电 子 科 技 大 学

UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

专业硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE

论文题目 基于深度学习的嵌入式语音降噪系统的设计与实现

专业学位类别	**电子信息**
学　　号	**202222280507**
作者姓名	**高冬煜**
指导老师	**于力　副教授**
学　　院	**电子科技大学（深圳）高等研究院**

分类号 TP29 密级 公开

UDC[注1] 681.8

学 位 论 文

基于深度学习的嵌入式语音降噪系统的设计与实现

（题名和副题名）

高冬煜

（作者姓名）

指导老师 **于力 副教授**

电子科技大学 成都

（姓名、职称、单位名称）

申请学位级别 **硕士** 专业学位类别 **电子信息**

提交论文日期 **2025 年 4 月 15 日** 论文答辩日期 **2025 年 5 月 22 日**

学位授予单位和日期 **电子科技大学 2025 年 6 月**

答辩委员会主席 **凡时财**

评阅人 **凡时财 汪海波 邱周静子 郑宏 沈泽微**

注 1：注明《国际十进分类法 UDC》的类号。

Design and Plementation of Embedded Speech Noise Reduction System Based on Deep Learning

A Master Thesis Submitted to
University of Electronic Science and Technology of China

Discipline:	**Electronic Information**
Student ID:	**202222280507**
Author:	**Gao DongYu**
Supervisor:	**A.Prof. Yu Li**
School:	**Shenzhen Institute for Advanced Study, UESTC**

摘　要

实时语音降噪系统的研究目标是研发出能通过实时语音降噪技术提高听感和可懂度的，对环境噪音、系统设备噪声以及突发噪声进行降噪的系统，广泛应用于各大领域。但目前该领域的研究存在以下问题：第一，传统语音降噪算法虽然计算复杂度低，但在处理非稳态噪声时，面临语音降噪效果不佳的难题，同时基于深度学习的语音降噪系统在处理非稳态噪声时有更好的效果，但这些系统的复杂性总是很高，限制了在低功耗设备上的部署应用。第二，现有大多数研究在理想情况下进行，能够部署在低功耗嵌入式设备上的语音降噪算法在低信噪比环境下降噪效果不佳，且模型参数和计算量较大。第三，现有大多停留在仿真阶段，实用价值不高。为此，为了解决这些问题本文进行了深入研究。主要工作如下：

(1) 针对传统语音降噪算法无法对非稳态噪声进行精确建模的问题，本文提出了一种基于 OMLSA-IMCRA 的 OMLSA-TCN-GRU 改进算法，建立了一种融合频谱特征与梅尔频谱特征的降噪模型。本文提出的改进算法首先采用傅里叶变换将语音信号从时域转换到频域，使用轻量 TCN-GRU 网络估计带噪信号的梅尔频谱特征从而得到估计噪声谱，解决了 OMLSA-IMCRA 算法中依赖理想先验假设无法有效计算非稳态噪声噪声谱的问题。仿真实验结果表明，所提出的算法相较于原始算法有着更好的效果，在 PESQ、STOI 和 COVL 三个语音质量指标得分上，分别提高了 12.2%、6.4%、5.9%。

(2) 针对现有的语音降噪算法在低信噪比环境下降噪效果不佳，以及模型参数和计算量较大的问题，本文提出了一种基于短时离散余弦变换 (STDCT) 和注意力机制的 STDCT-DNet 模型，作为语音降噪后处理模块。经 OMLSA-TCN-GRU 模型对带噪语音的频域幅度谱进行粗修复后，该后处理模块可对带噪语音的相位信息进行恢复，同时对幅度信息进行更精确的修复，从而提高低信噪比环境下的降噪效果。仿真实验结果表明，本文提出的模型在参数更少的情况相较于其他实时语音降噪模型取得几乎一样的评价指标分数，实现相似的降噪效果。

(3) 针对现有研究大多停留在仿真阶段，实用价值不高的问题，本文实现的实时语音降噪系统，能够完成实时语音降噪，并通过对降噪算法进行并行化加速的方法部署在便携的低功耗嵌入式设备上。经实验验证，该系统能够实时高效地实现语音降噪任务，较好地实现了设计性能指标。

关键词：深度学习，语音降噪，梅尔频谱，短时离散余弦变换

ABSTRACT

The goal of real-time speech noise reduction systems is to enhance listening quality and intelligibility by reducing environmental, system, and sudden noises, making them widely applicable. However, current research faces several challenges: (1) Traditional algorithms have low complexity but struggle with non-stationary noise, while deep learning-based methods perform better but are too complex for low-power devices. (2) Most studies focus on ideal conditions, and low-power algorithms perform poorly in low SNR environments. (3) Many remain in the simulation stage with limited practical value. This study aims to address these issues through in-depth research.The main work is as follows:

(1)Aiming at the problem that traditional speech noise reduction algorithms cannot accurately model non-steady-state noise, this thesis proposes an improved OMLSA-TCN-GRU algorithm based on OMLSA-IMCRA. The model integrates spectral and Mel-spectral features, using a lightweight TCN-GRU network to estimate noise spectra. This approach overcomes OMLSA-IMCRA’s reliance on ideal prior assumptions. The simulation experiment results show that the proposed algorithm has a better effect compared with the original algorithm. In terms of the scores of the three speech quality indicators of PESQ, STOI and COVL, it has increased by 12.2%, 6.4% and 5.9 % respectively.

(2)To enhance noise reduction in low-SNR environments, this thesis proposes the STDCT-DNet model, incorporating STDCT and attention mechanisms as a post-processing module. After the OMLSA-TCN-GRU model performs coarse spectral restoration, this module refines phase and amplitude information for better noise suppression. Simulation results show that with fewer parameters, the proposed model achieves similar performance metrics to other real-time denoising models.

(3)To enhance practicality, this thesis implements a real-time speech noise reduction system optimized for low-power embedded devices through parallelized acceleration. Experiments confirm its efficiency and effectiveness in meeting design goals.

Keywords: Deep learning, Noise reduction of speech, Mayer spectrum, Short-time discrete cosine transform (STDCT)

目　录

图目录

表目录

第一章 绪 论

1.1 研究工作的背景与意义

在当今数字化时代，使用音视频多媒体已成为人们分享生活、观看精彩内容、提升生活便利度的重要方式之一。无论是游戏直播、户外探险、线上课堂，还是助听器等，音视频多媒体的内容和形式日益丰富。比如，在直播过程中，观众的体验不仅取决于画面质量，还受到声音质量的影响，直播体验的一个关键方面往往被忽视，那就是声音。清晰、自然的语音能够让观众更好地理解主播的内容，并提升沉浸感。然而，声音环境往往较为复杂，环境嘈杂声、风声、回声以及其他干扰性声音常常影响语音的清晰度和可懂度。这些噪音不仅会分散观众的注意力，还可能掩盖真正的重要信息，从而影响干净语音的整体效果。为了提升语音质量，实时语音处理系统应具备高效的语音降噪技术，使观众能够更加清晰地获取到原本的语音信号，避免因噪声干扰而产生困扰。

在语音信号处理领域，语音降噪技术的主要目标是通过一定程度上的噪声抑制技术，提高带噪语音的音质和语音可懂度，并改善人耳对带噪语音的听觉感知体验。比如在直播场景、远程会议场景、医疗场景中，语音降噪技术能够帮助主播在各种复杂环境下提供更清晰、自然的语音输出。例如，在户外直播时，风噪声和环境噪声会显著降低语音的清晰度，而在远程会议时，回声问题、键盘敲击声问题可能会影响听感。如果没有有效的语音降噪处理，可能会因为噪声干扰而破坏声音中传播的信息。

实时语音降噪技术在多行业中具有重要作用，比如在直播行业能够有效提升主播语音的清晰度，增强观众体验。对于嵌入式系统而言，单通道语音降噪方法由于计算量小、实时性高，是当前最佳的解决方案之一。随着人工智能和嵌入式计算能力的进步，未来的语音降噪技术将更加高效和智能，为依赖于声音信号的行业带来更优质的声音体验。在多种领域如电话、视频会议、无线对讲等通信场景中，语音信号往往在受到环境噪声和传输干扰的影响下，导致通话质量下降,实时语音降噪技术能够有效提升通信清晰度，改善用户体验；智能助手、语音识别、智能家居等应用依赖高质量的语音输入，实时语音降噪技术能够提高语音识别率和交互体验[1,2]；医疗行业[3]，实时语音降噪技术可用于听力辅助、远程医疗等应用，提高患者的听觉体验，某些行业的应用产品如图1-1所示。

语音降噪技术包括使用麦克风阵列的多通道算法和使用单麦的单通道算法[4]。前者相比后者而言，可以通过获取各通道信号间的空间信息来实现说话人分离和

(a) (b) (c)

图 1-1 语音降噪技术在实际场景中的应用。(a)obsbot 直播相机；(b)dji 通信耳麦；(c) 安克助听器

语音降噪[4]，但系统中拾音阵列体积大，其算法计算量和复杂度也都相对较高。这种技术虽然在效果上可能更为优越，但其对硬件和计算资源的要求也更高，可能不适合所有应用场景。

对于应用在嵌入式系统上的语音降噪系统来说，较高的算法计算量和复杂度往往带来实时性的下降。实时性是语音降噪能够应用在各领域中非常重要的一个方面，任何延迟都可能导致效果和体验的下降。因此，对于嵌入式系统，需要寻找一种既能保证语音降噪效果，又能保持实时性的技术方案。

综上所述，实时单通道语音降噪技术已经渗透进各个领域，对实时语音降噪技术的深入研究具有迫切的学术和实践意义。

1.2 国内外研究现状

单通道语音降噪由于有着只能采集一路语音信号的限制条件, 所以无法利用多个通道采集的与语音之间的相关性, 这导致通过单通道语音信号来进行降噪的挑战性更大。

(1) 国外研究现状

常见的噪声可分为加性噪声和乘性噪声，乘性噪声一般由信道不理想引起, 与信号的关系是相乘，加性噪声与信号的关系是相加, 通常是现实环境中常会产生的背景噪声[5,6]。这类噪声是以加的形式叠加在干净语音上, 从而形成带噪语音[5,6]。现有研究大多假设噪声是加性噪声。

20 世纪 60 年代国外开始研究单声道语音降噪技术，并取得了一些基础性成果，在数字信号理论逐渐成熟之后，70 年代曾形成一个理论高潮，研究者提出了一些非常经典的语音降噪算法，促使语音降噪的研究变成了语音信号处理领域中

的一个重要分支。

第一阶段的语音降噪算法是基于传统数字信号处理技术，Boll [7] 在 1979 年提出了傅里叶变换域中的谱减法，该方法在假定加性噪声信号和语音信号不相关, 估计噪声的功率谱, 接着使用带噪语音功率谱中减去估计出的噪声功率谱, 再与原带噪语音的相位结合成去噪后的语音信号。但是在输入低信噪比情况下, 残留的音乐噪声往往较大，为了改进这个问题，多种方法被提出。同年 Beroui [8] 在传统谱减法的基础上增加了两个调节系数，分别控制残留噪声的大小和语音畸变的程度，提高了谱减法的性能, 但其系数是根据经验确定的, 适应性较差。过减法 [9,10] 也是一种改进策略，该方法通过对噪声谱进行过估计，然后减去过估计值来提高谱下限，这些算法的优点在于简单易懂，且计算速度较快。然而，由于这些方法是基于一些不切实际的假设，未能充分考虑语音信号和噪声的复杂关系，因此在实际应用中，语音效果往往不尽如人意。

第二阶段的语音降噪算法基于统计模型，为了解决传统数字信号数字处理技术的不足，学者们提出了谱估计方法用来实现预先语音降噪。二十世纪四十年代，数学家 Norbert Wiener 在基于最小均方误差准则下提出维纳滤波，随后维纳滤波被广泛应用于语音信号处理之中。在 1979 年，Lim 和 Oppenheim 提出了基于维纳滤波的语音降噪方法，引入 Wiener 滤波的平滑技术来削减音乐噪声的影响 [11]，并且提出了类似于 Beroui 所改进的谱减法所使用的调节系数以提高算法的灵活性，但是会为降噪后的语音引入音乐噪声，此后多种基于维纳滤波的方法被用在语音降噪中 [12]。1984 年，Y.Ephraim 提出了最小均方误差准则短时幅度谱估计的语音降噪算法 MMSE-STSA，该算法考虑到了频谱分量的幅度对于人的听觉的影响，较好地抑制了音乐噪声 [13]，并在 1985 年对 MMSE-STSA 算法进行了改进，提出了最小均方误差短时对数幅度谱估计的语音降噪算法 MMSE-LSA [14]。但在语音不确定的情况下, MMSE-LSA 的增益函数并不是最优的。2001 年，Cohen 等对 Y.Ephraim 等人提出的 LSA 算法进行了改进，提出了基于 MMSE 最优改进对数谱幅度估计 (optimally modified log-spectral amplitude estimator, OM-LSA) 算法 [15,16]。算法能够适应多种噪声环境, 避免音乐噪声残留, 并保护较弱的语音单元，并改进了传统噪声谱估计算法在低信噪比情况下不准确的问题 [17–19]，使用最小受控递归平均算法（Minima Controlled Recursive Averaging，MCRA） [20] 及其改进方法 [21] 估计噪声谱。此后该算法逐渐作为工业界的主流语音降噪方案的中的一个重要步骤。

