]，将信号分解成多个规模较小的子问题，并利用对称性质和递归思想来减少计算量。它通过不断地将信号划分为偶数点和奇数点进行迭代计算，然后通过组合这些计算结果来获得最终的频域结果。

（5）功率谱计算

根据傅里叶变换得到的频谱计算功率谱：

$$P_i(k) = \frac{1}{N}|X_i(k)|^2 \tag{2.5}$$

其中，$P_i(k)$ 表示信号在频率为 k 处的功率谱密度，N 是信号的持续时间，$X_i(k)$ 是傅里叶变换后得到的频谱值。这个公式描述了如何从信号的频谱信息中计算出信号在不同频率处的功率谱密度。功率谱密度表示了信号在不同频率上的能量分布情况，通过计算 $|X_i(k)|^2$ 可以得到信号在频率为 k 处的功率。

（6）梅尔滤波器组

在梅尔刻度上应用滤波器组，该刻度是根据人类听觉特性设计的。将梅尔刻度与传统频率刻度的关系可以用以下公式表示：

$$m = 2595\ \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.6}$$

其中 f 是频率，m 是梅尔刻度值。经由梅尔滤波器组，我们得到一系列滤波器结果。

（7）离散余弦变换

对梅尔滤波器组的对数输出应用离散余弦变换得到梅尔频率倒谱系数：

$$c_n = \sum_{m=1}^{M} l_i(m) \cos\left[n(m-0.5)\frac{\pi}{M}\right] \tag{2.7}$$

DCT 公式描述了通过将时域信号与一组余弦函数进行加权求和，将信号从时域转换到频域的过程。DCT 变换是实数到实数的变换，一般来说，取 DCT 结果的前 12 到 13 个系数用作特征，这些低序数的 MFCC 被认为包含了语音信号的主要能量分布。

（8）小波变换

小波变换是一种具有多尺度分辨率的信号分析方法，它允许信号在不同的尺度上被分析 [34]，可以同时提供时间和频率信息，使其在处理非平稳信号方面非常有优势。

连续小波变换的公式定义如下：

$$CWT(a,b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} x(t)\psi\left(\frac{t-b}{a}\right) dt \tag{2.8}$$

$CWT(a,b)$ 表示连续小波变换的结果，其中 a 和 b 是尺度和平移参数，用于调整小波函数的尺度和位置。$x(t)$ 代表原始信号，其中 t 是连续时间变量。在小波变换中，我们要分析或处理的信号通常是一个时间序列。$\frac{1}{\sqrt{|a|}}$ 是归一化因子，用于确保在不同尺度 a 下的小波函数满足能量归一化条件。连续小波变换提供了一种在尺度和时间上同时分析信号特性的方法，通过变化参数 a 和 b，可以获得信号在不同频率和时间尺度上的表示。通过计算连续小波变换，可以得到信号的时频图，帮助我们理解信号的局部特征和频率内容。

离散小波变换的公式定义如下：

$$DWT(j,k)=\frac{1}{\sqrt{2^j}}\int_{-\infty}^{\infty}x(t)\psi\left(\frac{t-k2^j}{2^i}\right)dt \tag{2.9}$$

$DWT(j,k)$ 表示在尺度 j 和位置 k 下的离散小波变换结果，$x(t)$ 是输入信号，通常为离散时间序列，如音频信号。DWT 的计算过程可分为低通滤波与下采样和高通滤波与下采样，重复这两个步骤，逐级进行分解，直到达到所需要的分解层次。

2.2 传统的语音降噪方法调研

2.2.1 统计滤波算法

统计滤波算法是用于语音降噪的一种非常重要的算法类别，它们基于统计模型来区分噪声和语音信号，并尝试减少噪声的影响，以提高语音的质量和可懂度。这些算法通常假设噪声和语音是统计上独立的，且具有不同的统计特性。以下是一些统计滤波算法的基本概念和公式。

（1）噪声功率谱估计

在滤波算法中，首先必须估算噪声功率谱。通过在没有语音活动时（即只有噪声时）对信号进行分析估算出噪声谱。常用的估算方法是计算信号短时傅里叶变换 (STFT) 的功率谱，并在认为噪声存在的帧上取平均。如果 $Y(k,l)$ 是第 1 帧的第 k 个频率分量的 STFT，噪声功率谱的估计 $\widehat{N}(k,l)$ 可能如下：

$$\widehat{N}(k,l)=\frac{1}{L_N}\sum_{l=l_s}^{l_s+L_N-1}|Y(k,l)|^2 \tag{2.10}$$

其中 L_N 是用于平均的帧数，l_s 是开始的帧索引。

（2）声音信号的功率谱估计

在统计滤波算法中，还需对正在处理的帧的语音信号的功率谱进行估计。这可以通过取当前信号帧的 STFT 的功率谱减去先前估计的噪声功率谱来完成：

$$\hat{S}(k,l)=max\left(|Y(k,l)|^2-\widehat{N}(k,l),\epsilon\right) \tag{2.11}$$

其中 ϵ 是一个很小的正数，用来确保估计的语音功率谱非负。

（3）维纳滤波器

维纳滤波器是统计滤波算法中非常著名的一种，它基于最小化信号的均方误差来优化 [35]。维纳滤波器的增益函数 $G(k,l)$ 通常定义为：

$$G(k,l)=\frac{\hat{S}(k,l)}{\hat{S}(k,l)+\hat{N}(k,l)} \tag{2.12}$$

维纳滤波器基于这样的假设，即信号和噪声在频域是累积的，并且它们统计上是独立的。

（4）频谱减法

频谱减法是另一种统计滤波算法，它直接通过减去估计的噪声功率谱从信号中减去噪声：

$$\hat{Y}_{clean}(k,l)=max\left(|Y(k,l)|-\alpha\sqrt{\hat{N}}(k,l),\delta\right)e^{j\theta_{Y(k,l)}} \tag{2.13}$$

$\hat{Y}_{clean}(k,l)$ 是去噪后信号的复数 STFT 值，α 是一个小于或等于 1 的过减因子，$\theta_{Y(k,l)}$ 是 $Y(k,l)$ 的相位，而 δ 提供一个保护楼层以避免导致负估计。

2.2.2 自适应滤波算法

自适应滤波算法是一种能够自我调整其滤波器系数以最优方式根据输入信号进行工作的算法。其核心原理是通过算法不断调整滤波器的参数 [36]（通常是权重或者系数），以最小化输出信号和某个期望信号之间的差异（即误差）。在语音降噪的应用中，目标通常是减少噪声成分，同时保持原始语音信号尽可能不变。

自适应滤波通常包括以下几个关键步骤:

（1）初始化: 设置滤波器的初始参数值，包括初始权重和步长等重要参数。

（2）误差计算: 在每一时刻，计算滤波器的输出和期望信号之间的差，也就是误差。

（3）权重更新: 根据已经计算出的误差来更新滤波器的权重。权重更新的目标是使得输出信号尽可能接近期望的信号。

（4）迭代优化: 不断重复上述过程，通过逐次迭代来优化滤波器的性能。

2.2.3 光谱减算法

光谱减法算法的语音降噪原理简要来说是通过估计语音信号中的噪声成分并在频域中将其从受噪声污染的语音信号中减去，以便获取更清晰的语音信号。

以下是光谱减法的简化流程：

（1）转换到频域：使用短时傅里叶变换（STFT）将时间序列的语音信号转换为频域表示。

（2）噪声频谱估计：在语音信号中找到非语音段（即只存在背景噪声的时间），计算其平均频谱作为噪声频谱的估计值。

（3）进行光谱减法：从每段含噪语音信号的频谱中减去估计的噪声频谱，得到去噪后的语音频谱。为避免产生负值，通常会对减去的结果设置一个最小的阈值。

（4）逆变换到时域：将去噪后的语音频谱通过逆短时傅里叶变换回到时域，得到降噪后的语音信号。

光谱减法简单有效，但可能会导致失真和“音乐噪声”[37]（由于噪声估计不准确引起的一些不自然的声音）。为此，存在各种改进算法来减少这些负面影响。

2.3 基于神经网络的语音降噪方法调研

2.3.1 卷积神经网络

卷积神经网络（Convolutional Neural Network，CNN）是一种深度学习模型，常用于处理图像和音频等数据。CNN 主要由卷积层、池化层、全连接层和激活函数等组成。在应用于语音降噪中，CNN 可以用于学习音频中的噪声模式和语音特征，帮助降低噪声并提取语音信息 [38]，CNN 的整体架构如图2.2所示：

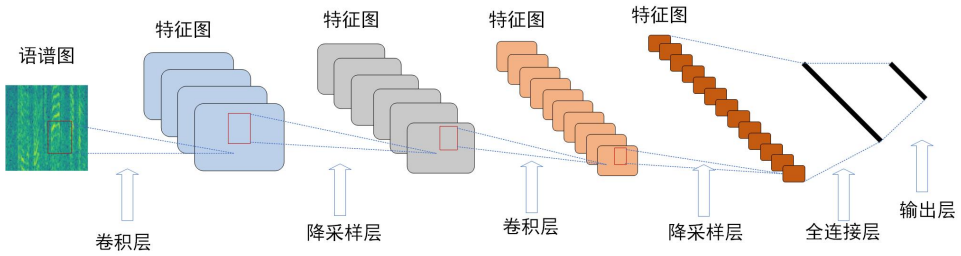


图 2.2 卷积神经网络整体架构图

下面是 CNN 的各层介绍及其在语音降噪中的应用：

（1）卷积层：

卷积层通过一系列滤波器（卷积核）对输入数据进行滤波处理，提取数据中的特征。在语音降噪中，卷积层可以帮助网络学习音频中的噪声特征和语音信号特征，以便更好地降噪和提取干净的语音信号。

（2）池化层：

池化层用于减少数据维度，保留主要特征，同时减少计算量，提高模型的鲁棒性。在语音降噪中，池化层可以帮助网络对特征进行降维处理，提高系统的运行效率。

（3）全连接层：

全连接层将前一层的所有神经元与当前层的所有神经元相连接，用于整合前一层提取的特征。在语音降噪中，全连接层可以帮助网络进一步学习特征和调整权重，从而更好地降噪和提取干净的语音信号。

（4）激活函数：

激活函数引入非线性因素，提高网络的表达能力，使其能够学习更加复杂的模式。常用的激活函数包括 ReLU、Sigmoid、Tanh 等。

在语音降噪研究中，研究者们通常会将 CNN 与其他技术（如循环神经网络 RNN[39]、长短时记忆网络 LSTM 等）结合使用，构建更复杂的模型，以提高降噪效果和语音还原的质量。同时，为了提升模型的性能，也会不断尝试新的网络结构和训练方法。

2.3.2 循环神经网络

循环神经网络（Recurrent Neural Network，RNN）是一种专门用于处理序列数据的神经网络。与传统的前馈神经网络不同，RNN 具有循环连接，可以处理变长的输入序列并保留时间信息。RNN 的重要特性是其隐藏层的状态可以在不同时间步之间传递，从而捕捉到序列数据的动态特征 [40][41]。在语音降噪任务中，RNN 可以学习语音信号的长期依赖关系和时序特征，用于降低噪声并提取出清晰的语音信息，RNN 的网络结构 [42] 如图2.3所示：