但是对真实的非平稳声学条件进行精确建模往往非常困难 [22]。近年来，随着深度学习技术的飞速发展，其强大的拟合能力使得单通道语音在非平稳噪声环境下的降噪性能得到了显著提升。

第三阶段的语音降噪算法基于深度学习技术。这些算法可以大致分为两个方向，即基于复域的带噪语音恢复[23]和基于时域的带噪语音恢复[24–28]。2017 年，微软研究院的 Tashev 等[29]提出了一种将深度学习与语音传统去噪结合的语音增强方法，同年，Pascual 等[24]提出 SEGAN, 利用全卷积神经网络在时域上进行语音降噪。2018 年，Valin[30]提出了一种混合数字信号处理和深度学习方法的语音增强方法，并将项目进行了开源。其效果优于传统的 MMSE 语音增强算法的同时仍能在基于 CPU 的运算上满足实时性的要求。2019 年，Yuan 等[31]提出利用 CycleGAN 针对不成对的语音数据进行语音降噪。该方法提取带噪语音的对数功率谱特征和幅度谱特征, 利用 DNN 对干净语音进行建模，同年，Nicolson 利用双向 LSTM 网络，通过输入带噪语音幅度谱来估计语音先验信噪比，并且将网络输出的先验信噪比估计值应用到基于统计模型的传统算法中[32]，其实验结果表明该算法语音增强效果优于传统算法。2020 年，Takahashi[33]等构建了 MMDenseLSTM 网络，利用 LSTM 强大的时序建模能力对时频转换后的语音信号幅度谱进行增强。2021 年 Pandey 等[34]使用注意力机制和稠密连接在时域上估计干净语音, 并且同时用来估计噪声。2022 年，Schroter 等人提出了 Deepfilternet2 模型[35]算法，该算法成为复杂度低、高性能的音频增强工具，可以部署在低功耗嵌入式设备上。

(2) 国内研究现状

国内语音降噪算法和技术的研究较国外来说起步较晚，但在进入二十一世纪以来也发展迅速。

早期国内研究现状也是基于传统信号处理算法和统计模型，2002 年，侯正风[36]提出了综合应用小波变换和维纳滤波的降噪方法，能较好的抑制白噪声的同时抑制传统维纳滤波产生的音乐噪声。2004 年，蔡斌[37]等对 MMSE 语音降噪方法进行了改进，能较好的抑制音乐噪声。2011 年，刘凤增[38]等提出了一种结合 OMLSA 与小波阈值去噪的语音降噪算法，能够高效的去除 OMLSA 算法的残留噪声。2017 年，张建伟[39]等提出了一种基于改进谱平滑策略的 IMCRA 算法，能够更好地跟踪噪声信号变化，改善语音质量。

国内研究的第二阶段是基于深度学习，对于深度学习在语音降噪领域中的应用，国内研究发展迅速，发展方向同样是基于复域的带噪语音恢复[40,41]和基于时域的带噪语音恢复[42]。2016 年，贾海蓉[43]等提出了一种基于 DNN 的子空间语音降噪算法，该算法在测试增强阶段根据噪声估计和 DNN 模型去除非平稳噪声，提高了语音可懂度。2018 年，阴法明等[44]提出了一种基于深度置信网络的语音降噪算法, 增强后的语音质量优于 LOGMMSE 与 OM-LSA 算法。MetricGAN 的生成器利用 BLSTM 强大的序列建模能力对语音进行增强。2021 年，Fu 等[45]提

出将 Transformer 模型中的位置编码改为卷积层用于语音降噪，再使用 Boosting 策略 [46], 根据 MetricGAN 的框架对 Transformer 模型进行微调。随后，Fu 等 [47] 提出的 MetricGAN+ 进一步提升了语音降噪的能力，并在公开数据集 VoiceBank-DEMAND 上达到了当时的最高效果 (state of the art,SOTA) 2021 年，Li 等人提出了两阶段算法 [48]，将原来的单阶段优化任务分解为两个更简单、更渐进的子任务。具体来说，第一阶段负责幅度谱增强，以粗去除噪声。随后，第二阶段进一步预测复频谱的残差分量，从而抑制残差噪声，恢复语音相位。在此之后相位恢复得到了更多的关注 [49–53]。

1.3 本文的研究内容与创新

本文根据目前嵌入式实时语音降噪系统存在的问题，尝试以算法研究和系统实现为核心，重点针对语音降噪算法和算法计算复杂度两部分进行改进，并设计了基于海思 SS928 芯片平台的语音降噪系统，在该系统上对改进的算法进行效果验证。本文追求算法实现和工程应用上共同实现，文章的主要研究内容如下：

1. 针对传统语音降噪算法无法对非稳态噪声进行精确建模而基于深度学习的语音降噪系统难以部署在低功耗设备上的问题，本文提出了一种基于 OMLSA-IMCRA 的 OMLSA-TCN-GRU 改进算法，建立了一种融合频谱特征与梅尔频谱特征的降噪模型。本文提出的改进算法首先使用快速傅里叶变换，将时域的带噪语音信号转换到频域，使用轻量 TCN-GRU 网络估计带噪信号的梅尔频谱特征从而得到估计噪声谱，解决了 OMLSA-IMCRA 算法中依赖理想先验假设无法有效计算非稳态噪声噪声谱的问题。并且通过仿真实验证明了所提出的改进算法的优越性。

2. 针对现有的能够部署在低功耗嵌入式设备上的语音降噪算法在低信噪比环境下降噪效果不佳，且模型参数和计算量需求较大的问题，本文提出了一种基于短时离散余弦变换 (STDCT) 的 STDCT-DNet 模型，作为语音降噪后处理模块。经 OMLSA-TCN-GRU 模型对带噪语音的频域幅度谱进行粗修复后，该后处理模块可对带噪语音的相位信息进行恢复，同时对幅度信息进行更精确的修复，从而提高低信噪比环境下的降噪效果。最后通过仿真实验结果表明，本文提出的模型在参数更少的情况相较于其他语音降噪模型取得几乎一样的评价指标分数，实现相似的降噪效果。

3. 搭建了基于海思 SS928 芯片平台的自动对焦系统，完成包括音频设备驱动、数模转换模块在内的多电路软硬件设计，编写算法程序，并对算法计算进行并行加速，最终完成了基于该实时语音降噪系统的功能测试和算法性能测试，证明了该系统的实用性。

1.4 本论文的结构安排

本文一共六个章节，具体结构安排如下：

第一章，绪论。本章主要阐述本课题研究的背景与意义，系统回顾了国内外在语音降噪系统方面的研究现状，分析了该技术的重要性及其广阔的应用前景。最后，对本文的整体结构进行了概述，明确了各章节的主要内容和组织框架。

第二章，相关理论基础。从语音分帧、语音加窗、常见的声学特征方面详细介绍了语音信号处理相关理论；介绍了几种常用的深度学习模型，为后文所建立的基于深度学习的语音降噪模型的研究与设计奠定了基础。

第三章，基于 TCN-GRU 网络的改进 OMLSA-IMCRA 算法。探究了传统 OMLSA-IMCRA 算法在处理非稳态噪声时的不足，基于此设计了一种基于 TCN-GRU 网络的改进 OMLSA-IMCRA 算法，搭建了基于 TCN-GRU 网络与 OMLSA 幅度增益估计法的联合降噪模型，同时引入混合注意力机制进一步提高了模型性能。

第四章，基于 STDCT 变换的语音增强后处理模型。引入了低复杂度的 STDCT-DNet(Short Time Discrete Cosine Transform Denoise Net) 网络模型作为后处理模块来进一步抑制第三章模型输出中的残余噪声，提高模型的抗噪性能和降噪后语音的主观质量，所使用的网络模型在第三章提出的网络模型的基础上引入注意力机制进一步提高了模型性能，提高了在非理想条件下降噪性能。

第五章，实时嵌入式语音降噪系统实现。首先基于海思 SS9278 芯片平台完成了能够完成语音信号输入采集、语音信号预处理、语音降噪算法模型推理、结果保存的嵌入式语音降噪系统的软硬件设计实现，然后在设备启动并运行算法时，在在实际会议场景下测试算法并验证了嵌入式单通道实时语音降噪系统的实时性能以及降噪性能。

第六章，全文总结与展望。本文对前几章的研究工作进行了系统总结，全面回顾了本课题的研究内容与主要成果。在此基础上，分析了当前工作中存在的不足与限制，并结合实际应用需求，提出了若干具有发展潜力的改进方向和未来研究的重点，为后续的深入探索奠定了基础。

第二章 相关理论基础概述

语音信号处理，是信号处理领域中非常重要的技术，对于各类基于语音信息的应用有着重要意义。在当今数字化时代，语音交互技术在人机交互、智能语音助手、语音识别、语音合成以及通信系统等众多领域中扮演着核心角色。本章围绕语音信号分帧、加窗、常见的声学特征以及深度学习模型等方面展开相关探讨。语音信号分帧是将连续的语音信号划分为短时间内的小段，以便对每一帧进行独立处理，这是后续分析的基础；加窗则是为了减少帧边缘的不连续性，从而降低信号处理中的伪影。常见的声学特征提取，如语谱，梅尔频谱等，能够从语音信号中提取出最具代表性的信息，为后续的分析和识别提供关键依据。而深度学习模型的引入，更是为语音信号处理带来了革命性的变化，它能够自动学习复杂的语音特征，极大地提高了语音处理的性能和效率。本章旨在为后续的语音降噪算法提供坚实的理论基础和实践指导，为后续的算法设计和优化奠定坚实的基础。

2.1 语音信号处理相关理论

2.1.1 语音分帧

由于语音信号在整体上呈现出非平稳特性，其本质特征参数具有明显的时变性，因此难以直接采用传统数字信号处理方法进行有效处理。然而，语音的产生是由于人类口腔肌肉的运动构成了特定的声道形状，从而激发出相应的语音响应。需要注意的是，与语音信号的频率相比，这种肌肉运动的变化过程相对缓慢，因此，从另一个角度来看，在较短的时间范围内（通常为 10～30ms），语音信号具有短时平稳性。

因此，语音信号的分析与处理需基于‘短时分析’原理，即将整体语音信号划分为若干时长较短的帧（通常为 10～30ms），在每一帧内提取相对稳定的特征参数，从而实现对非平稳语音信号的有效建模与处理。帧长一般截取为 10 35Ms。因此，从整体上看，语音信号在经过短时分析后，其特征提取过程实际上形成了由帧级特征参数组成的时间序列，从而为后续建模与识别提供基础。在语音信号处理中通常采用短时分析的方法，即将语音信号划分为多个短时帧（Frame），以便进行后续处理。语音分帧是语音信号处理中的关键步骤，它的合理性直接影响到后续处理的效果。

分帧过程中通常采用重叠分段的策略，以保证帧间的平滑过渡和信号的时间连续性。相邻帧之间的时间间隔，即重叠滑动的步长，称为‘帧移’。。分帧的主

要参数包括帧长（Frame Length）即单帧包含的采样点数，通常取 10ms 至 30ms；帧移（Frame Shift），即相邻帧的起始位置间隔，通常为帧长的 1/2 或 1/3；重叠率（Overlap Rate），即相邻帧之间的重叠部分比例，这会影响平滑性。而分帧方法一般会采取固定长度分帧或自适应分帧方法，前者所有帧的长度相同，帧移也固定不变。这种方法计算量较小，适用于大多数语音处理任务。后者根据语音信号的特性（如短时能量变化）动态调整帧长，以便更精确地捕捉语音的变化，但是计算量较大。图2-1是语音信号分帧示意图。

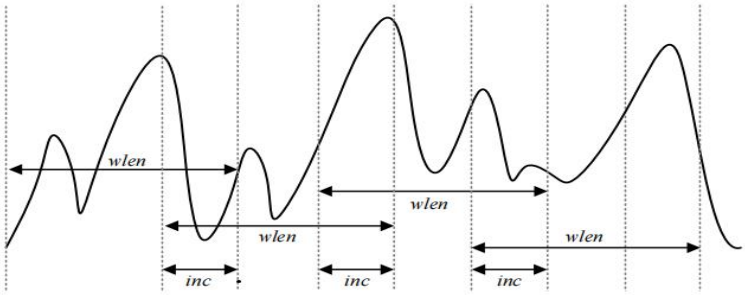


图 2-1 信号分帧示意图

2.1.2 语音加窗

语音信号具有短时平稳性（通常认为 10-30ms 内近似平稳），但长时非平稳。为了分析处理，需将信号分帧，但直接分帧会导致帧两端突变，傅里叶变换假设信号是周期无限的，而实际分帧后的信号边界存在不连续点，导致频域出现虚假频率分量。引发频谱泄漏（Spectral Leakage）。加窗的核心目的是通过平滑帧边缘，降低截断效应带来的频域干扰。给每一帧语音信号加窗的过程是将每帧与一个窗函数相乘，其计算公式可以由式2-1表示，式中 $x(n)$ 为加窗后的音频函数，$s(n)$ 为原始语音信号，$w(n)$ 为窗函数。

$$x(n) = s(n) * w(n) \tag{2-1}$$

语音信号处理加窗中常用的窗函数包括矩形窗、汉宁窗（Hanning）、汉明窗（Hamming）、布莱克曼窗（Blackman）等。它们时域和频域特性各有不同，适用于不同的应用场景。

(1) 矩形窗

矩形窗是最简单的窗函数，其所有值均为 1, 矩形窗的主瓣宽度较窄，但旁瓣幅度较高，频谱泄漏较为明显, 其计算公式如式2-2所示。

$$w(n)=\begin{cases}1, & 0\le n\le N-1\\ 0, & \text{otherwise}\end{cases} \tag{2-2}$$

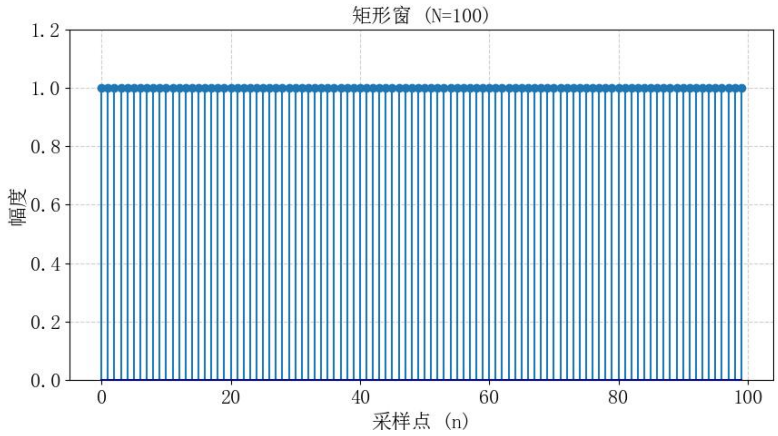


图 2-2 矩形窗示意图

(2) 汉宁窗

汉宁窗是一种余弦窗，两端平滑过渡到零，汉宁窗的主瓣宽度较宽，但旁瓣幅度较低，能够有效减少频谱泄漏，其计算公式如式2-3所示。

$$w(n)=\begin{cases}0.54-0.46\cos\left(\frac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & \text{otherwise}\end{cases} \tag{2-3}$$

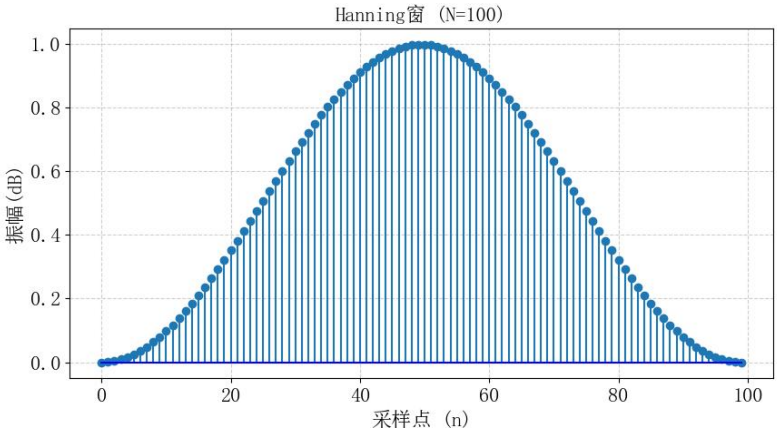


图 2-3 hanning 窗示意图

(3) 汉明窗

汉明窗与汉宁窗类似，但其两端不为零，汉明窗的主瓣宽度与汉宁窗相近，但旁瓣幅度更低，适合频谱分析，其计算公式如式2-4所示。

$$w(n)=\begin{cases}0.5-0.5\cos\left(\frac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & \text{otherwise}\end{cases} \tag{2-4}$$

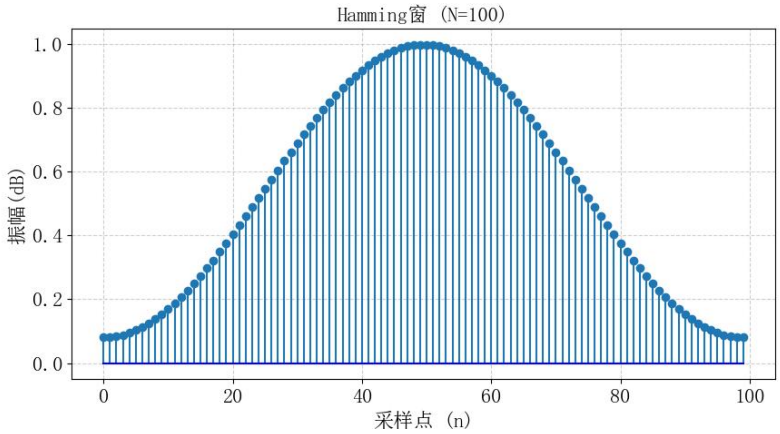


图 2-4 hamming 窗示意图

(4) 布莱克曼窗

布莱克曼窗是一种三余弦窗，具有更宽的主瓣和更低的旁瓣, 布莱克曼窗的频谱泄漏最小，但主瓣宽度较大，计算复杂度较高，适用于测量单频信号，寻找更高次谐波以及高精度频谱分析，其计算公式如式2-5所示

$$w(n)=\begin{cases}w(n)=0.42-0.5\cos\left(\frac{2\pi n}{M-1}\right)+0.08\cos\left(\frac{4\pi n}{M-1}\right), & 0\le n\le N-1\\ 0, & \text{otherwise}\end{cases} \tag{2-5}$$

2.1.3 常见的声学特征

声学特征的选择和优化在语音降噪任务中对提高语音增强效果及其质量起着非常重要甚至决定性作用。声学特征能够有效地从噪声较强的语音信号中提取有用的语音信息，因此，在面对噪声较大或干扰较强的语音信号时，选择鲁棒性更强的特征显得尤为重要。此外，对于实时语音增强应用，如何在保证降噪效果的同

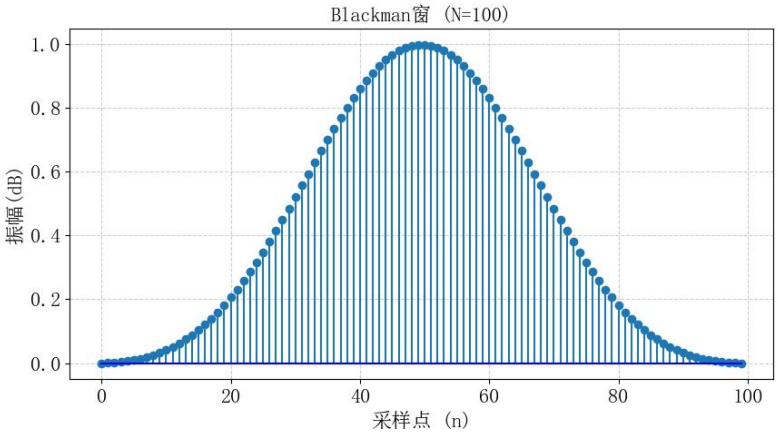


图 2-5 blackman 窗示意图

时，实现低延迟的实时处理，也成为了一个关键问题。因此，声学特征不仅需要具备较强的信息提取能力，还需考虑处理效率，以满足不同应用场景的需求。根据特征在时间和频率域上的表现，声学特征可以大致分为时域特征和频域特征两类，每一类特征在语音增强中的作用各不相同，优化这些特征的提取与处理方法，将直接影响语音增强系统的性能表现。

(1) 时域特征

时域特征是通过分析语音信号的波形形状来描述其特性，主要反映语音信号在时间轴上的变化模式。这类特征通常可以通过时域滤波和统计分析方法提取，能够提供关于语音信号能量分布、振幅变化以及周期性的关键信息。在语音增强中，常见的时域特征包括短时平均幅度、短时能量、自相关函数和过零率等。其中，短时能量用于衡量语音信号在短时间窗内的能量分布，短时平均幅度反映信号振幅的变化趋势，过零率用于描述信号的频率成分和噪声特性，而自相关函数则用于分析语音信号的周期性特征。通过合理选择和优化时域特征，可以更有效地提升语音增强算法的性能，尤其是在低信噪比环境或特定语音处理任务中。

(2) 基于傅里叶变换的频域特征

声音信号本质上是沿时间轴变化的一维时域波形，其频率成分的动态变化难以通过原始波形直观呈现。传统傅里叶变换虽能将信号映射至频域揭示其频谱特征，但这种全局视角的频谱分析本质上剥离了时间维度，导致频率分布的时间演化规律无从观测。为此，现代信号处理领域发展出一系列时频分析方法，通过时频联合域分析实现对信号频率特征的动态追踪，有效结合了时域与频域的双重信息优势。

1) 语谱图（spectrogram）

语谱图是一种用于表示语音信号频谱随时间变化的图形，其横轴表示时间，

纵轴表示频率，而不同频率成分在特定时刻的强弱则通过灰度或色调的变化来直观呈现。作为一种结合了频谱分析和时域波形信息的可视化工具，语谱图能够清晰地展示语音信号的动态频谱特性，使语音在时间轴上的频率分布及其变化趋势一目了然。

语谱图不仅包含了丰富的语音学信息，还能够反映语音的语句特性，例如语音的共振峰、基音频率、音素边界以及发音方式等关键特征。现代语谱图分析技术的进步，使得语音研究人员能够更加精准地解读语音的发声特征，并利用这些信息进行语音识别、语音增强、说话人识别等任务。一些经验丰富的语音学家甚至可以通过观察语谱图的形态，结合语音学知识，对语音的产生过程进行分析和解释。这种技术的应用不仅有助于语音信号处理的优化，还在语言学研究、语音病理分析以及人机交互等领域发挥着重要作用。