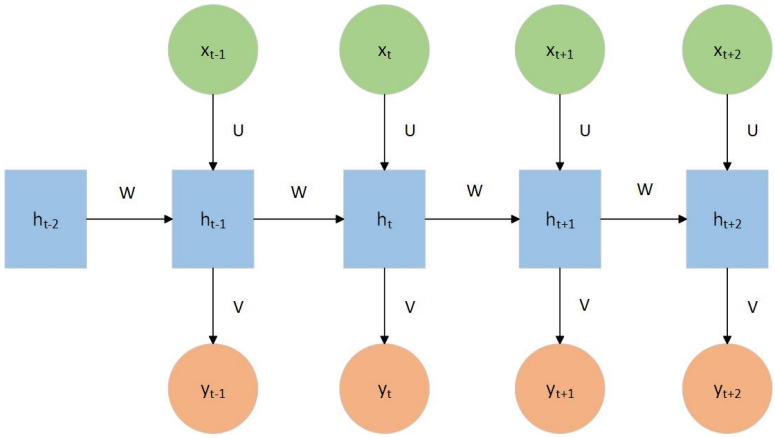


图 2.3　循环神经网络结构图

下面是 RNN 的各层介绍及其在语音降噪中的应用：

（1）输入层：

RNN 的输入层接收序列数据，如经过傅里叶变换得到的音频频谱特征。对于语音降噪任务而言，输入层将语音频谱作为网络的输入，以便网络学习语音信号的频谱特征。

（2）隐藏层：

RNN 的隐藏层是 RNN 网络的核心部分，它包含了循环连接，能够捕捉时间序列信息。隐藏层的状态会根据当前时间步的输入和前一时间步的状态进行更新，从而保留了序列数据的历史信息。在语音降噪中，隐藏层可以帮助网络学习语音信号的长期依赖关系和时序信息，以便更好地降噪和还原干净的语音信号。

（3）输出层：

RNN 的输出层通常是一个全连接层，负责生成输出结果。在语音降噪中，输出层可以通过一个线性变换或非线性激活函数，将隐藏层的状态映射到语音信号的估计输出，即降噪后的语音信号。

在实际的语音降噪任务中，为了提高降噪效果和处理更复杂的语音场景，人们常常使用变种的 RNN 结构，例如长短时记忆网络（LSTM）和门控循环单元（GRU）。这些变种结构在隐藏层中引入了门控机制，能够更好地捕捉和管理长期依赖关系。

RNN 在语音降噪中的应用除了单独使用，还常常与其他深度学习模型结合，如卷积神经网络（CNN）或注意力机制，以进一步提升降噪效果。同时，研究者们

也持续尝试新的网络结构和优化方法，以改进 RNN 在语音降噪任务中的性能。

2.3.3 长短期记忆网络

在长短时记忆网络（LSTM）中，核心的信息存储单元负责保持和传递长期依赖关系的信息。该存储单元类似于 LSTM 网络的记忆单元，负责在序列数据中维持长期信息，LSTM 的网络结构 [43] 如图2.4所示：

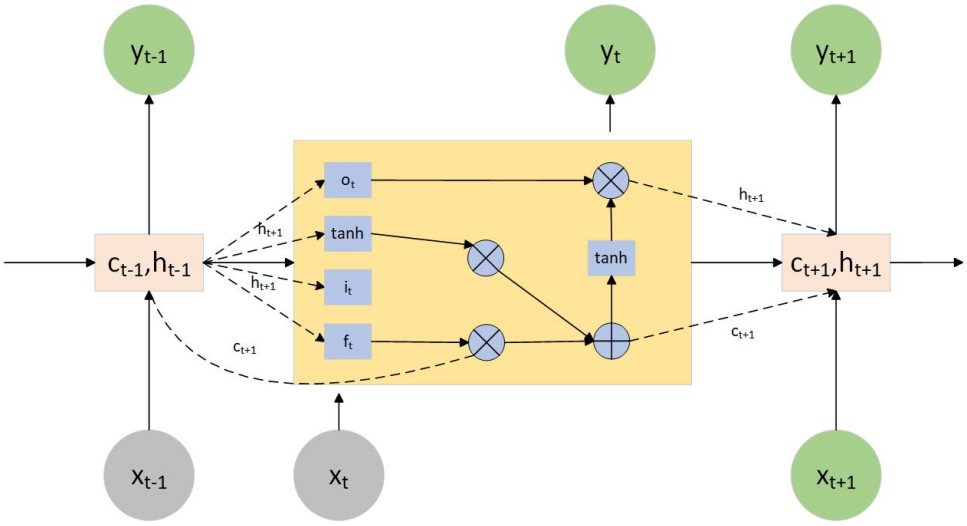


图 2.4　长短期记忆网络结构图

在 LSTM 网络中，信息存储单元受到输入门、遗忘门和输出门的控制和更新。输入门确定哪些信息会更新到存储单元中，遗忘门选择性地遗忘旧信息，输出门决定了如何将当前状态的信息传递给隐藏层和最终输出。

在语音降噪任务中，LSTM 的存储单元在处理音频频谱特征序列时发挥着关键作用：

（1）输入阶段：存储单元通过输入门接收新的频谱特征信息，根据门控机制更新当前时间步的状态。

（2）遗忘阶段：遗忘门决定哪些语音信号特征不需要保留在存储单元中，以适应不同的噪声环境。

（3）状态更新：存储单元负责记录和传递长期依赖关系的信息，帮助网络处理长序列数据中的语音信息。

（4）输出结果：通过输出门，网络可以生成输出结果，即降噪后的语音信号，基于当前时间步的存储单元状态和隐藏状态的信息。

综上所述，在 LSTM 网络中，信息存储单元是一个关键组成部分，通过输入门、遗忘门和输出门的调控 [44][45][46]，能够在语音降噪任务中捕捉和管理长期依赖关系，提高降噪效果。

2.3.4 轻量级 U 型网络

U-Net 是一种常用于图像分割任务的卷积神经网络结构，具有 U 字形的特殊设计，包括编码器和解码器部分 [47]，U-Net 的网络结构 [48] 如图2.5所示：

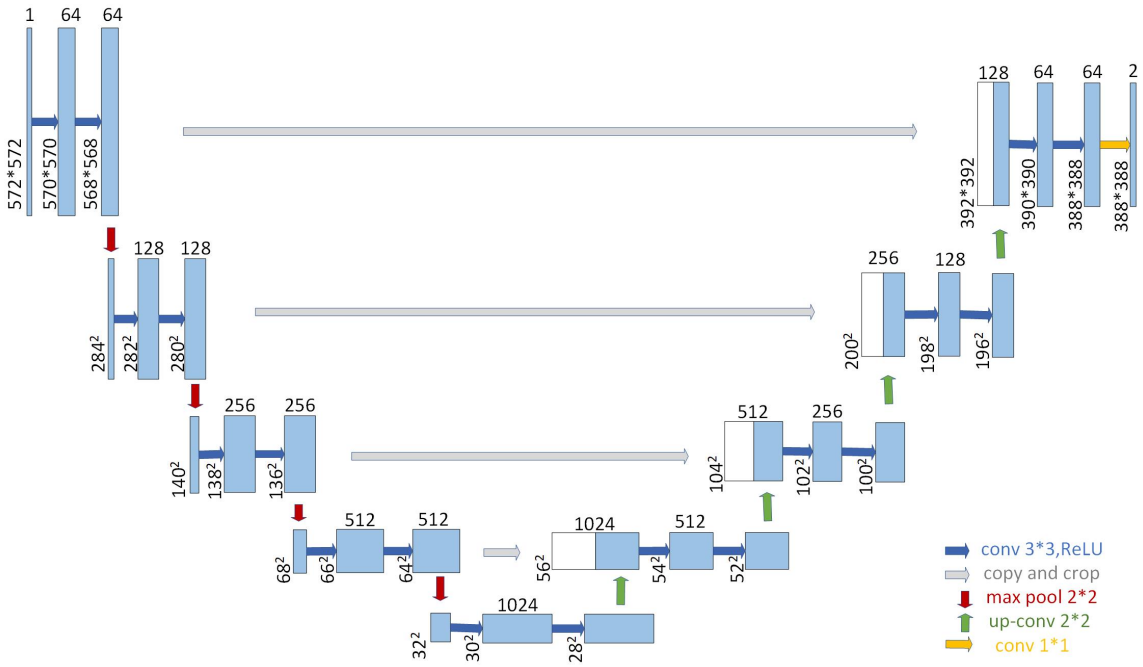


图 2.5　U-Net 网络结构图

下面将详细介绍 U-net 的各层结构，并解释它如何应用于语音降噪任务中：

（1）编码器：

编码器部分由卷积层、池化层和激活函数（通常为 ReLU）组成，用于逐步提取输入数据中的高级特征。卷积层负责捕获局部特征，而池化层则用于下采样，减小特征图的尺寸。

（2）解码器：

解码器部分由转置卷积层、跳跃连接和激活函数组成，用于逐步将编码器提取的特征图恢复到原始输入图像的大小。跳跃连接有助于融合不同层级的特征信息，以改善分割结果的精度。

（3）跳跃连接：

跳跃连接是 U-Net 的关键设计特点之一，将编码器中的特征图与解码器中对应层级的特征图连接在一起。这样做可以帮助网络在解码过程中保留更多细节信息，提高分割准确性。

（4）损失函数：

U-net 通常使用像素级别交叉熵损失函数来衡量模型输出与真实标签之间的差异，用于指导网络学习提供更准确的分割结果。

在语音降噪任务中，可以将 U-Net 应用于频谱图像的分割，实现语音信号的降噪效果 [47][49][50]。具体而言，U-Net 在语音降噪中的应用过程如下：

步骤一：输入处理：将语音信号转换为频谱图像，作为 U-Net 的输入数据。

步骤二：特征提取：U-Net 的编码器部分负责提取频谱图像中的关键特征，包括语音信号和噪声信号的特征。

步骤三：信息融合：通过解码器部分和跳跃连接，U-Net 能够融合高层次和低层次的特征信息，有助于准确地分割出语音信号部分。

步骤四：输出结果：U-Net 的输出结果是经过降噪处理的频谱图像，可以进一步进行逆变换，还原成清晰的语音信号。

通过以上方式，U-Net 在语音降噪任务中展现出了良好的特征提取和分割能力，有助于提高降噪效果并减少噪声对语音信号的干扰。这使得 U-Net 成为处理语音信号的有效工具之一。

2.4 语音降噪方法评价指标

语音降噪算法的性能通常通过客观评价指标和主观评价指标来衡量。客观评价指标是量化的性能指标，而主观评价指标通常涉及到实际的人类听众对去噪效果的主观感受。以下是一些常用的评价指标：