一段带噪语音波形图和语谱图如图2-6所示，它的横轴表示时间，纵轴表示频

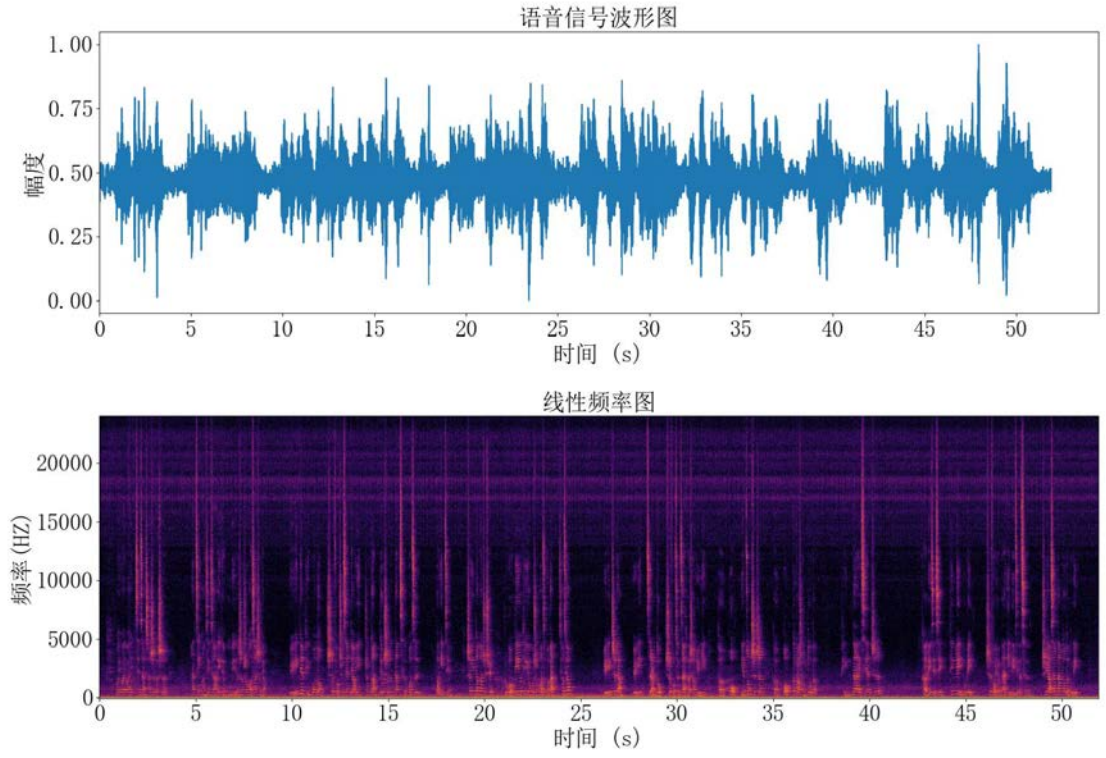


图 2-6 带噪语音波形-语谱图

率，而颜色则表示信号的能量，用时间 n 作为横坐标，w 作为纵坐标，将 Pn(w) 的值表示为灰度级所构成的二维图像即语谱图 (Spectrogram)，为时域语音信号经过短时傅里叶变换之后所得。其计算公式如式2-6所示，其中 $W[n]$ 是窗函数，R 是重叠步长，N 是窗长度。

$$X[k,m]=\sum_{n=0}^{N-1}x[n]\cdot w[n-mR]\cdot e^{-j\frac{2\pi}{N}kn} \tag{2-6}$$

2) 梅尔频谱

为了得到合适大小的声音特征，往往把语谱信息通过梅尔标度滤波器组（mel-scale filter banks），变换为梅尔频谱。梅尔频谱（Mel Spectrogram）是语音处理、语音识别和音频分析中常用的一种特征表示方法。梅尔频率尺度是基于人类听觉的特性设计的。人耳对低频的变化更为敏感，而对高频的变化不那么敏感。梅尔频率通过对线性频率轴进行压缩，使得较低频率部分更精细，而较高频率部分则被压缩。梅尔频率尺度的这种设计使得梅尔频谱比传统的线性频谱更能反映出人类听觉系统的特点，从而能够更好地捕捉语音和音乐中的有用信息。在语音信号中，许多信息主要集中在较低的频率范围，而较高的频率部分相对较少且不太重要。梅尔频谱通过将频率信息映射到梅尔尺度，能够有效地提取与语音相关的特征，梅尔频率（fm ）与线性频率（f）之间的关系可以通过式2-7表示.

$$f_m=2595\cdot\log_{10}\left(1+\frac{f}{700}\right) \tag{2-7}$$

与此同时使用梅尔频谱可以帮助降维和去除一些高频噪声，通过对频谱进行梅尔尺度的转换，可以减少高频部分的干扰，并使得低频信息更加突出，适用于语音增强和噪声抑制任务，在深度学习模型中广泛应用，尤其是在语音识别和语音降噪任务中。梅尔频谱能够提取出有意义的特征，同时可以减少数据的冗余。梅尔频谱的计算依赖于梅尔滤波器组。梅尔滤波器组通常由一组重叠的三角形滤波器组成，这些滤波器的中心频率均匀分布在梅尔尺度上，如图2-7所示。

每个滤波器的带宽和形状旨在模拟人耳对音频信号的响应，通过梅尔滤波器组计算梅尔频谱的计算公式如式2-8所示。式中 $M[m]$ 表示梅尔频谱在第 m 个梅尔频率上的能量值，$\sum_k H_m(f_k)$ 是第 m 个梅尔滤波器在频率 f_k 表示该频率对梅尔频谱的贡献。每个滤波器的响应通常是三角形的，且在梅尔尺度上均匀分布。$|X[k]|^2$ 是频谱的能量，表示在频率 f_k 上的信号能量。

$$M[m]=\sum_k H_m(f_k)\cdot|X[k]|^2 \tag{2-8}$$

一段带噪语音波形图和梅尔频谱图如图2-8所示。

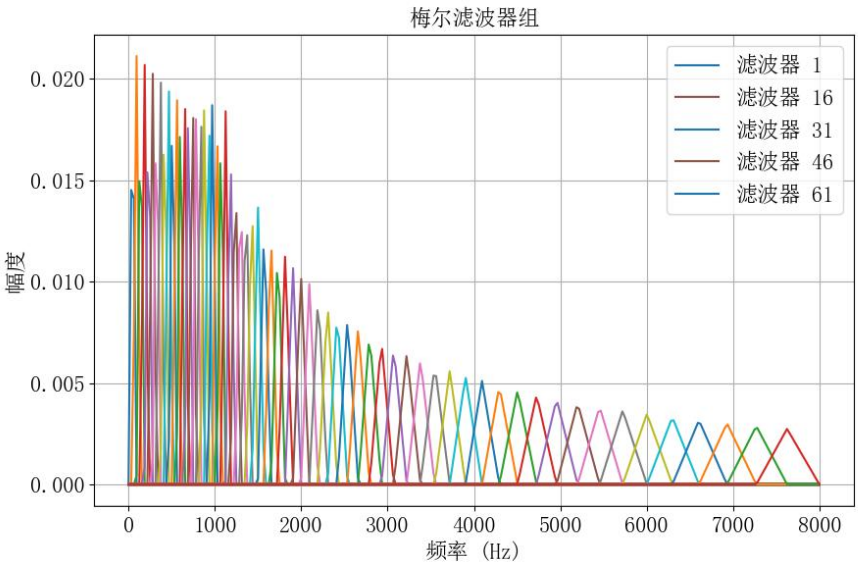


图 2-7 梅尔滤波器组图

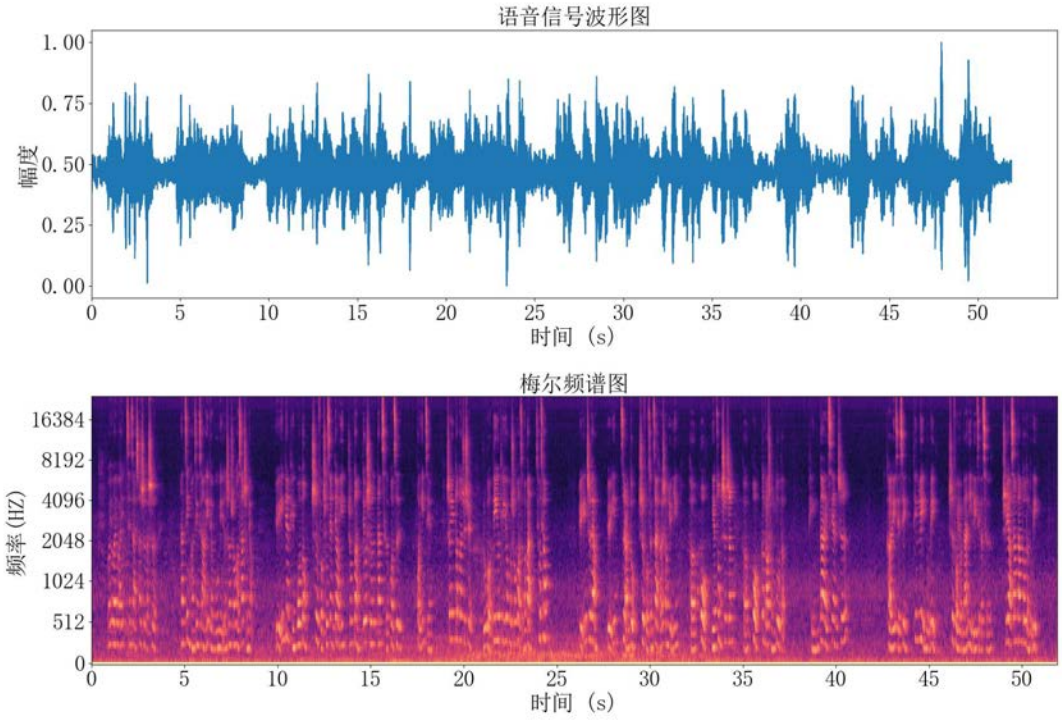


图 2-8 带噪语音波形-梅尔频谱图

(3) 基于余弦变换的频域特征

STDCT 对信号进行短时分析，将其分为短时间帧并对每帧应用离散余弦变换，有效表示信号的频谱特性，尤其是在信号能量集中在低频部分时。短时离散余弦变换特别适用于信号的压缩，因为它能够将信号的能量集中到少数几个低频系数

上，这对于语音编码如 MP3 或音频压缩格式非常有效。较高频率的系数通常包含较少的信息，可以去除或量化，达到压缩效果，在许多情况下提供更好的频域分辨率，这使得它在语音处理中比傅里叶变换更能适应非平稳信号的特性。

2.2 深度学习模型

2.2.1 卷积神经网络

(1) 传统卷积神经网络

卷积神经网络 (CNN) 是深度学习中一种极具代表性的算法，广泛应用于音频多个领域如语音降噪、语音增强、语音识别等领域。卷积神经网络的核心架构通常是卷积层、池化层和全连接层三个部分。

CNN 可以通过卷积层中的卷积核计算感受野内数据的点积，并通过激活函数进行非线性变换，常用的激活函数有 Sigmoid 函数，Tanh 函数，ReLU（Rectified Linear Unit）函数。这一过程能够有效提取输入数据中的局部模式或特征，如图像中的边缘、纹理或其他形状信息。

池化层一般位于卷积层之后，主要功能是对卷积层输出的特征图进行降采样，减少数据的空间维度，同时降低计算量和网络的参数数量，进而提高计算效率，减少过拟合现象。全连接层通常位于 CNN 的末尾，可以通过连接的权重矩阵和激活函数的作用，最终输出预测结果。

全连接层常常用于最终的分类或回归任务，它的作用是将经过卷积和池化处理后的高维特征图映射到具体的类别标签上。通常，经过若干个卷积层和池化层后，提取到的特征图会被展平（Flatten），然后输入到全连接层进行进一步的处理。

(2) 时域卷积网络

时域卷积网络（Temporal Convolutional Network，TCN）由 Shaojie Bai et al. 在 2018 年提出的，可以用于时序数据处理[54]。深度学习背景下的序列建模主题主要与递归神经网络架构 (如 LSTM 和 GRU) 有关，S. Bai 等人认为，这种思维方式已经过时，在对序列数据进行建模时，应该将卷积网络作为主要候选者之一加以考虑，使用卷积网络而不是递归网络可以提高性能，因为它允许并行计算输出，这可以充分利用现代硬件的并行加速能力。

在语音任务中，常常处理一维信息。对一维信息进行卷积称为 1D 卷积，大值方法可以分为全 (full) 卷积、同 (same) 卷积和有效 (valid) 卷积。时域卷积网络由具有相同输入和输出长度的扩张的、因果的 1D 卷积层组成，其可以用于时序数据处理依赖于因果卷积，膨胀卷积、残差连接这三个特性。

因果卷积如图2-9所示。

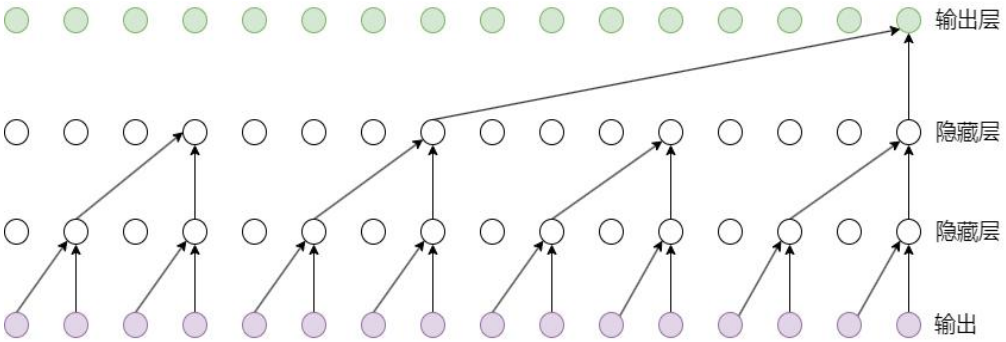


图 2-9 TCN 因果卷积示意图

因果卷积（Causal Convolution）在时间序列建模中具有严格的时序依赖特性。与传统的卷积神经网络不同，因果卷积在计算当前时刻 t 的输出时，仅依赖于该层在时刻 t 及其之前的输入，而不考虑未来的信息。这种单向的信息传递机制确保了模型在处理序列数据时遵循“先因后果”的时间逻辑，避免了对未来数据的泄露，因此被称为因果卷积。它是一种具备严格时间约束的建模方式，特别适用于需要保持时间一致性的应用场景。