1. 客观评价指标：

（1）信噪比：SNR 衡量的是信号的功率与背景噪声功率的比值 [51]。通过去噪过程，期望 SNR 值提高，表示噪声水平相对于信号的水平减少。

（2）改善的信噪比：ISR 是指去噪前后信噪比的差值。它表示去噪算法提高信噪比的能力。

（3）语谱失真：语谱失真测量去噪语音与原始干净语音在频谱上的差异。较低的 SD 值表示去噪语音频谱接近于原始干净语音频谱。

（4）对数似然比改善：对数似然比改善是基于对数似然比的变化来衡量去噪效果，该指标可以反映畸变的程度。

（5）语音清晰度指数：STOI 是一个用于预测语音清晰度的指标，即去噪语音对于理解的可懂度。STOI 的值通常在 0 和 1 之间，值越大代表语音越清晰可懂。

（6）PESQ：PESQ 是 ITU-T 推荐的一种语音质量评价算法，被广泛应用于语音编码和降噪效果的评价 [52]。它提供了一个从-0.5 到 4.5 的分数，分数越高代表语音质量越好。

2. 主观评价指标：

（1）MOS：MOS 是一种通过听力测试进行的评价，听众对于语音质量进行打分，一般范围在 1 到 5 之间。MOS 的平均值是衡量语音质量的传统和可靠的方法。

（2）CMOS：当比较两种处理过的语音时，可以使用比较均值意见得分（CMOS）。听众会给出它们之间相对质量的评分，表示哪一个听起来更好。

（3）DOS：当需要评价去噪前后语音质量退化的程度时，可以使用 DOS。听众对比去噪前后的语音，对质量退化程度进行评分。

各种评价指标可能会根据特定的需求和场景被选择和使用。通常客观评价指标更易于自动计算和比较，但主观评价指标被认为更贴近于实际的人类听觉体验。最佳的做法是结合客观和主观评价指标来全面评估语音降噪算法的性能。

2.5 本章小结

本章系统地介绍了语音降噪领域的基础理论知识，包括声学特征分析、预处理工作流程以及特征提取流程。这些基础理论为后续讨论提供了必要的背景知识和理论基础。

通过调研传统的语音降噪方法，包括统计滤波算法、自适应滤波算法和光谱减算法，我们发现传统方法在一定程度上可以减少噪声对语音信号的影响，但在处理复杂的噪声环境和保留语音信号细节方面仍存在局限性。

针对传统方法的局限性，我们调研了基于神经网络的语音降噪方法，包括卷积神经网络、循环神经网络、长短期记忆网络以及本研究所优化模型对应的基础模型 U-Net。这些神经网络模型在语音降噪任务中展现出了很好的潜力，能够更好地捕捉语音信号的复杂特征并有效降低噪声干扰。

本章节介绍了基础模型 U-Net 的层级架构，后续本研究对传统 U-Net 进行了改进和优化，为轻量级神经网络应用于语音降噪提供了新的思路和方法。

第三章　基于多级嵌套的 U-Net 神经网络语音降噪方法

3.1　引言

本章节着眼于基于 MSE 重构误差的多级嵌套 U-Net 模型，通过在 U-Net 网络的编解码器部分引入深度级别的损失函数计算和即时局部权重更新，以提高特征的完整性和重构精度。

在语音降噪领域，模型的重构精度对于恢复清晰的语音信号至关重要 [53][54]。为了有效提高重构精度并确保特征的完整性，本文提出了基于 MSE 重构误差的多级嵌套 U-Net 网络模型。通过在 U-Net 网络的编解码器部分引入深度级别的损失函数计算和即时局部权重更新，本模型旨在优化特征提取和信息重建的过程，从而实现更精准的语音信号恢复。多级嵌套的设计使得模型在不同层级能够更全面地捕获语音信号的特征，并保持特征的连续性和一致性。深度级别的损失函数计算有助于指导网络在不同深度层次上学习更准确的特征表示，从而提高重构的准确性和稳定性。与传统的 U-Net 模型相比，多级嵌套 U-Net 在特征提取和重构过程中具有更强的鲁棒性和效率性。

通过缜密的实验验证和结果分析，本章研究的多级嵌套 U-Net 网络模型，在语音降噪任务中取得了更好的性能表现，为提高语音重构的精度和完整性提供新的方法和思路。本章将详细介绍模型的设计原理、实验设置以及结果分析，展示多级嵌套 U-Net 在语音降噪领域的潜在应用和价值。

3.2　基于 MSE 重构误差的损失函数

MSE 重构误差是衡量模型重构输出与预期输出或原始数据之间差异的一种方法。它是最常用的量化重构质量的指标之一。MSE 的定义是预测值与实际值差异的平方的平均值。在语音降噪的任务中，MSE 重构误差可以这样计算：

$$L_{MSE}(y,\hat{y}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3.1}$$

其中 y 是目标信号的实际值，例如原始的、未损坏的信号。$\hat{y}$ 是模型重构的信号，也就是通过神经网络、算法或任何其他方法预测的值。N 是样本点的总数，L_{MSE} 是计算得到的 MSE 重构误差。

这个函数计算了每个样本点的误差平方，并将它们求和后再取平均得到最终的 MSE 值。MSE 重构误差越小，表明模型的重构输出与真实信号越接近，即模型性能越好。在语音降噪的语境中，重构的干净信号应当尽可能准确地反映原始信号 [55]，而 MSE 重构误差则提供了量化模型性能的一种手段。

3.3 引入稳定器的多级嵌套 U-Net 神经网络模型

该模型的设计用于更有效地进行语音信号的降噪。模型独到之处在于在 U-Net 网络的编解码器部分，通过在每层卷积后引入深度级别的损失函数计算和即时局部权重更新来确保特征的完整性和提高重构精度，形成多级嵌套的 U-Net 网络模型。这种结合了传统信号处理和深度学习方法，以及引入局部损失监督和权重更新机制的创新 U-Net 网络架构，旨在提高去噪后语音信号的质量，同时也增加了模型训练的效率和效果。改进 U-Net 网络的模型结构如图3.1所示。

在本研究提出的改进 U-Net 网络模型中，通过每个阶段的上采样过程来计算一个从输入至当前层的局部损失，进而更新权重参数，从而有助于达到更好的重构效果。具体为:

（1）阶段损失计算:

改进的 U-Net 模型特点之一在于每个反卷积阶段都与深度监督的损失函数相关联，损失函数用于计算在当前阶段重构的语音信号与实际纯净信号之间的差异。在训练过程中，损失函数值的梯度将沿网络被反向传播，用以更新当前阶段以及之前所有阶段的权重。

（2）局部权重更新:

改进模型中在编码器阶段也有即时局部权重更新的机制。这意味着网络不会仅在全局损失（来自整个网络输出）的基础上更新权重，而是能够利用局部损失（来自每个上采样阶段的输出）更细致地调整权重。这有助于网络在解码器的每一个阶段都更精确地调整特征图，以接近期望的清晰信号。

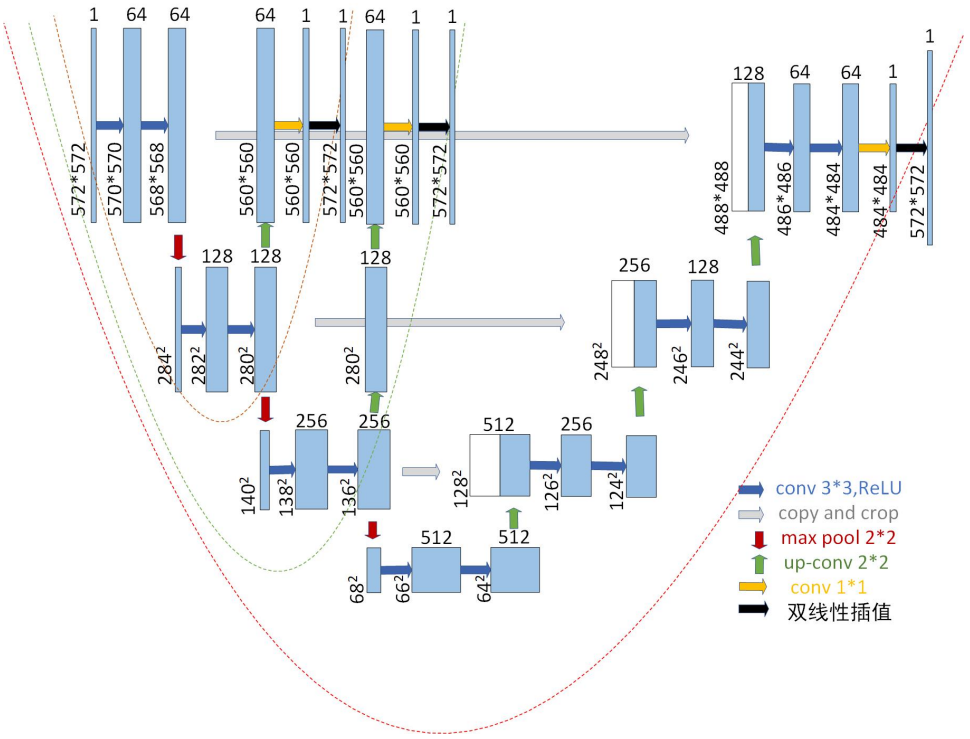


图 3.1　引入稳定器的多级嵌套 U-net 结构图

3.3.1　编码器阶段

1. 编码器阶段的结构

改进的编码器保留了 U-Net 的基本结构，即一个向下采样的过程，用于逐步提取空间特征和减少数据的空间维度。但是，在改进的网络结构中，编码器被修改为在每两层卷积之后进行一次反卷积操作，并基于 MSE 重构误差计算局部损失，来优化权重参数。以下是编码器的基本流程细节：

（1）卷积层:

卷积层负责提取特征，该层使用 ReLU 作为激活函数，来引入非线性。经过两层卷积后，特征图的尺寸和深度会根据卷积操作的设置而改变。

（2）反卷积与误差计算:

在编码阶段经过两层卷积之后，会执行一次反卷积（也称作转置卷积）操作，该操作的目的是对前两层卷积运算得到的特征图做一次上采样，用于恢复到接近原始输入数据的维度。计算编码阶段损失一的过程如图3.2所示。