膨胀卷积（Dilated Convolution），又称为空洞卷积，是在因果卷积基础上的一种扩展。虽然因果卷积能够保证时间顺序的合理性，但其感受野仍然受到卷积核大小的限制，难以有效建模较长时间跨度的依赖关系。为了解决这一问题，引入了膨胀卷积机制，通过在卷积操作中引入间隔（即空洞），在不增加参数数量的前提下扩大感受野，从而提升模型对长时依赖特征的捕捉能力。为了解决这个问题，研究人员提出了膨胀卷积，如图2-10所示。

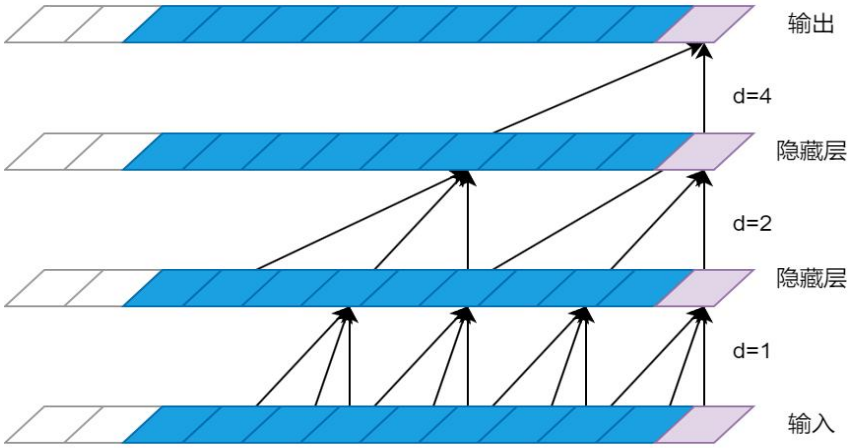


图 2-10 TCN 膨胀卷积示意图

膨胀卷积允许在卷积期间对输入进行间隔采样，并且采样率由 d 控制。底层的 $d=1$ 表示在输入的过程中对每个点进行采样，中间层的 $d=2$ 表示在输入过程

中对每 2 个点采样一次作为输入。层级越高，d 的数值越大。通过这种方法，卷积网络可以使用较少的层，就能获得大的感受野。

残差连接被证明是训练深度网络的有效方法，它允许网络以跨层方式传输信息，在时间卷积网络中同样采用了残差连接，如图2-11所示。

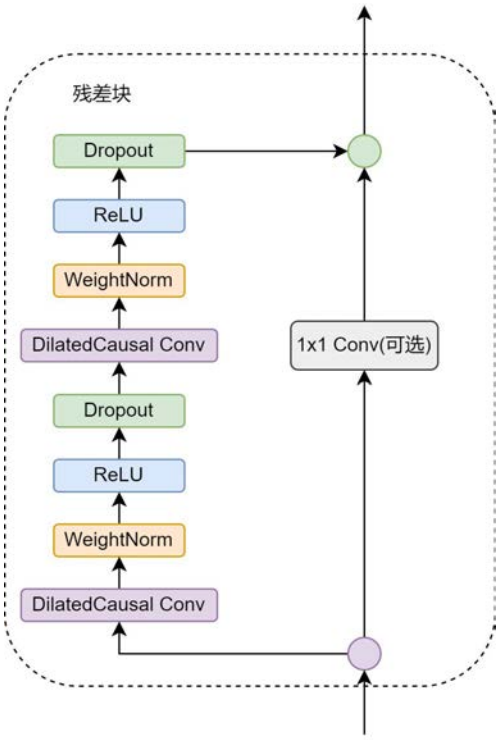


图 2-11 TCN 残差连接示意图

2.2.2 循环神经网络

循环神经网络（Recurrent Neural Networks, RNN）作为一种具备短时记忆能力的神经网络结构，在处理序列问题中展现出显著优势，因而在自然语言处理等领域得到了广泛应用。RNN 通过其独特的反馈连接机制，能够在处理序列数据时有效捕捉序列中各元素之间的时序关系和依赖性，从而提升模型对上下文信息的理解能力。在处理音频信号输入时，RNN 能够充分挖掘当前时刻数据输入和前后时刻的数据间的依赖关系。RNN 结构示意图如图2-12所示。

图中 X_t 表示 t 时刻的输入，S_t 表示 t 时刻的隐藏状态，O_t 表示 t 时刻的输出，公式分别如式2-9和式2-10所示。

$$s_t = f(Ws_{t-1} + Ux_t + a) \tag{2-9}$$

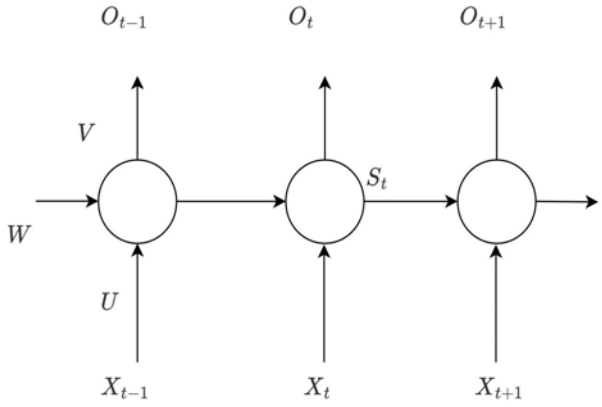


图 2-12 RNN 结构示意图

$$o_t = g(Vs_t + b) \tag{2-10}$$

式中，U、W、V 表示权重矩阵，a、b 表示偏置，$f(x)$ 和 $g(x)$ 分表表示隐藏层和输出层的激活函数。

GRU（Gated Recurrent Unit，门控循环单元）是一种循环神经网络（Recurrent Neural Network, RNN）变体，旨在处理序列数据，通过使用门控机制有效解决了传统 RNN 存在的梯度消失和梯度爆炸问题，尤其适合处理长时间依赖的数据。GRU 的设计目的是在保持计算效率的同时，拥有较高的性能，适用于广泛的序列处理任务。GRU 的结构相对简单，仅包含两个门（更新门和重置门）而不是三个门（输入门、遗忘门和输出门）。这种结构的简化使得 GRU 在保持效果的同时提高了计算效率，GRU 门控结构如图2-13所示。

2.2.3 多头自注意力机制

多头自注意力机制首先在 Transformer 的编码器（Encoder）和解码器（Decoder）中使用 Transformer[55] 的核心包括编码器（Encoder）和解码器（Decoder），核心网络结构如图2-14(a)所示。

编码器中包含多头自注意力（Multi-head Self-Attention）和前馈（Feed Forward）层组成。在多头自注意力中，输入序列可以通过多个自注意力头捕获不同的上下文信息，如图2-14(b)所示。解码器中也是用了多头自注意力，用于处理解码器的输入序列，其计算过程见式2-11、式2-12和式2-13。

$$Q_i = XW_i^Q, \quad K_i = XW_i^K, \quad V_i = XW_i^V \tag{2-11}$$

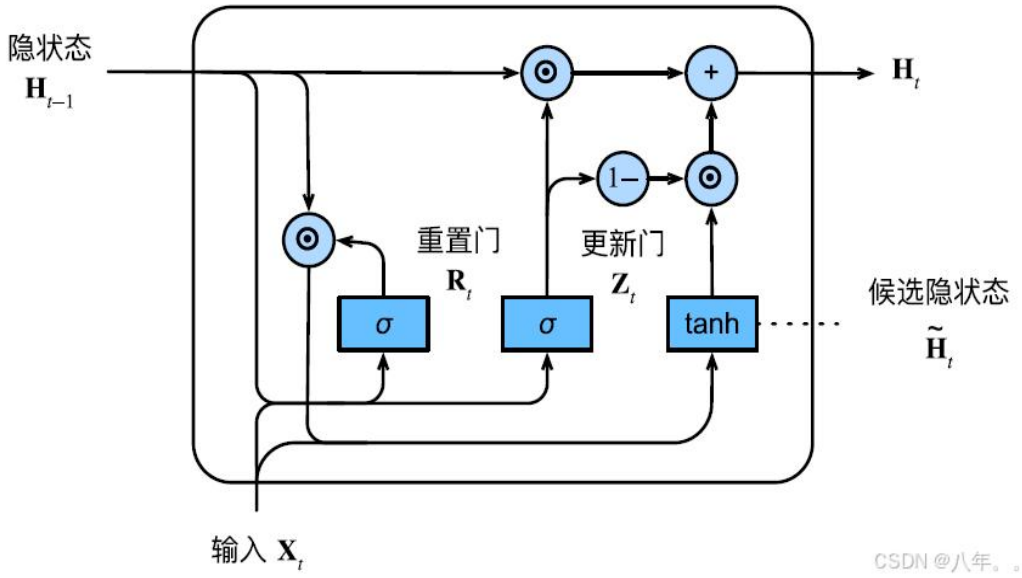


图 2-13 GRU 门控结构图

$$\text{head}_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d}}\right) V_i \tag{2-12}$$

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h) W^O \tag{2-13}$$

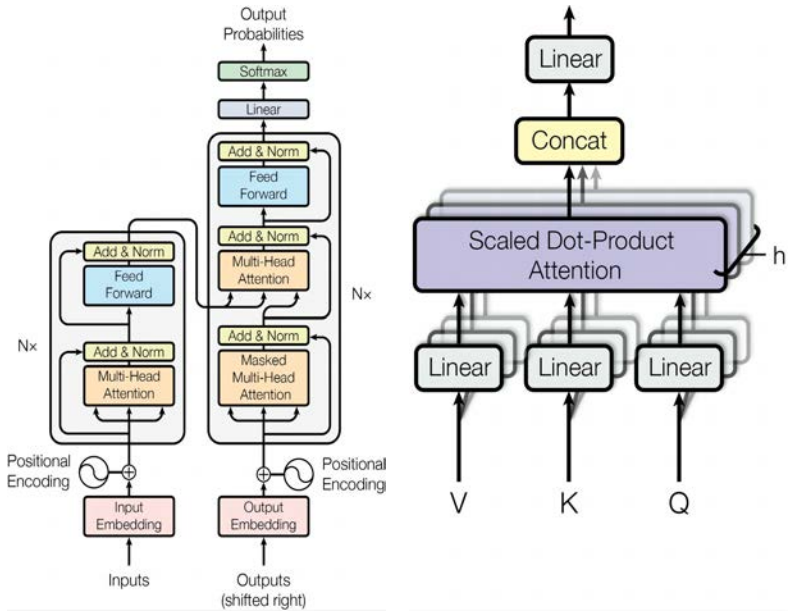
卷积神经网络的感受野有限，即每个神经元只能感知到输入数据中的部分区域，限制了其获取上下文信息的能力。相较之下，Transformer 中所使用的多头自注意力机制允许模型在不同位置的输入序列之间建立直接的连接，从而有效地捕捉长距离依赖关系，这种机制在处理长时间信息的任务时具有优势，已在语音相关领域如语音降噪[56]，语音识别[57]等领域有着大量的应用。

同时，自注意力层可以并行计算，这在一定程度上提高了模型的计算效率，也使得在低功耗嵌入式设备上部署成为可能。基于此，本章提出的 STDCT-DNet 语音降噪后处理模块将引入多头自注意力机制。

2.3 语音降噪相关理论

2.3.1 稳态噪声

稳态噪声指的是在观察时间段内其统计特性（如均值、方差、功率谱密度等）不随时间变化的噪声。换句话说，稳态噪声的频谱和时域特性在时间上是稳定的，不会发生明显的变化。稳态噪声通常在大规模环境中较为常见，比如机械噪声、电气设备噪声等。



(a) (b)

图 2-14 Transformer 与多头自注意力机制图。(a)Transformer 架构；(b) 多头自注意力机制；

稳态噪声的频率分布在时间上保持一致，通常可以使用一个固定的频谱模型来描述，稳态噪声的均值、方差等统计量在时间上不会有明显变化。由于其统计特性稳定，稳态噪声的建模和估计通常较为简单。例如，在语音增强中，我们可以利用噪声的稳态特性通过噪声估计算法来估计噪声功率谱，并进行噪声抑制。比如风扇或空调产生的风声常常在一定时间内保持不变，因此被认为是稳态噪声，电力设备、计算机硬盘等产生的噪声，其频谱在一定时间内保持恒定。由于稳态噪声的统计特性已知，语音增强方法可以通过估计和建模噪声的频谱来进行噪声抑制。例如，常见的 Wiener 滤波和谱减法方法便是通过噪声功率谱估计来滤除噪声。稳态噪声的噪声模型通常是静态的或缓慢变化的，这使得噪声估计过程相对简单。