同样地，再进行两次卷积之后，会计算编码阶段的第二个损失，进一步更新权

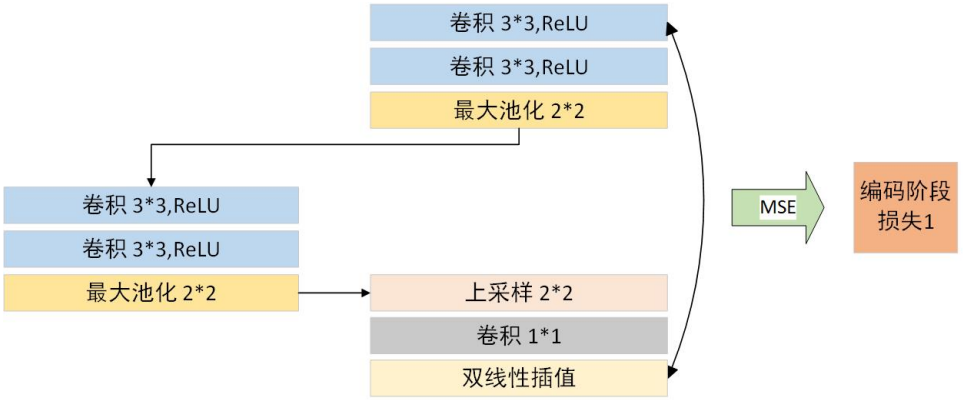


图 3.2　计算编码阶段损失 1 流程图

重参数，如图3.3所示，在反卷积后，通过计算得到的特征图与原始输入数据间的均方误差（MSE）。这个误差反映了当前编码器输出与初始信号之间的差异。

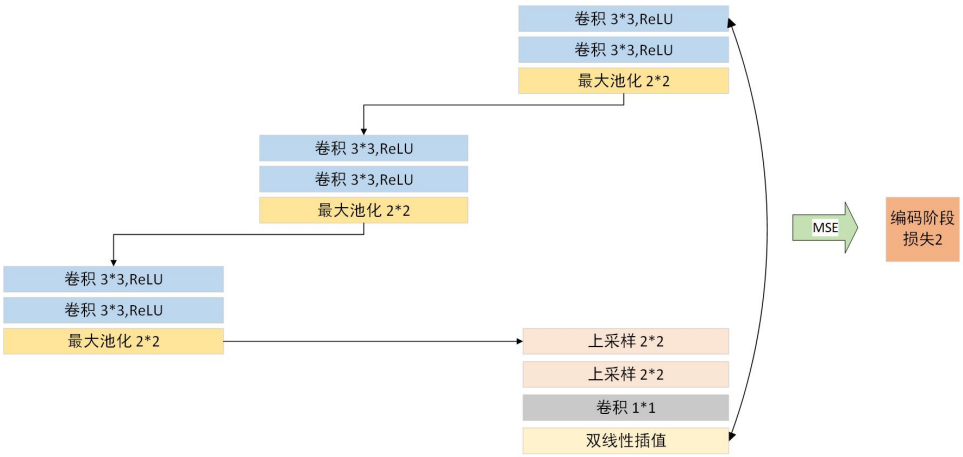


图 3.3　计算编码阶段损失 2 流程图

（3）权重更新:

计算得到均方误差（MSE）后，采取一定的梯度下降策略（本研究使用 Adam）来更新网络权重。这个更新不仅影响反卷积层本身，还影响从原始输入到当前层间所有相关的卷积层的权重。该过程类似于应用了一个局部损失函数，目的是校正每一阶段的特征提取以优化整个模型的性能。

（4）池化层:

编码器中包含最大池化层，用于在卷积操作之间降低特征图的尺寸，提高模型的抽象能力，减少计算量，并增加特征图的接受场。

（5）特征图尺寸恢复

在以 U-Net 为基础模型进行语音降噪时，需要确保输出的特征图与输入尺寸一致，避免丢失语义信息。本研究使用双线性插值法进行特征图的恢复。双线性插值是一种常用的图像插值方法，可以根据已知点的信息推算出未知点的像素值，从而实现图像的放大或缩小。下面是利用双线性插值进行语谱图大小恢复的步骤：

步骤一：计算缩小比例：将原始语谱图的大小除以目标大小，得到缩小比例。

步骤二：创建目标大小的空白画布：生成一个与目标大小相匹配的空白画布，用于存储处理后的语谱图。

步骤三：遍历目标画布中的每个像素，计算在原始语谱图中的坐标：将目标画布中每个像素的坐标乘以缩小比例，得到在原始语谱图中对应的坐标。

步骤四：找到最近邻的四个像素的坐标：根据原始语谱图中的坐标，找到其上下左右最近的四个像素的坐标。

步骤五：计算权重：根据坐标的小数部分，计算插值需要的权重值。

步骤六：双线性插值计算：利用最近邻的四个像素的值和权重，进行双线性插值计算，得到目标像素的值。

步骤七：将像素值填充到目标画布：将计算得到的目标像素值填充到目标画布中对应的像素处。

步骤八：返回恢复后的语谱图：处理完所有像素后，返回处理后的语谱图即目标画布。

2. 多级嵌套 U-Net 设计特点

此设计旨在保证编码器不仅仅关注在深层次的抽象特征，还保留了关于输入的重要信息，这对于信号重构或分割任务特别重要。理论上，这种方法允许更早地识别和更正特征提取过程中可能出现的错误，因为每个阶段都尝试尽可能好地重构出输入信号。通过计算局部损失并及时更新权重，这种结构有助于改善梯度消失问题，并在训练过程中提供更好的收敛速度。

3.3.2 解码器阶段

在本研究提出的改进 U-Net 网络模型中，解码器阶段负责将压缩后的特征信息逐步解压，恢复到原始信号的尺寸。下面是解码器阶段的详细介绍:

解码器阶段主要负责利用编码器阶段提取的特征来进行信号的上采样和重构。它由一系列卷积、上采样操作构成，目的是不断地增大特征图的空间分辨率，最终得到与原始输入数据相同尺寸的输出。以下是解码器的基本流程细节:

（1）接收编码层特征: 解码器从编码器的最深层接受压缩后的特征图。这包括在解码器每一步的起始进行跳跃连接，从而将高级特征与低级特征结合。

（2）卷积层: 类似编码器阶段，解码器的卷积层也用于进一步细化特征表示。经过激活函数后，可以保留和增强重要的特征信息。

（3）反卷积与误差计算: 解码器在每两层卷积之后执行一次反卷积操作，其目的是对特征图进行上采样，使其在空间上扩大，接近原始信号的大小。

最后，解码器在输出层会计算出一个全局损失（这里的损失计算过程类似 U-Net 原型，只是本研究所提出的改进 U-Net 模型与 U-Net 原型的网络深度不同）。这个损失反映了解码器的输出与目标输出之间的差异。全局损失的计算过程如图3.4所示。

（4）权重更新: 根据 MSE 重构误差，同样采用梯度下降策略来更新从输入到输出层的所有相关权重。这有助于在重构过程中优化特征图的空间细节，使重构的信号能更准确地匹配原信号。

3.4 实验与结果分析

3.4.1 数据生成

本研究使用了两种数据集用于生成带噪声的语音训练样本：首先，实验使用 WSJ0 作为干净无噪声的人声语音数据。WSJ0 是一个广泛使用的公开数据集，包括了多个说话人在安静环境下录制的英语语音。该数据集常用来训练和测试语音处理相关的算法和模型。其次，实验使用了自制的噪声数据集（简称 ZMJAUD），该数据集录制自实际的矿区工作环境，多为采煤机和液压支架等大型设备的工作

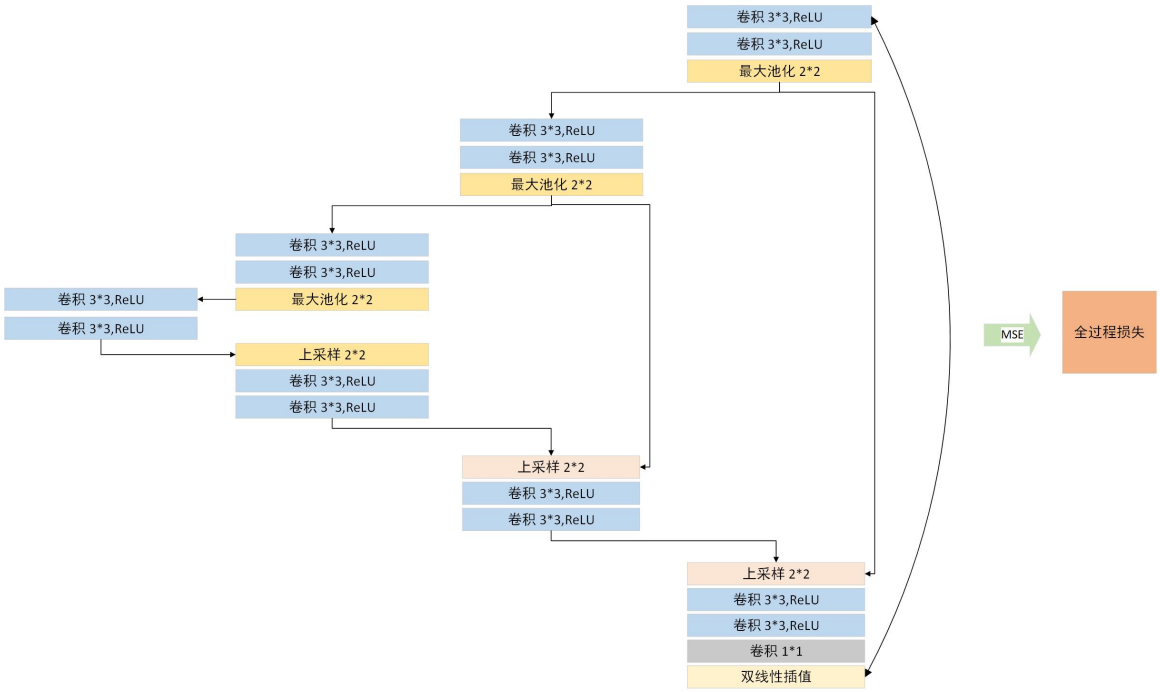


图 3.4　计算解码阶段全局损失流程图

噪声。纯净人声数据集和噪声数据集特点如表3.1所示：

表 3.1　WSJ0 数据集与 ZMJAUD 数据集特点对比

WSJ0	ZMJAUD
单声道	单声道
采样率 16kHz	采样率 16kHz
清晰的录音质量	噪声多具非线性特点
多个说话人的录音	噪声丰富，覆盖多种大型设备
记录了详细的元数据	包含复杂的声音峰值和不规则的声波形状

将 WSJ0 数据集中的纯净人声和 ZMJAUD 数据集中的噪声相结合，通过线性叠加的方式，制造出具有背景噪声的语音样本，以模拟真实环境中噪声干扰下的语音信号。在现实世界应用中，由于噪声类型和信噪比的变化可以非常大，通过调整噪声分量的能量，便可以生成不同信噪比（SNR）水平的训练样本，从而让模型更好地泛化到真实情况。

3.4.2 实验设置

数据加载包含噪声音频和对应的干净音频，通过 CWRUDataset 类，将加载的数据封装成 PyTorch 的 Dataset 对象，便于后续以批处理的方式进行模型训练。在 Dataset 类中，将数据的维度进行了调整，以适应模型的输入格式。模型建立在 U-Net 式的架构上，特别地，在编码阶段，添加了两个用于早期输出的分支，它们在编码器的不同层上接分支，使得模型可以学习不同层次的特征。使用 Adam 优化器进行参数优化，初始学习率设置为 0.1。采用 StepLR 调度器，每 30 个 epoch 学习率衰减为原来的 0.5，有助于模型在训练后期的精细调整。

3.4.3 实验结果与分析

1. 对比实验

图3.5展示了不同模型使用相同数据集 WSJ0+ZMJAUD 进行降噪的损失下降图，可以看到本文模型（The proposed model）中损失以最快的速度得到收敛，相比其他模型具有更好的性能。

此外，模型从 PESQ、STOI、CSIG、CBAK 和 COVL 这些指标上也能看到明显的优势，统计结果如表3.2所示。

表 3.2　在 WSJ0+ZMJAUD 数据集上的语音降噪得分

模型	Domain	PESQ	STOI(%)	CSIG	CBAK	COVL
SEGAN	T	2.52	-	3.76	2.98	2.83
TSTNN	T	2.97	91.87	4.41	3.53	3.76
PHASEN	T	3.13	-	4.25	3.57	3.64
CAUNet	T	3.09	92.53	4.19	3.48	3.61
DEMUCS	T	3.21	94.16	4.24	3.43	3.67
本文模型	**T**	**3.31**	**94.88**	**4.36**	**3.57**	**3.72**