2.3.2 非稳态噪声

非稳态噪声在时间分布上不连续且具有特定形态特征的噪声。与稳态噪声在时间上保持恒定或变化缓慢不同，非稳态噪声在时间上具有较强的波动性，其统计特性随着时间发生明显变化。因此，非稳态噪声的处理和分析相对复杂，常见的类型包括起伏噪声、间歇噪声和脉冲噪声。

起伏噪声是指在一定的观察时间内，噪声的声级变化较大。具体来说，使用声级计进行慢档动态测量时，噪声的声级变化幅度通常大于 3dB，但一般不会超过 10dB。起伏噪声通常出现在一些动态环境中，其声级在一定范围内不断波动。

这类噪声常见于交通流量变化较大的地方，例如繁忙的街道或道路上，噪声的变化受到车辆、行人以及交通状况等因素的影响。

间歇噪声是一种断续性出现的噪声，呈现出断断续续的特征。间歇噪声通常发生在一些具有周期性或突发性的环境噪声源中，常见于建筑业、维修作业以及某些工业生产过程中。例如，建筑施工中的机械设备噪声，常常会在短暂的时间内产生突出的噪声，然后又快速回到较低的背景噪声水平。

脉冲噪声是一种瞬时的噪声，常见于爆炸声、武器发射声等。与脉冲噪声类似，撞击噪声也是一种瞬时噪声，但其声压上升和下降的时间较脉冲噪声长。撞击噪声常见于工业环境，如锤锻、冲压等作业过程中产生的噪声。

由于非稳态噪声具有较强的时间波动性，其测量方法与稳态噪声有所不同。测量非稳态噪声时，需要特别注意噪声的瞬时变化和动态特性，因此常常需要采用高时间分辨率的仪器进行精确的声级监测。此外，由于非稳态噪声的声级变化较大，测量时需要根据噪声类型调整测量策略，确保能够捕捉到噪声的快速变化和短时特征。

2.4 本章小结

首先，本章介绍了语音分帧、加窗等语音信号预处理技术，并且介绍了在语音处理领域中常用的声学特征。其次，介绍了几种常用的深度学习模型，为后文所建立的基于深度学习的语音降噪模型的研究与设计奠定了基础。最后，对不同特性的噪声进行了简要的概述和分析。

第三章 基于 TCN-GRU 网络的改进 OMLSA-IMCRA 算法

现代传统语音降噪方法虽然在某些场景下表现良好，但在应对非稳态噪声时因为依赖于先验假设而无法对进行精确建模，导致降噪效果下降。传统的对数谱幅度（Log-Spectral Amplitude, LSA）估计算法以及最优改进 OMLSA-IMCRA 算法虽然能在一定程度上抑制噪声干扰，但在实际非稳态噪声场景中仍然面临一些挑战。针对非稳态噪声环境中的问题，本章提出了一种基于 TCN-GRU 网络改进的 OMLSA-IMCRA 算法，该算法基于梅尔频谱特征学习网络，能够更好的估计带噪语音的噪声谱。通过对改进算法与现有的 OMLSA-IMCRA 算法进行仿真实验和比较分析，验证了本文算法在语音降噪方面的有效性。

3.1 OMLSA-IMCRA 算法

时域积分方程时间步进算法的阻抗元素直接影响算法的后时稳定性，因此阻抗元素的计算是算法的关键之一，采用精度高效的方法计算时域阻抗元素是时域积分方程时间步进算法研究的重点之一。

3.1.1 对数谱幅度估计算法

在语音信号处理中，通常会将语音信号从时域转换到频域，以便更好地分析其频率成分和能量分布。频谱表示了信号在不同频率上的能量分布情况，是频域分析的重要结果。为了获得频谱，通常使用短时傅里叶变换（STFT）或其他类似的变换方法。通过这些方法，信号被分解成不同频率的分量，便于对其进行处理和分析。

Boll 在 1979 年提出了傅里叶变换域中的谱减法[7]，该方法在假定加性噪声信号和语音信号不相关, 估计噪声的功率谱, 接着使用带噪语音功率谱中减去估计出的噪声功率谱, 再与原带噪语音的相位结合成去噪后的语音信号[7]，其计算方法如式3-1所示。

$$|\hat{X}(k,l)| = \begin{cases} |Y(k,l)|^p - |\hat{D}(k,l)|^p, & \text{if } |Y(k,l)|^p > |\hat{D}(k,l)|^p \\ 0, & \text{Other} \end{cases} \tag{3-1}$$

式中 $\hat{X}(k,l)$ 表示增强的语音谱，$\hat{D}(k,l)$ 表示噪声谱的估计。

但是经典的谱减法并不适用于变化较为剧烈的噪声，因此学者们提出了基于

统计模型的谱估计方法用来实现预先语音增强。其中，基于贝叶斯估计的方法，因为利用了先验知识，所以一般来说其性能要比最大似然估计的方法更好。在众多贝叶斯估计方法中，典型的代表就是基于最小均方误差的幅度估计方法[13]，其优化目标是最小化增强语音幅度与纯净语音幅度之间的贝叶斯均方误差，进而计算计算最优幅度谱增益系数，如式3-2所示。

$$BMSE = E\left\{\left(|\hat{X}(k,l)| - |X(k,l)|\right)^2\right\} \tag{3-2}$$

尽管 MMSE 幅度估计算法在数学上易于推导，但就人耳听觉感知而言越来越多的研究表明非线性的对数谱均方误差算法对语音处理更为合适[58]。研究者从一些语音信号处理的经验出发，将线性的幅度谱上的误差平方改为了对数幅度谱上的误差平方，如式3-3所示。

$$|\hat{X}(k,l)| = \begin{cases} |Y(k,l)|^p - |\hat{D}(k,l)|^p, & \text{if } |Y(k,l)|^p > |\hat{D}(k,l)|^p \\ 0, & \text{Other} \end{cases} \tag{3-3}$$

经过之前研究者的主观听音测试可以发现，对数谱的估计器与幅度谱的估计器相比，残留噪声更少，并且语音失真也更小。

3.1.2 OMLSA-IMCRA 降噪算法

Cohen 提出的最优改进对数幅度谱估计 (OM-LSA) 算法[16]，能够适应较强噪声环境，保护较弱的语音单元。此后该算法因为针对稳态噪声卓越的降噪效果逐渐作为工业界的主流语音降噪方案的中的一个重要步骤，并且因为较小的计算量常作为低功耗嵌入式设备上的降噪方案。

OMLSA 算法的计算准则基于对数谱幅度估计算法，所以实际上也是最小化实际干净语音和估计出来的干净语音的差异，这一差异如果使用公式可以表示为:

$$Delta = E\left(\left|\log A(k,l) - \log \hat{A}(k,l)\right|^2\right) \tag{3-4}$$

其中 $A(k,l)$ 是不带噪的干净语音信号 x 的频谱幅值，$\hat{A}(k,l)$ 是估计出来的频谱幅值。如果使用传统的幅度谱估计方法，那么可由式3-5表示。

$$\hat{A}(k,l) = G(k,l)|Y(k,l)| \tag{3-5}$$

其中 $G(k,l)$ 为带噪语音的估计增益，$Y(k,l)$ 是含噪语音的幅度谱。在经过 Cohen 的改进后，估计频谱幅值变成了一种更加复杂的形式，可由式3-6表示。

$$\hat{A}(k,l) = \left(G_{\min}|Y(k,l)|^{(1-p(k,l))}\right) \times \left(G_{H1}(k,l)|Y(k,l)|^{p(k,l)}\right) \tag{3-6}$$

其中 $G_{\min}$ 为语音不存在时的增益函数，一般将 $G_{\min}$ 的值设为带噪信号的最小信噪比。$G_{H1}(k,l)$ 是语音存在时的增益函数，其计算方法可由式3-7表示。

$$G_{H1}(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-x}}{x}\,dx\right), \text{with } v(k,l) = \frac{\xi(k,l)\gamma(k,l)}{1+\xi(k,l)} \tag{3-7}$$

其中 $\xi(k,l)$ 是一帧语音的先验信噪比，其计算方法可由式3-8表示。

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l)) \exp\left(-\frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)}\right)\right\}^{-1} \tag{3-8}$$

公式3-6中 $p(k,l)$ 的是语音存在的后验概率。计算语音存在后验概率的公式可由式3-9表示。

$$\xi(k,l) = \alpha G_{H1}^2(k,l-1)\gamma(k,l) + (1-\alpha)\max\{\gamma(k,l)-1, 0\} \tag{3-9}$$

可以看到，公式3-8和公式3-9的计算都需要首先计算 $\gamma(k,l)$，表示的是一帧语音的后验信噪比。所以要使用改进的 OM-LSA 算法，需要首先计算语音先验信噪比和语音后验信噪比，而当前帧的语音先验信噪比计算依赖于上一帧的语音后验信噪比，所以计算语言后延信噪比非常重要。语音后验信噪比的计算方法可由式3-10表示。

$$\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_d(k,l)} \tag{3-10}$$

其中 $\lambda_d(k,l)$ 表示噪声谱估计，所以使用 OM-LSA 算法进行语音降噪还需要估计噪声谱。

噪声谱估计有多种方法，OM-LSA 算法的提出者研究出了一种噪声谱估计方法 IMCRA。该算法首先需要计算带噪语音功率谱的第一次平滑估计，其计算方法可由式3-11表示。

$$S(k,l) = \alpha_s S(\lambda - 1, k) + (1 - \alpha_\epsilon) S_f(\lambda, k) \tag{3-11}$$

其中 α_s 为平滑参数。并通过式3-12来跟踪平滑功率谱并更新平滑功率谱的最小值，其中 $S(k,l-1)$ 表示上一帧含噪语音的功率。

$$S_{\min}(k,l) = \min\{S_{min}(k,l-1), S(k,l)\} \tag{3-12}$$

接下来计算指示函数 $I(\lambda,k)$ 来进行语音活动检测，可用式3-13表示。

$$I(\lambda,k) = \begin{cases} 1, & \text{if } \gamma_{\min}(k,l) < \gamma_0 \text{ and } \zeta(k,l) < \zeta_0 \text{ (speech is absent)} \\ 0, & \text{otherwise (speech is present)} \end{cases} \tag{3-13}$$

其中 γ_0 和 ζ_0 为阈值参数，可自己选择。其计算公式如式3-14所示。

$$\gamma_{\min}(k,l) = \frac{|Y(k,l)|^2}{B_{\min} S_{\min}(k,l)} \quad , \quad \zeta(k,l) = \frac{S(k,l)}{B_{\min} S_{\min}(k,l)} \tag{3-14}$$

其中 $B_{\min}$ 是噪声谱最小值估计偏置补偿因子，在很多文献中取值为 1.66。接下来需要在第一次平滑的基础上使用语音活动检测的结果进行二次平滑，其计算方法可由式3-15表示。

$$\tilde{S}_f(\lambda,k) = \begin{cases} \dfrac{\sum_{i=-\omega}^{\omega} b(i) I(k-i,l) |Y(k-i,l)|^2}{\sum_{i=-\omega}^{\omega} b(i) I(k-i,l)}, & \text{if } \sum_{i=-\omega}^{\omega} I(k-i,l) \neq 0 \\ \tilde{S}(k,l-1), & \text{otherwise} \end{cases} \tag{3-15}$$

其中 $\tilde{S}(k,l-1)$ 为第一次平滑的结果。在计算完第二次平滑的结果后，需要通过式3-16来跟踪二次平滑功率谱并更新二次平滑功率谱的最小值。

$$\tilde{S}_{\min}(k,l)=\min\left\{S_{\min}(k,l-1),\tilde{S}(k,l)\right\} \tag{3-16}$$

并据此估计语音不存在的概率 $\hat{q}(k,l)$，其计算可用式3-17来表示。

$$\hat{q}(k,l)=\begin{cases}1, & \text{if } \tilde{\gamma}_{\min}(\lambda,k)\le 1 \text{ and } \tilde{\zeta}(\lambda,k)<\zeta_0\\ \frac{\gamma_1-\tilde{\gamma}_{\min}(\lambda,k)}{\tilde{\gamma}_1-1}, & \text{if } 1<\tilde{\gamma}_{\min}(\lambda,k)\le 1 \text{ and } \tilde{\zeta}(\lambda,k)<\zeta_0\\ 0, & \text{otherwise}\end{cases} \tag{3-17}$$