2. 消融实验

为进一步验证本文提出的基于多级嵌套的 U-Net 神经网络模型对于语音降噪性能的提升，本文在 WSJ0+ZMJAUD 数据集上进行了消融实验，首先，引入稳定器的 U-Net 模型与原 U-Net 模型的损失下降对比如图3.6所示。

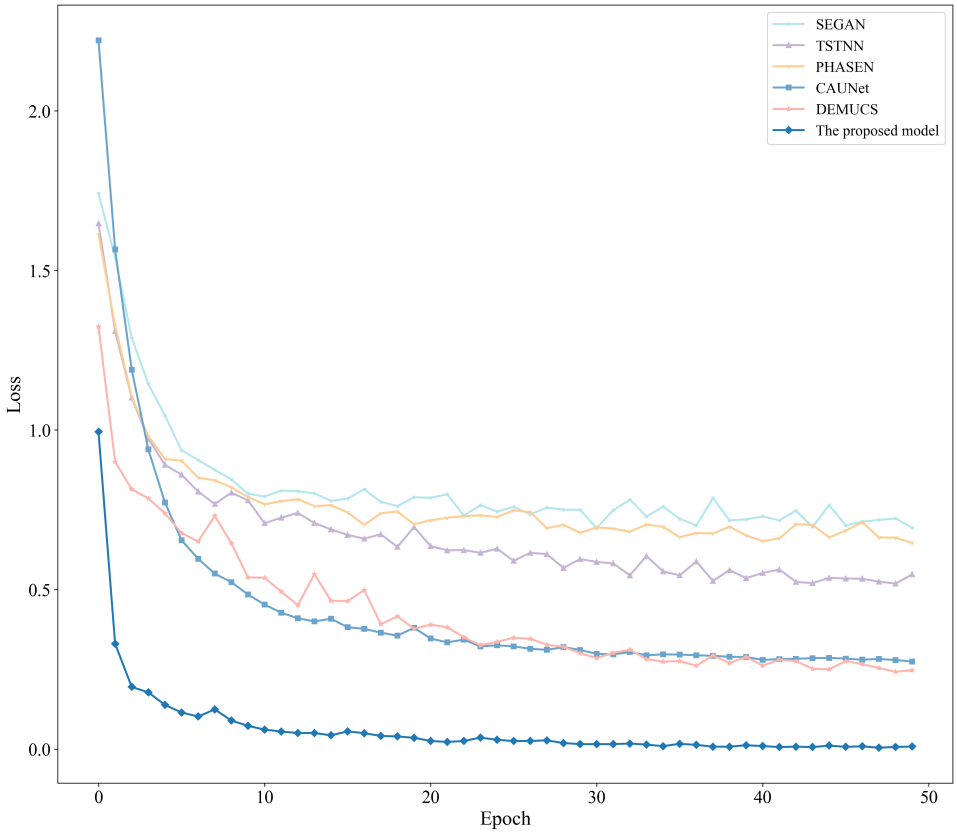
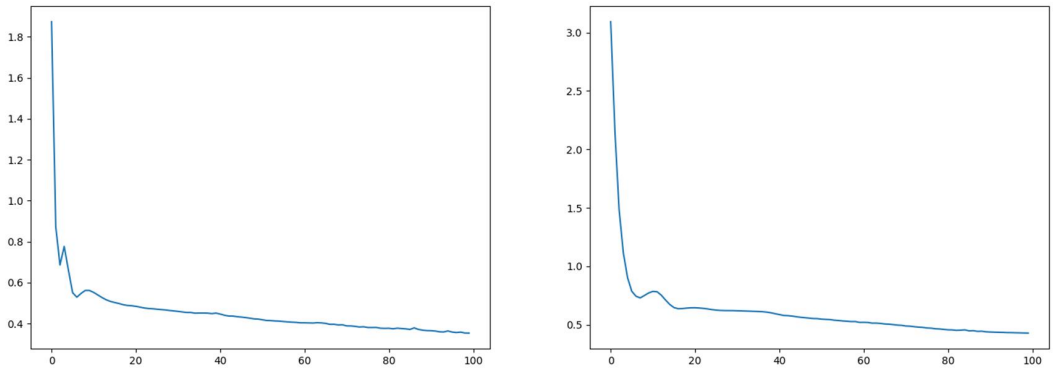


图 3.5　不同模型对比实验结果图



a. U-Net模型损失下降图　　b. 多级嵌套的U-Net模型损失下降图

图 3.6　U-Net 模型与多级嵌套的 U-Net 模型效果对比图

从损失下降图表中观察到，相比于原始的 U-Net 模型，改进的 U-Net 模型在训练过程中损失下降得更快，显示出更快的收敛速度。损失快速下降通常意味着模型能够更快地学习到数据中的特征和模式，并且可能在较短的训练周期内达到更

优的性能表现。

此外，本文做了三组实验，分别为 U-Net 原型、一级嵌套 U-Net 模型、二级嵌套 U-Net 模型（本章模型）。一级嵌套 U-Net 模型即为仅在编码阶段计算一个阶段性损失，局部更新一次权重参数，二级嵌套 U-Net 模型即为本文所提出的多级嵌套 U-Net 模型，它在编码器阶段分别计算两个阶段性损失，执行两次局部的权重参数更新。三个模型使用相同的实验设置，分别在 PESQ 和 STOI 上比较了模型的降噪性能，消融实验结果统计如表3.3所示：

表 3.3　消融实验

模型	PESQ	STOI(%)
U-Net	2.80	86.72
一级嵌套 U-Net	3.26	92.14
二级嵌套 U-Net	3.31	94.88

通过比较三个模型的得分情况，可以判断一级嵌套 U-Net 相比 U-Net 原型应用于语音降噪，有明显的性能提升，且二级嵌套 U-Net 在一定程度上进一步优化了模型。

3.5　本章小结

本章介绍了对 U-Net 进行改进并应用于语音降噪的研究。在 U-Net 网络的编解码器部分，通过在每层卷积后引入深度级别的损失函数计算和即时局部权重更新，以确保特征的完整性和提高重构精度。具体而言，每两次卷积之后进行一次反卷积操作，并计算一个局部的损失，然后更新一次权重参数，以更充分地融合特征。

该研究进行了对比实验和消融实验来验证改进模型的有效性。对比实验是将改进的 U-Net 模型与原始 U-Net 模型及其他模型进行性能对比，以评估改进的效果。消融实验则是逐步减少网络模型的嵌套等级，减少局部特征更新次数，以验证它们对模型性能的影响。通过这些实验，本研究验证了改进模型在语音降噪任务中的性能提升，并证明了改进方法对特征保留和重构精度的有效性。

第四章　基于多尺度融合的 U-Net 神经网络语音降噪方法

4.1　引言

本章节将专注于基于多尺度融合的 U-Net 神经网络语音降噪方法。具体而言，本章在第三章提出的多级嵌套 U-Net 模型的基础上，将 N 级小波分解的特征注入到 U-Net 模型的下采样层，以更全面地融合特征信息，提高语音降噪效果。

为进一步提升语音降噪效果，本文探索基于多尺度融合的 U-Net 神经网络方法。在第三章优化的模型基础上，本章引入了 N 级小波分解的特征，并将其注入到 U-Net 模型的下采样层，旨在更细致地融合多尺度的特征信息，从而提高语音重构的精度和准确性。

多尺度信息对于语音信号的处理和恢复至关重要，不同尺度的特征可以提供更丰富的语音内容表示。通过将小波分解的特征与 U-Net 模型相结合，我们可以在不同级别上捕获语音信号的细微特征，实现特征的多尺度融合和交互，从而更好地还原清晰的语音信息。

本章将深入研究多尺度融合的 U-Net 神经网络方法的原理和实现过程，探讨如何有效结合小波分解的特征和 U-Net 模型，使得模型能够更全面地理解和恢复语音信号。通过实验验证和结果分析，本研究期望展示多尺度融合 U-Net 在语音降噪任务中的优势和潜力，为改善语音重构效果提供新的思路和方法。本章将系统地介绍方法设计、实验结果和讨论分析，旨在为读者展示基于多尺度融合的 U-Net 方法在语音降噪中的创新性和实用性。

4.2　融合小波分解多级特征

多级小波分解是通过递归的方式对信号进行多个层次的分解。在每个层次上，信号都被分为两个部分：近似信号和细节信号。

这一过程使用离散小波变换（DWT）来实现，并且是基于过滤器组的思想进行的。使用两个互补的过滤器 [6]（一个低通过滤器和一个高通过滤器），以及下采

样的操作，来分别提取信号的近似和细节成分。

在多级小波分解中，离散信号 $x[n]$ 通过迭代使用低通和高通过滤器以及下采样操作分解为多个尺度的近似和细节。以下是多级小波分解的数学描述：

假设我们有滤波器对 $g[n]$（低通）和 $h[n]$（高通），它们通过卷积被应用到输入信号 $x[n]$ 上，然后进行下采样以提取近似系数和细节系数 [56]。如果我们记低频系数（近似）为 $a[n]$，高频系数（细节）为 $d[n]$，那么第一级的分解可以如下表示：

$$a_1[n] = \sum_{k=-\infty}^{\infty} x[k]g[2n-k] \tag{4.1}$$

$$d_1[n] = \sum_{k=-\infty}^{\infty} x[k]h[2n-k] \tag{4.2}$$

在这里，我们下采样因子为 2，也就是说我们只保留每个滤波器结果的偶数项。上述两式的 n 通常取偶数，这导致 $a1[n]$ 和 $d1[n]$ 的长度大约是 $x[n]$ 的一半。

对于信号的下一级分解，我们递归地应用相同的过程到近似系数 $a_{j-1}[n]$ 上。于是，对于第 j 级分解，我们得到：

$$a_j[n] = \sum_{k=-\infty}^{\infty} a_{j-1}[k]g[2n-k] \tag{4.3}$$

$$d_1[n] = \sum_{k=-\infty}^{\infty} a_{j-1}[k]h[2n-k] \tag{4.4}$$

经过 j 级的分解后，我们会得到一系列近似系数 $a_j[n]$ 和细节系数 $d_j[n]$（这里 j 从 1 到 j）。这些系数代表了原始信号在不同尺度下的信息 [57][58]。通常，近似系数 $a_j[n]$ 表示信号的粗略结构，而细节系数 $d_1[n]$ 捕获了在各个级别的高频部分，这些通常与信号中的快速变化有关。