其中 γ_1 为阈值参数，可自己选择。$\tilde{\gamma}_{\min}$ 和 $\tilde{\zeta}(\lambda,k)$ 可由式3-18计算得出。

$$\tilde{\gamma}_{\min}(k,l)=\frac{|Y(k,l)|^2}{B_{\min}\tilde{S}_{\min}(k,l)} \quad , \quad \zeta(k,l)=\frac{\tilde{S}(k,l)}{B_{\min}\tilde{S}_{\min}(k,l)} \tag{3-18}$$

据此可以根据公式3-8计算出语音存在的后验概率，然后根据语音存在的后验概率进行噪声谱估计，计算过程可由式3-19表示。

$$\tilde{\lambda}_d(k,l+1)=\tilde{\alpha}_d(k,l)\tilde{\lambda}(k,l)+[1-\tilde{\alpha}_d(k,l)]|Y(k,l)|^2 \tag{3-19}$$

其中 $\tilde{\lambda}_d(k,l+1)$ 为时变平滑因子，考虑到了语音存在概率的影响，其计算方法如式3-20所示。

$$\tilde{\alpha}_d(k,l)=\alpha_d+(1-\alpha_d)p(k,l) \tag{3-20}$$

其中 α_d 可取决于具体的应用场景进行选择，有的文献建议将 α_d 设置为 0.7-0.9 之间。为了避免噪声谱估计的过低，需要采用一个偏置因子对噪声谱估计进行补偿，至此得到最终的噪声谱估计，其计算方法如式3-21所示。

$$\hat{\lambda}_d(k,l+1)=\beta\tilde{\lambda}_d(k,l+1) \tag{3-21}$$

3.1.3 OMLSA-IMCRA 降噪算法性能仿真

为进一步分析 OMLSA-IMCRA 降噪算法的性能，在此建立仿真实验进行比较。在仿真实验中，假设采用单通道麦克风进行语音信号采样，麦克风的采样频

率为 16KHz，假设经模数转换模块得到 16K 语音数据。

设置三组噪声条件，分别检验该降噪算法在 15dB 稳态噪声、15dB 非稳态噪声、0dB 非稳态噪声环境下的降噪效果。

实验一：分析 OMLSA-IMCRA 算法在 15dB 稳态噪声环境下的降噪效果。

纯净语音为一段语音信号测试女声，该语音信号和语谱图如图3-1所示。

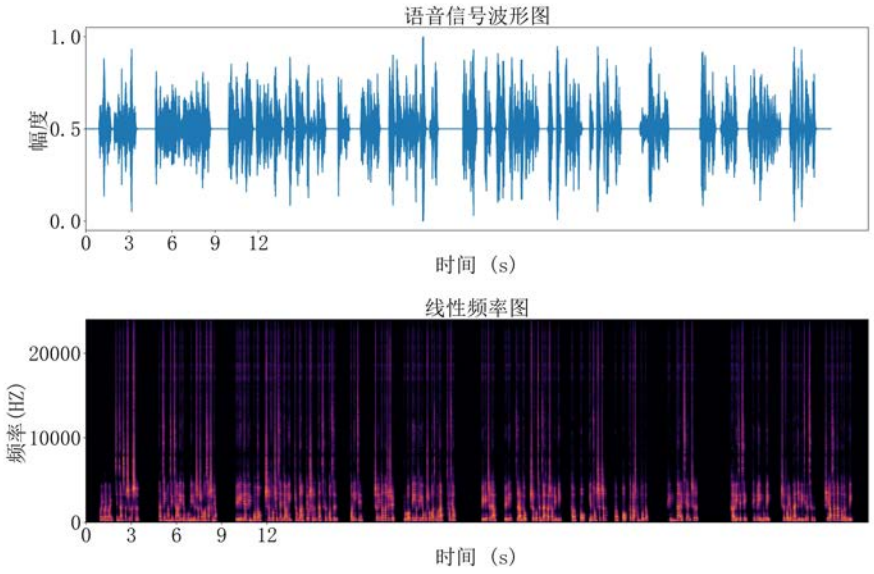


图 3-1 纯净语音 (女声) 波形-语谱图

设置环境噪声为高斯白噪声，信噪比为 15dB，采用加性噪声的方式融合纯净语音引号和噪声信号，带噪信号和其语谱图如图3-2所示。

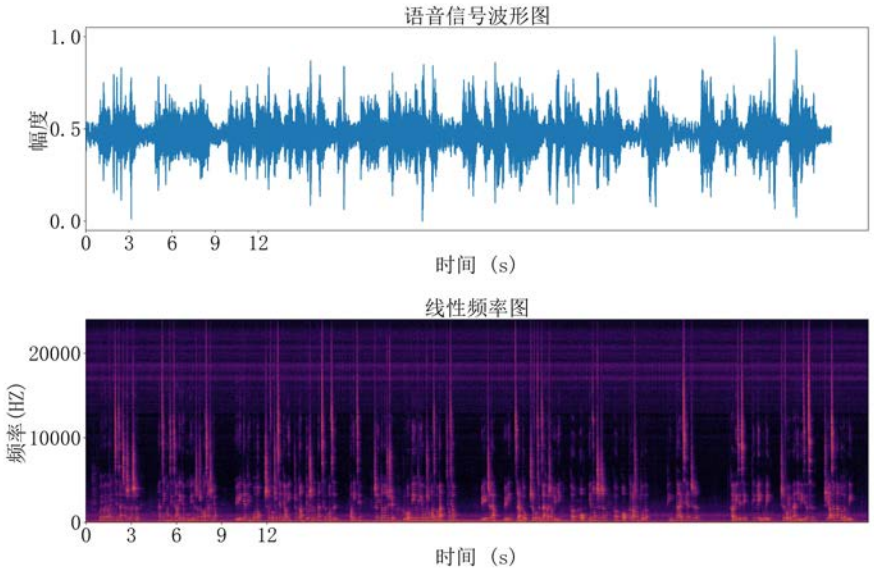


图 3-2 混合稳态噪声语音波形-语谱图

使用 OMLSA-LSA 算法对该带噪信号进行降噪，仿真实验得到的降噪后语音信号及其语谱图如3-3所示。

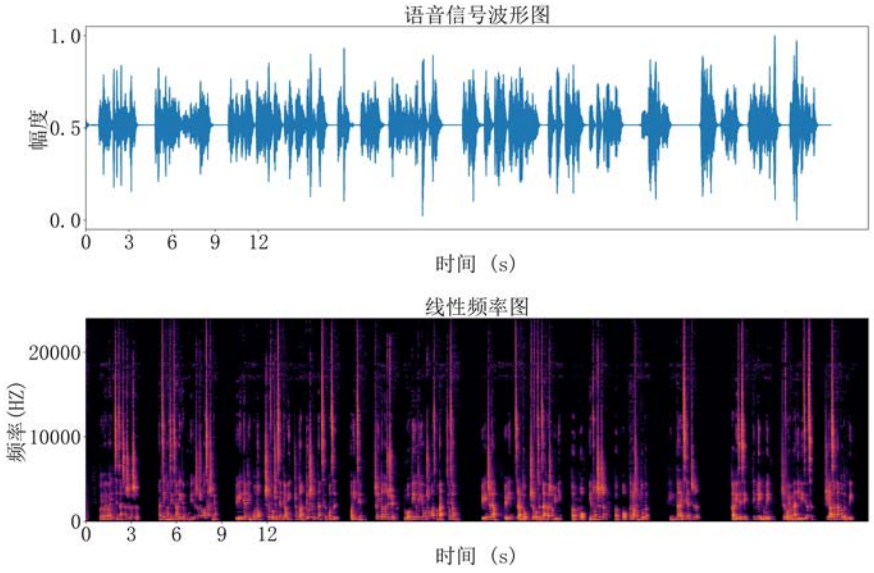


图 3-3 OMLSA-IMCRA 降噪算法应用后语音波形-语谱图

从降噪后的语音信号的语谱图中可以看出，该降噪算法对信噪比为 15dB 的混合白噪声信号的降噪效果较好，可以很好的还原原始语音信号。

实验二：分析 OMLSA-IMCRA 算法在 15dB 非稳态噪声环境下的降噪效果

纯净语音为一段语音信号测试男声，该语音信号和语谱图如图3-4所示。

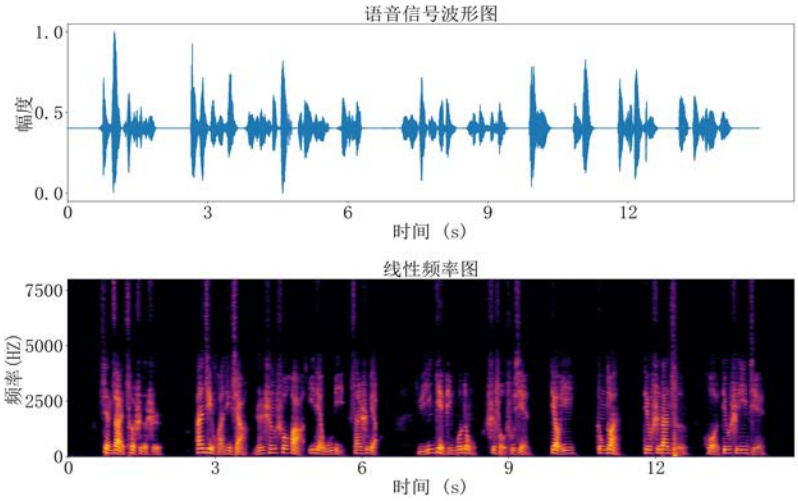


图 3-4 纯净语音 (男声) 波形-语谱图

设置环境噪声为 Noise-92 噪声集中的 factory1 噪声信号，该噪声信号录制于工厂，包含稳态噪声信号如鼓风机风声，也包含非稳态信号如器械撞击声。按照

信噪比为 15dB 调制带噪语音信号。带噪信号和其语谱图如图3-5所示。

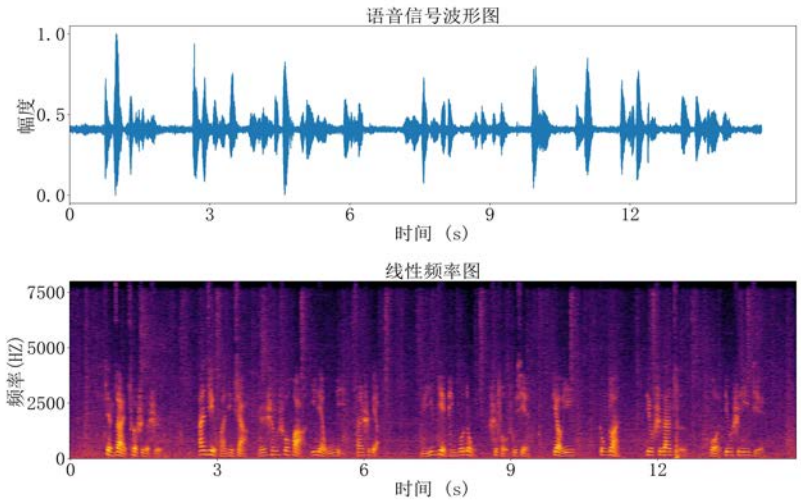


图 3-5 信噪比 15db 混合工厂噪声波形-语谱图

使用 OMLSA-LSA 算法对该带噪信号进行降噪，仿真实验得到的降噪后语音信号及其语谱图如图3-6所示。

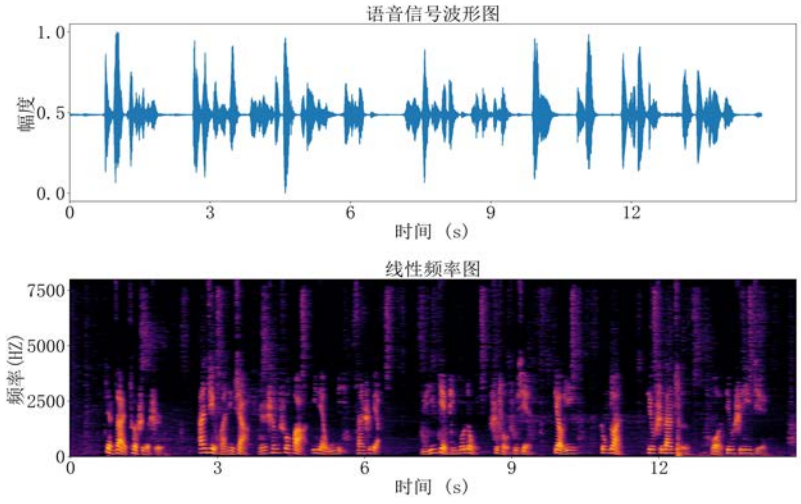


图 3-6 对 15db 信噪比混合噪声应用 OMLSA-IMCRA 算法后的语音波形-语谱图

从降噪后的语音信号的语谱图中可以看出，该降噪算法对信噪比为 15dB 的混合白噪声信号的降噪效果一般，可以消除 factory1 噪声信号中的部分稳态噪声，但是无法对非稳态噪声进行有效的消除，无法有效的重建纯净语音信号。