4.3 引入多尺度融合的 U-Net 神经网络模型

小波分解是一种在多个尺度上分析信号的方法，可以有效地捕获信号的局部时频特征。将这些多级特征作为附加信息输入到 U-Net 模型的跳跃连接点，可增强

模型对细节信息的捕捉能力，这对于精准的语音信号重构尤为重要。引入多尺度融合的 U-Net 神经网络模型架构如图4.1所示：

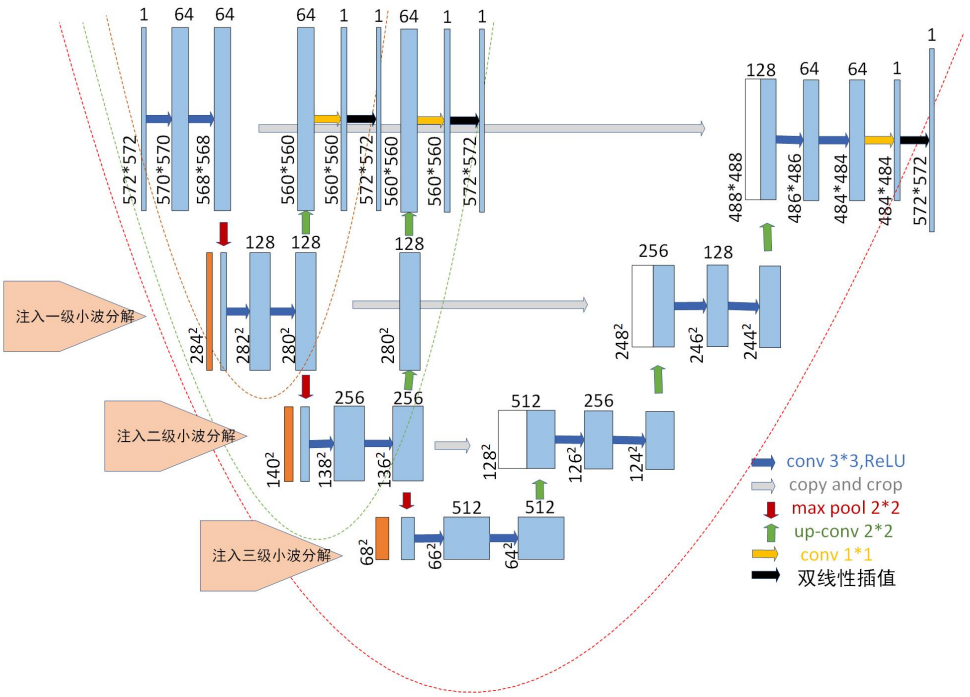


图 4.1　融合 N 级小波分解特征的 U-Net 神经网络模型架构

4.3.1　多级特征重排

在小波分解后得到的多级特征需要与 U-Net 网络中的特征图的尺寸和维度相匹配才能被注入到跳跃连接点。这种处理的目的是使得小波特征可以与卷积网络特征图进行有效地结合。下面是需要进行的处理步骤：

（1）调整特征尺寸：小波分解的各级特征可能与 U-Net 中各层的特征图尺寸不同。为了匹配大小，需要上采样或者下采样小波特征。这可以通过池化操作来实现。

（2）扩展维度：在 U-Net 的批次中使用了多个样本，则需要确保小波特征与之对应，即在每个特征图上重复小波特征以匹配批次尺寸。

（3）特征通道的整合：小波分解产生的近似系数和细节系数通常只有一个通道，而本研究使用的 U-Net 特征图有多个通道。在注入到下采样层之前，需要将小波特征沿通道维度扩展，以配合 U-Net 特征图的多通道结构。

（4）数据类型转换：确保小波特征数据类型与 U-Net 处理的数据类型一致，比如浮点数表示形式，避免数据类型不匹配带来的问题。

（5）归一化处理：重要的是，注入到网络的所有特征都应该有相似的值范围和分布，这样有助于网络的训练稳定性。根据所使用 U-Net 网络特征图的归一化方式，需要对小波特征进行归一化处理，以使其与网络特征图的值范围一致。

（6）特征选择与组合：针对不同的下采样层，可能并不需要使用所有级别的小波特征。本研究有选择性地使用部分近似系数和细节系数。

实现这些步骤后，处理过的小波特征就可以通过下采样层与 U-Net 编码器对应层的特征图进行拼接了。本研究的拼接方式是将小波特征作为额外的通道直接添加到现有的特征图上，这些调整使网络能够收到含有原信号更多完整信息的输入，可以帮助模型在进行像语音降噪这样的复杂任务时，更好地学习和重建信号。

4.3.2 注入至下采样层

根据多级嵌套的 U-Net 架构各层的特征图尺寸和小波变换的等级，确定需要将小波分解特征加入的下采样层：

（1）注入一级小波分解

语谱图经过一级小波分解（DWT）之后，将得到两个输出：近似系数和细节系数。这些系数细分了原语谱图中的不同频率成分。对于一个语谱图，假设其原始尺寸为 MxN（M 代表时间轴上的帧数，N 代表频率轴上的频带数量），一级小波分解会产生四组输出，如表4.1所示：

表 4.1 一级小波分解输出

输出	大小
近似系数 (LL)	M/2 x N/2
水平细节系数 (LH)	M/2 x N/2
垂直细节系数 (HL)	M/2 x N/2
对角细节系数 (HH)	M/2 x N/2

这些输出代表了原始信号通过低通和高通滤波器得到的不同组合，且都进行了下采样。因此，每个输出的时间维度和频率维度都是原始语谱图的一半。考虑到

多级嵌套的 U-Net 网络经过一次最大池化操作后，特征图的尺寸减半，因此，一级小波分解后输出的特征图的尺寸最接近一次池化操作后的下采样层的特征图尺寸，将一级小波分解的特征注入至一次池化操作后的下采样层最合适。如图4.2所示：

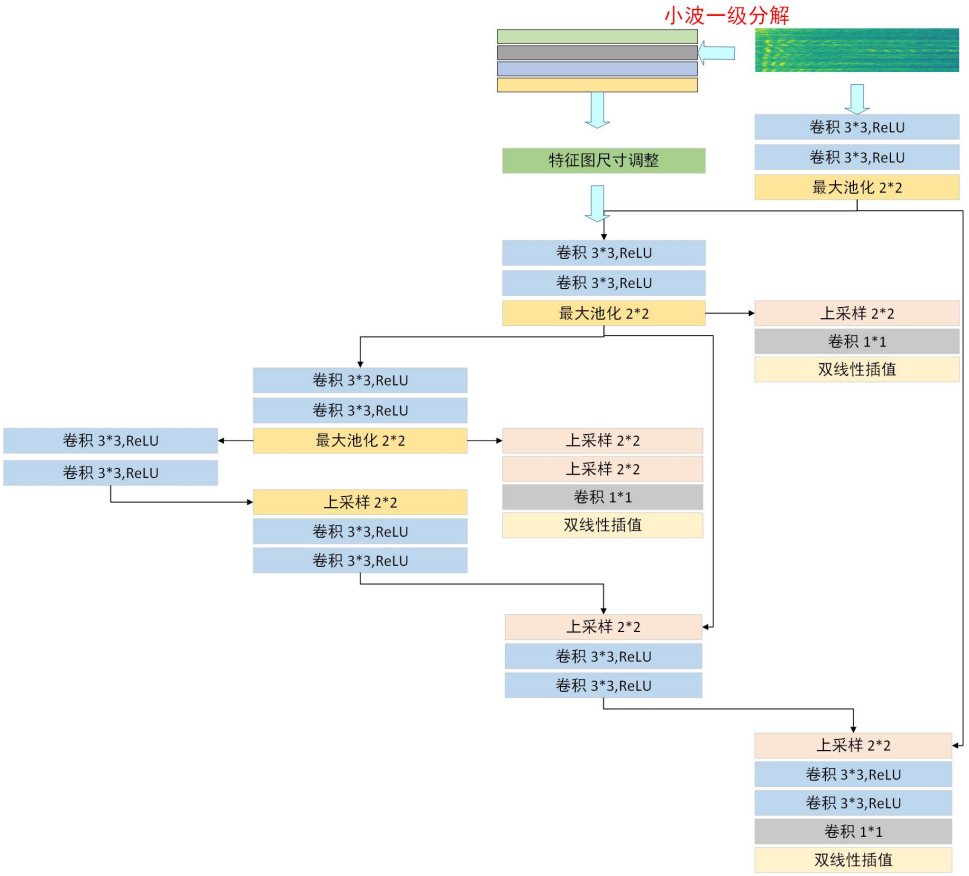


图 4.2　注入一级小波分解

（2）注入二级小波分解

二级小波分解进一步细分了一级小波分解的结果，为我们提供了更详细的频率和时间层次的信息。在二级小波分解中，我们首先对原始语谱图进行一级小波分解，然后再对一级分解的近似系数（LL）进行第二次小波分解。这样，每一级的分解都会产生四个输出，即近似系数（LL），水平细节系数（LH），垂直细节系数（HL），和对角细节系数（HH）。第二级小波分解会产生四组输出，如表4.2所示：

同样地，因为二级小波分解后输出的特征图的尺寸最接近二次池化操作后的下采样层的特征图尺寸，因此，将二级小波分解的特征注入至二次池化操作后的下采样层最合适。如图4.3所示。

表 4.2　二级小波分解输出

输出	大小
近似系数 (LL2)	M/4 x N/4
水平细节系数 (LH2)	M/4 x N/4
垂直细节系数 (HL2)	M/4 x N/4
对角细节系数 (HH2)	M/4 x N/4

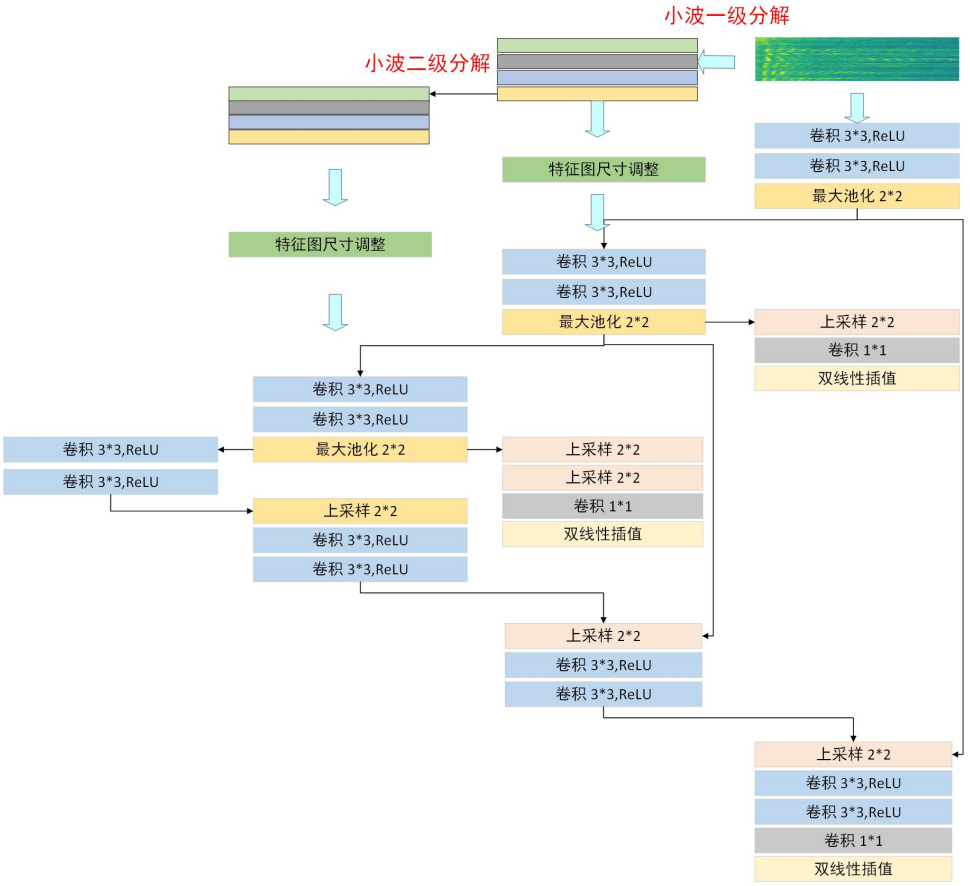


图 4.3　注入二级小波分解

（3）注入三级小波分解

三级小波分解进一步提炼了二级小波分解的结果，使我们能够获得更深层次的频率和时间信息。第三级小波分解会产生四组输出，如表4.3所示。

同样地，将三级小波分解的特征注入至三次池化操作后的下采样层最合适。如图4.4所示。

（4）特征融合：

表 4.3　三级小波分解输出

输出	大小
近似系数 (LL3)	M/8 x N/8
水平细节系数 (LH3)	M/8 x N/8
垂直细节系数 (HL3)	M/8 x N/8
对角细节系数 (HH3)	M/8 x N/8

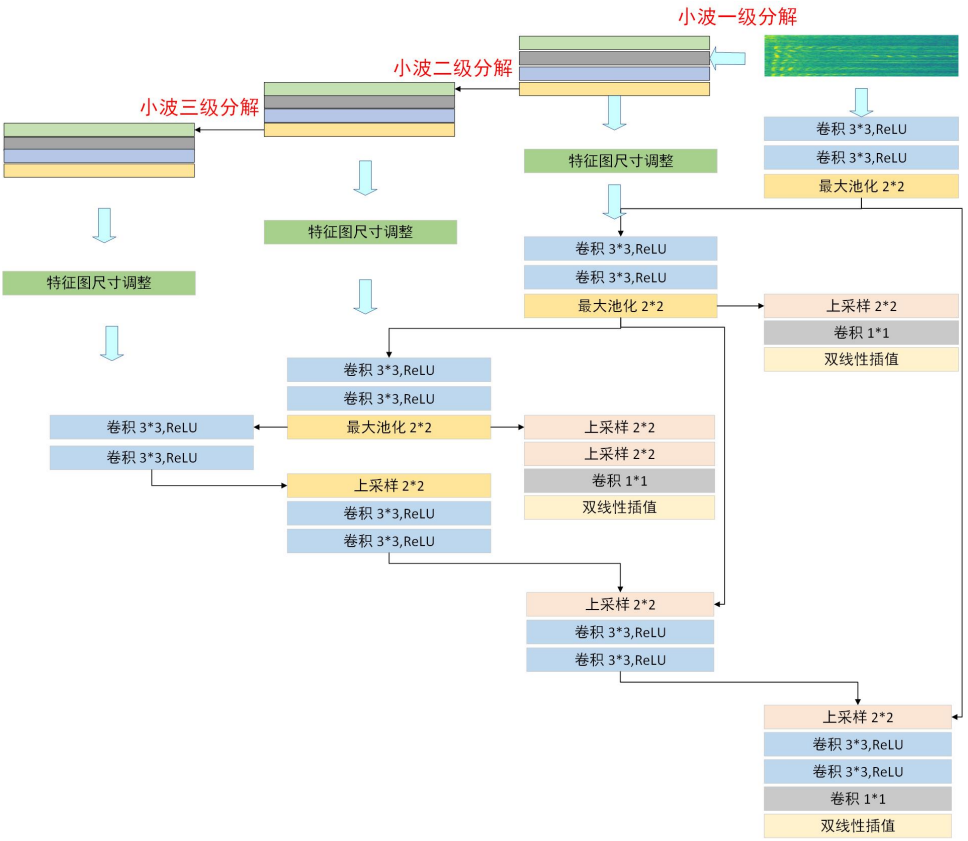


图 4.4　注入三级小波分解

特征组合后，需要一个融合步骤来整合原特征图和小波特征图。本研究采用卷积层来完成。卷积层可以学习如何综合这些特征，并减少通道数量，以使其适应后续解码器层的需要。

4.4 实验与结果分析

4.4.1 数据生成

本次实验中的纯净人声仍然使用公开的数据集 WSJ0，噪声数据集增加公开的 NOISE-92 数据集，以增强模型的泛化能力。从该数据集中选用六种噪声，分别为：babble、destroyerengine、factory1、leopard、volvo、white。此外，分别在-5dB、-0dB、5dB 的信噪比条件下对纯净人声和噪声进行融合，验证改进模型的降噪性能。其中 80% 的数据集用于训练，20% 的数据集用于测试。

4.4.2 实验设置

基于改进的 U-Net 模型，在模型下采样层引入多尺度融合。在模型中将一、二、三级小波分解的输出经过池化操作调整特征图的尺寸，然后分别注入对应的下采样层。使用训练集对改进的 U-Net 模型进行训练。选定均方差误差（MSE）作为损失函数，以及 Adam 优化器作为优化算法，初始学习率设置为 0.1，每 30 个 epoch 学习率衰减为原来的 0.5，迭代次数为 100。

4.4.3 实验与结果分析

（1）对比实验

将引入多尺度融合的 U-Net 模型分别在-5dB、0dB、5dB 的信噪比下与 SEGAN、TSTNN、PHASEN、CAUNet、DEMUCS 进行对比，分析改进模型在降噪性能上的提升，该对比实验使用了 WSJ0+NOISE-92 数据集，在不同信噪比条件下对比 PESQ 得分如表4.4所示。

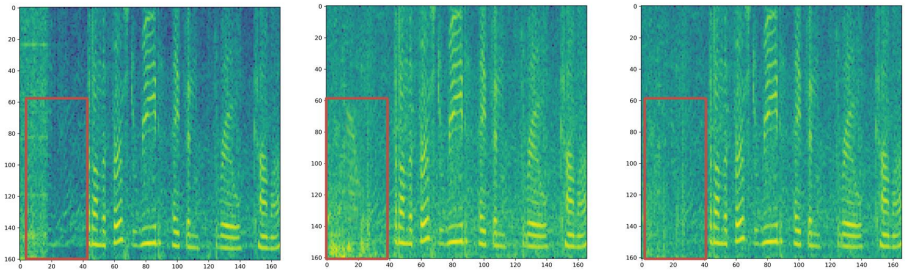
对比表4.4的结果，可看出，本文模型相较于其他对比模型，在不同信噪比条件下都取得了最好的 PESQ 分数，相较于第三章提出的多级嵌套的 U-Net 网络，模型的性能也有明显的进一步提升。

最终，使用引入多尺度融合的多级嵌套 U-Net 网络进行语音降噪前后的语谱图对比如图4.5所示，可以看出红框内的噪声被明显消除了，反映出该降噪方法的可行性。

（2）消融实验

表 4.4　不同信噪比下模型 PESQ 对比

模型	-5dB	0dB	5dB
SEGAN	2.23	2.29	2.52
TSTNN	2.26	2.53	2.97
PHASEN	2.38	2.76	3.13
CAUNet	2.77	2.83	3.09
DEMUCS	2.83	3.04	3.21
本文模型	**2.96**	**3.34**	**3.49**



a. 纯净人声语谱图　　b. 降噪前语谱图　　c. 降噪后语谱图

图 4.5　降噪前后语谱图对比

首先测试了没有小波融合的多级嵌套 U-Net 模型，记录其性能作为对照基准。然后，分别注入了 N 级小波分解的各个不同级别至改进的 U-Net 模型中，并记录了每一种设置的模型性能。进一步地，探索了注入多个小波分解级别对模型性能的影响，以找出最优的级别组合。消融实验结果如表4.5所示：

表 4.5　消融实验

模型	PESQ	STOI(%)
多级嵌套 U-net	3.29	94.83
多级嵌套 U-net + 融合一级小波特征	3.36	94.96
多级嵌套 U-net + 融合二级小波特征	3.41	95.03
多级嵌套 U-net + 融合三级小波特征	3.49	95.10

根据表4.5可以看出，当多级嵌套的 U-Net 模型在下采样层融合小波分解三级特征时，模型的 PESQ 得分最高，且 STOI 得分值也有所增长。因此，可以证实融

合多尺度特征的操作使得模型能够更全面地理解和恢复语音信号，对于提升语音降噪性能切实有效。

4.5 本章小结

本章对多尺度融合在 U-Net 模型中的应用进行了详细的讨论和实验验证，证实了小波分解的一个或多个级别结合改进的 U-Net 模型能够带来降噪性能的提升。消融实验揭示了该方法的关键因素，并验证了本章节的创新点的有效性。此外，此实验也存在一定局限性，比如进一步探索多级融合及其参数优化等，这也对未来的工作提出了方向。

第五章　总结与展望

5.1　工作总结

5.1.1　论文主要创新点

1. 引入稳定器的多级嵌套 U-Net 神经网络模型

论文改进了传统的 U-Net 结构，以通过多级嵌套 U-Net 来优化网络。具体为在编码器阶段，每经过 2 层卷积之后执行反卷积操作，并计算一个基于均方误差 (MSE) 的损失值。这个损失用于更新自网络输入到当前层的所有权重参数。此外，在解码器阶段在输出层计算一个全局的损失。此方法确保了在各个编码层中都包含了完整的特征信息，从而实现更为彻底的特征提取。

2. 引入多尺度融合的 U-Net 神经网络模型

在以上提出的改进 U-Net 模型的基础上，我们进一步融入了多级小波变换。通过以下步骤实现：首先，执行小波变换以获取信号的多级特征信息。其次，对这些特征进行适当的处理和重新排列以匹配 U-Net 网络的下采样层的维度和大小。然后，将处理过的小波特征注入到改进 U-Net 模型的下采样层，实现不同尺度特征的充分融合。这一创新的融合策略极大地提高了网络对信号细节的捕捉能力，并优化了降噪过程。

5.1.2　研究结果与结论

通过实验验证，本研究证实了改进的 U-Net 网络模型在信号重构质量上超越了传统的 U-Net 模型，且在与其他常用的语音降噪模型对比时，具有明显优势。基于 MSE 损失函数的权重更新机制理论上可以实时监控和优化特征提取过程，具体表现在重构信号的清晰度和降噪能力方面的提升。多级小波变换的引入，则进一步增强了模型对信号不同尺度和细节的处理能力，有效提升了降噪算法的整体性能。

综上，我们的研究表明，通过在 U-Net 网络中集成基于损失反馈的即时权重更新以及多级特征融合，能够显著提高信号的降噪效果，为相关领域提供了一种有效的深度学习模型改进方法。未来的研究可以进一步探索这些创新点在不同类型

的信号及嘈杂环境下的应用潜力。

5.2 未来展望

（1）模型优化和深度学习创新

未来的研究可以集中在对本研究中提出的改进 U-Net 模型的进一步优化上，探索更多种类的损失函数和网络架构的修改以提升模型性能。随着深度学习技术的不断发展，更高效的优化算法和更加复杂的网络结构，如深度可分离卷积和自适应学习率调整等技术也可以被结合到模型中以取得更好的结果。

（2）跨领域应用

改进的 U-Net 模型及其与小波变换的结合展示了在信号降噪方面的潜力。与此同时，这种方法也有可能应用于其他领域中类似的问题，例如图像和视频压缩、异常检测等。跨领域应用将极大扩展我们的研究工作的影响范围。

（3）数据集和现实场景

目前，深度学习的性能在很大程度上依赖于大量的、高质量的训练数据。在未来的工作中，研究者可以在更多样化和复杂的数据集上验证模型的泛化能力，以确保其在现实世界应用中的实用性和有效性。

参考文献

[1] 周伟. 基于结构相似性的语音信号增强 [D]. 陕西: 西安电子科技大学,2014. DOI:10.7666/d.D725931.

[2] 叶中付, 朱媛媛, 贾翔宇. 基于字典学习和稀疏表示的单通道语音增强算法综述 [J]. 应用声学,2019,38(4):645-652. DOI:10.11684/j.issn.1000-310X.2019.04.022.

[3] 小 波 变 换 [J]. 北 方 建 筑,2022,7(4):25. DOI:10.3969/j.issn.2096-2118.2022.04.009.

[4] 樊一帆, 张丽丹. 强噪环境基于谱减法的录音数字音频信号降噪. 计算机仿真,2023(11).

[5] Sisi S, Kuldip P, Andrew B. On DCT-based MMSE estimation of short time spectral amplitude for single-channel speech enhancement [J]. Applied Acoustics, 2023, 202:109134.

[6] Özen Acarbay Elif,Özkurt Nalan.Performance analysis of the speech enhancement application with wavelet transform domain adaptive filters[J].International Journal of Speech Technology,2023,26(1):245-258.

[7] Kalpana G ,Arti K .Single-Channel Speech Enhancement Using Single Dimension Change Accelerated Particle Swarm Optimization for Subspace Partitioning[J].Circuits, Systems, and Signal Processing,2023,42(7):4343-4361.

[8] 严涛, 江开忠, 姜新盈等. 基于高斯混合聚类采样的不平衡数据处理方法 [J]. 计算机应用与软件,2023,40(12):305-311.

[9] Yang X ,Liming S ,Lisby J H , et al.A speech enhancement algorithm based on a non-negative hidden Markov model and Kullback-Leibler divergence[J].EURASIP Journal on Audio, Speech, and Music Processing,2022,2022(1):

[10] Xia Y ,Wang J .Low-dimensional recurrent neural network-based Kalman filter for speech enhancement[J].Neural Networks,2015,67131-139.

[11] A regression approach to speech enhancement based on deep neural networks[J].IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),2015,23(1):

[12] A Hybrid Approach for Speech Enhancement Using MoG Model and Neural Network Phoneme Classifier[J].IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),2016,24(12):

[13] Han W ,Zhang X ,Min G , et al.A Perceptually Motivated Approach for Speech Enhancement Based on Deep Neural Network.[J].IEICE Transactions,2016,99-A(4):835-838.

[14] Multimedia; New Findings in Multimedia Described from Wuhan University (A Near-end Listening Enhancement System By Rnn-based Noise Cancellation and Speech Modification)[J].Computers, Networks Communications,2019,

[15] Li G ,Hu R ,Wang X , et al.A near-end listening enhancement system by RNN-based noise cancellation and speech modification[J].Multimedia Tools and Applications,2019,78(11):15483-15505.

[16] Mingfei S .Noise Suppression Based on RNN with a DBSCAN Classifier for Speech Enhancement[J]. 法政大学大学院紀要. 情報科学研究科編,2019,141-6.

[17] Zhenqing L ,Abdul B ,Amil D , et al.Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network.[J].PloS one,2024,19(1):e0291240-e0291240.

[18] Dayal S V ,Vaishnavi A ,Ponnapalli T , et al.Convolutional gated recurrent unit networks based real-time monaural speech enhancement[J].Multimedia Tools and Applications,2023,82(29):45717-45732.

[19] Yu W ,Zhou J ,Wang H , et al.SETransformer: Speech Enhancement Transformer[J].Cognitive Computation,2021,14(3):1-7.

[20] S G B ,Nikhil S ,A K C R , et al.A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone.[J].IEEE access : practical innovations, open solutions,2019,778421-78433.

[21] Ashutosh P ,DeLiang W .A New Framework for CNN-Based Speech Enhancement in the Time Domain[J].IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),2019,27(7):1179-1188.

[22] Nursadul M ,Soheil K ,L H J H .Convolutional Neural Network-based Speech Enhancement for Cochlear Implant Recipients.[J].Interspeech,2019,20194265-4269.

[23] Mustaqeem ,Kwon S .A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition[J].Sensors,2019,20(1):183-183.

[24] 韦高梧, 冯祖勇. 基于去噪技术的 DSP 语音识别系统设计 [J]. 传感器与微系统,2017,36(01):108-110+118.DOI:10.13873/J.1000-9787(2017)01-0108-03.

[25] 许春冬, 战鸽, 应冬文等. 基于隐马尔可夫模型的非监督噪声功率谱估计 [J]. 数据采集与处理,2015,30(02):359-364.DOI:10.16337/j.1004-9037.2015.02.014.

[26] Tu J ,Xia Y .Effective Kalman filtering algorithm for distributed multichannel speech enhancement[J].Neurocomputing,2018,275144-154.

[27] Zhang T ,Geng Y ,Sun J , et al.A Unified Speech Enhancement System Based on Neural Beamforming With Parabolic Reflector[J].Applied Sciences,2020,10(7):

[28] Xueli S ,Zhenxing L ,Shiyin L , et al.Multichannel Speech Enhancement in Vehicle Environment Based on Interchannel Attention Mechanism[J].Journal of Advanced Transportation,2021,2021

[29] Jeyasingh P ,M I M ,Madhan M .Microphone Array Speech Enhancement Via Beamforming Based Deep Learning Network[J].International journal of electrical and computer engineering systems,2023,14(7):781-790.

[30] Randall A ,Toon W V ,Marc M .Correction to: An integrated MVDR beamformer for speech enhancement using a local microphone array and external microphones[J].EURASIP Journal on Audio, Speech, and Music Processing,2021,2021(1):

[31] Firoozabadi D A ,Irarrazaval P ,Adasme P , et al.Multiresolution Speech Enhancement Based on Proposed Circular Nested Microphone Array in Combination with Sub-Band Affine Projection Algorithm[J].Applied Sciences,2020,10(11):

[32] Lin F ,Yao L L ,Lan S L , et al.Multimodal speech emotion recognition based on multi-scale MFCCs and multi-view attention mechanism[J].Multimedia Tools and Applications,2023,82(19):28917-28935.

[33] Kumar G G ,Sahoo K S ,Meher K P .50 Years of FFT Algorithms and Applications[J].Circuits, Systems, and Signal Processing,2019,38(12):5665-5698.

[34] Kuwałek P ,Jęśko W .Speech Enhancement Based on Enhanced Empirical Wavelet Transform and Teager Energy Operator[J].Electronics,2023,12(14):

[35] N. W M ,D. T A ,N. P S , et al.Drone audition: Audio signal enhancement from drone embedded microphones using multichannel Wiener filtering and Gaussian-mixture based post-filtering[J].Applied Acoustics,2024,216109818-.

[36] Burra ,Srikanth,Sankar , et al.A family of split kernel adaptive filtering algorithms for nonlinear stereophonic acoustic echo cancellation[J].Journal of Ambient Intelligence and Humanized Computing,2022,(prepublish):1-18.

[37] C. B S .Spectral Subtraction[J].SPECTROSCOPY,2021,36(5):14-19.

[38] Saleem N ,Gunawan S T ,Dhahbi S , et al.Time domain speech enhancement with CNN and time-attention transformer[J].Digital Signal Processing,2024,147104408-.

[39] Chao S ,Min Z ,Ruijuan W , et al.A convolutional recurrent neural network with attention framework for speech separation in monaural recordings[J].Scientific Reports,2021,11(1):1434-1434.

[40] Ashutosh P ,DeLiang W .Self-attending RNN for Speech Enhancement to Improve Cross-corpus Generalization.[J].IEEE/ACM transactions on audio, speech, and language processing,2022,301374-1385.

[41] Wenhao Y .Incorporating group update for speech enhancement based on convolutional gated recurrent network[J].Speech Communication,2021,(prepublish):

[42] 杨丽, 吴雨茜, 王俊丽, 等. 循环神经网络研究综述 [J]. 计算机应用,2018,38(z2):1-6,26.

[43] FELIX A. GERS, JURGEN SCHMIDHUBER, FRED CUMMINS. Continual Prediction using LSTM with Forget Gates[C]. //Neural Nets WIRN Vietri-99. 1999:133-138.

[44] Md. A R ,Salekul I ,A.K.M. I M , et al.An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition[J].Expert Systems With Applications,2023,218

[45] Junyi F ,Jibin Y ,Xiongwei Z , et al.Real-time single-channel speech enhancement based on causal attention mechanism[J].Applied Acoustics,2022,201

[46] Girirajan S ,Pandian A .Real-Time Speech Enhancement Based on Convolutional Recurrent Neural Network[J].Intelligent Automation Soft Computing,2023,35(2):1987-2001.

[47] Zheng K ,Zhihua H ,Chenhua L .Speech Enhancement Using U-Net with Compressed Sensing[J].Applied Sciences,2022,12(9):4161-4161.

[48] Ronneberger O ,Fischer P ,Brox T .U-Net: Convolutional Networks for Biomedical Image Segmentation.[J].CoRR,2015,abs/1505.04597

[49] Guimarães R H ,Nagano H ,Silva W D .Monaural speech enhancement through deep wave-U-net[J].Expert Systems With Applications,2020,158:113582.

[50] Phase-aware Speech Enhancement with Deep Complex U-Net.[J].CoRR,2019,abs/1903.03107

[51] 贾海蓉, 王卫梅, 吉慧芳. 信噪比信息与时频特征修正相位的语音增强 [J]. 西安电子科技大学学报,2019,46(05):162-170.DOI:10.19665/j.issn1001-2400.2019.05.023.

[52] Pereira E N ,Queiroz A M D ,Vieira J F , et al.Model predictive PESQ-ANFIS/FUZZY C-MEANS for image-based speech signal evaluation[J].Speech Communication,2023,154

[53] Junyi F ,Jibin Y ,Xiongwei Z , et al.Real-time single-channel speech enhancement based on causal attention mechanism[J].Applied Acoustics,2022,201:109084.

[54] Li A ,Peng R ,Zheng C , et al.A Supervised Speech Enhancement Approach with Residual Noise Control for Voice Communication[J].Applied Sciences,2020,10(8):

[55] Saadoune A ,Amrouche A ,Selouani S .Perceptual subspace speech enhancement using variance of the reconstruction error[J].Digital Signal Processing,2014,24187-196.

[56] Raj V T ,Padmakar S M .Automated Dual-Channel Speech Enhancement Using Adaptive Coherence Function with Optimised Discrete Wavelet Transform[J].Journal of Information Knowledge Management,2022,21(03):

[57] Mourad T ,Salim M B .A New Speech Enhancement Technique Based on Stationary Bionic Wavelet Transform and MMSE Estimate of Spectral Amplitude[J].Security and Communication Networks,2021,2021

[58] Bobai Z ,Qinglong L ,Qian L , et al.A Spectrum Adaptive Segmentation Empirical Wavelet Transform for Noisy and Nonstationary Signal Processing[J].IEEE ACCESS,2021,9106375-106386.