实验三：分析 OMLSA-IMCRA 算法在 0dB 非稳态噪声环境下的降噪效果

纯净信号为实验二所用信号，设置环境噪声为 Noise-92 噪声集中的 factory1

噪声信号，按照信噪比为 0dB 调制带噪语音信号，带噪信号和其语谱图如图3-7所示。

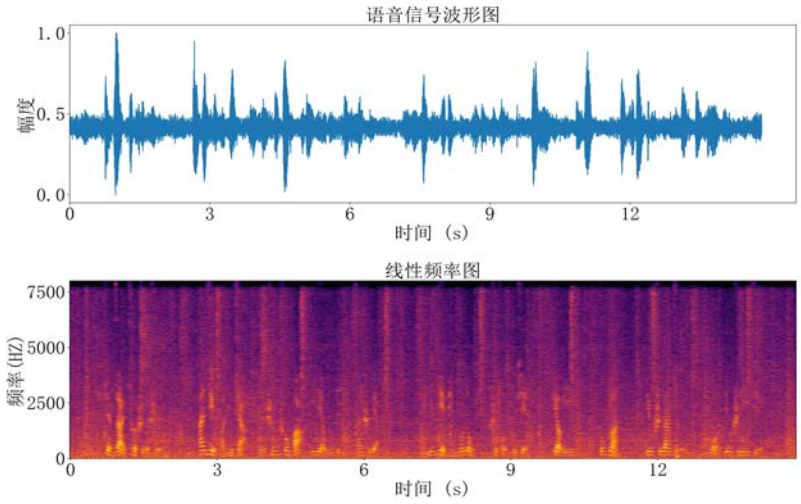


图 3-7 信噪比 0db 混合工厂噪声波形-语谱图

使用 OMLSA-LSA 算法对该带噪信号进行降噪，仿真实验得到的降噪后语音信号及其语谱图如图3-8所示。

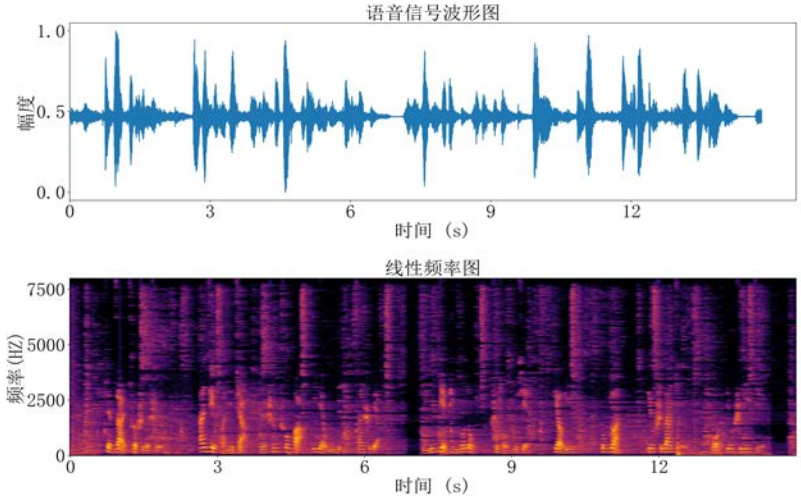


图 3-8 信噪比 0db 混合工厂噪声波形-语谱图

从降噪后的语音信号的语谱图中可以看出，该降噪算法对信噪比为 0dB 的混合白噪声信号的降噪效果非常差，同样可以消除 factory1 噪声信号中的部分稳态噪声，此部分噪声表现在带噪语音的高频部分。但是完全无法对非稳态噪声进行消除，对 OMLSA-IMCRA 算法的客观语音质量的定量分析将在之后的实验章节中展示。

3.2 基于 TCN-GRU 网络的改进 OMLSA-IMCRA 算法

根据上述仿真实验结果可得，OMLSA-IMCRA 算法在处理非稳态噪声时效果不佳，低信噪比环境是效果更差。这是由于之前的研究工作大多集中在改进估计器的增益函数，旨在提升模型的整体性能。然而，最近的一些研究结果表明，统计模型方法的效果受先验信噪比和后验信噪比估计精度的影响最大[59]。尤其是后验信噪比的估计，它依赖于噪声谱的准确估计，因此，如何提升后验信噪比估计的准确性就成为了一个关键问题。因此，从统计模型方法的优化角度来看，解决的核心问题已经转变为如何提高先验信噪比和噪声谱估计的精度。这一问题的解决将直接影响模型的上限，并推动噪声抑制技术的进一步发展。

为解决这一问题，本文提出了基于 TCN-GRU 网络的改进 OMLSA-IMCRA 算法，旨在通过 TCN-GRU 模块解决原算法噪声谱估计精度低的问题，替代噪声估计算法 IMCRA 来跟踪噪声谱，提高带噪信号的噪声谱估计精度。

3.2.1 算法整体框架

图3-9展示了本章节所提出的语音降噪模型（OMLSA-TCN-GRU）的整体框架。其中主要包括特征变换模块、幅度谱增益计算模块、噪声谱估计深度学习骨干网络和特征反变换模块。输入的带噪语音信号波形经过一定的预处理后，根据不同任务进行不同的高级特征变换，其中，带噪语音信号的每个声道都能提取出对应的频谱特征、梅尔频谱特征。OMLSA 估计器使用短时傅里叶变换频谱特征来估计增益，梅尔频谱特征则作为深度学习网络的输入从而估计噪声的梅尔频谱。

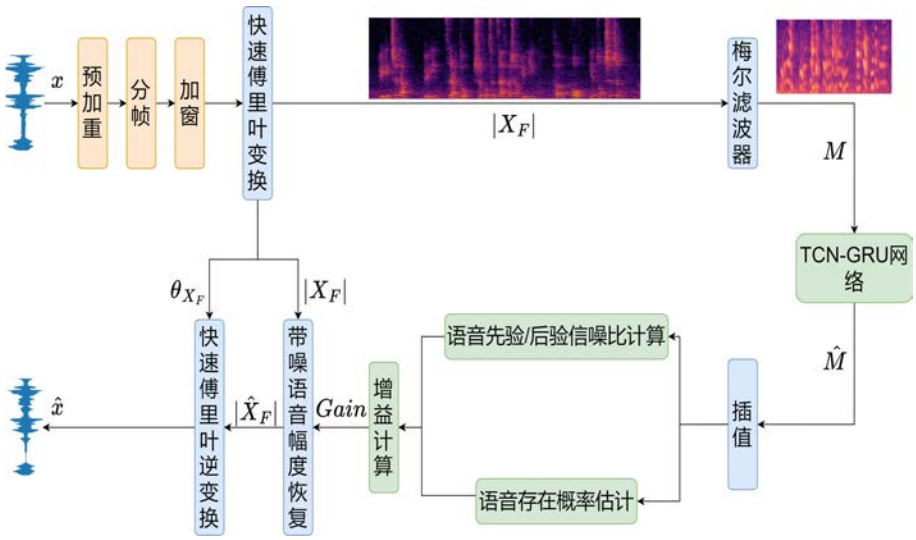


图 3-9 OMLSA-TCN-GRU 降噪流程图

梅尔频谱特征作为噪声谱估计网络的输入，经网络估计后得到噪声的梅尔频

谱特征，再通过插值得到噪声幅度谱信息，最终幅度谱信息交由 OMLSA 估计器估计增益，然后将增益与带噪语音幅度谱相乘得到降噪后的语音幅度谱，最终经傅里叶逆变换得到时域语音信号。

3.2.2 梅尔频谱特征

梅尔频谱（Mel Spectrogram）是语音处理、语音识别和音频分析中常用的一种特征表示方法。梅尔频率尺度是基于人类听觉的特性设计的。人耳对低频的变化更为敏感，而对高频的变化不那么敏感。梅尔频率通过对线性频率轴进行压缩，使得较低频率部分更精细，而较高频率部分则被压缩。梅尔频率尺度的这种设计使得梅尔频谱比传统的线性频谱更能反映出人类听觉系统的特点，从而能够更好地捕捉语音和音乐中的有用信息。对梅尔频谱的分析已在 2.1.3 小节完成，在此不做过多分析。

与此同时使用梅尔频谱可以帮助降维和去除一些高频噪声，通过对频谱进行梅尔尺度的转换，可以减少高频部分的干扰，并使得低频信息更加突出，适用于语音增强和噪声抑制任务。

在深度学习模型中广泛应用，尤其是在语音识别和语音降噪任务中，梅尔频谱能够提取出有意义的特征，同时可以减少数据的冗余，所以适合用于深度学习语音降噪任务，所以本文提出的模型将基于带噪信号和噪声信号的梅尔频谱特征完成降噪任务。

3.2.3 TCN-GRU 模块

针对传统单结构神经网络处理复杂预测任务的不足，本文提出了一种新的 TCN-GRU 模块来学习带噪语音噪声谱的梅尔频谱特性，该模型有效地集成了 GRU 网络和 TCN 网络, 首先通过通过 TCN 模块来捕捉语音特征信息中的长期依赖性，再通过 GRU 模块帮助模型理解语音局部特征信息的动态变化，这能够帮助捕捉持续时间较短的非稳态噪声。然后通过输出层将特征信息映射到输出空间从而预测出这一段带噪语音的噪声梅尔谱特征。基于 TCN-GRU 网络模块的框架图如图3-10所示。

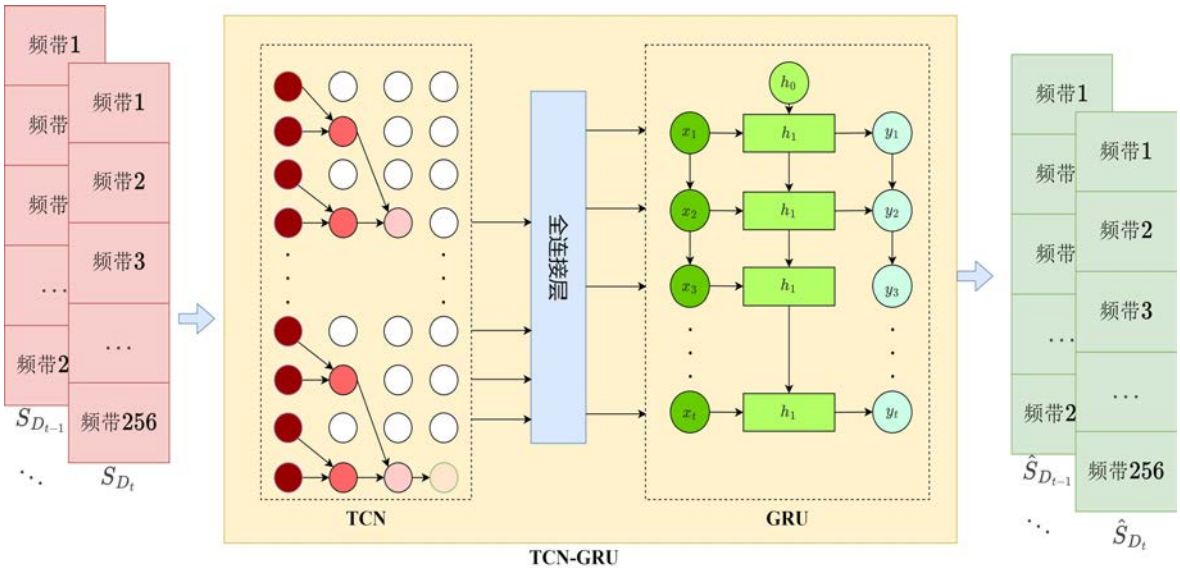


图 3-10 TCN-GRU 网络模块的框架图

首先，每一帧带噪语音的梅尔频谱特征将被发送到 TCN 网络中。时序卷积网络（Temporal Convolutional Network，TCN）是在卷积神经网络（Convolutional neural Network, cnn）的基础上发展起来的一种有效处理时间序列数据的神经网络。其主要功能是处理不同时间尺度上的数据，具有变长输入和输出。在复杂的语音相关任务中，TCN 帮助模型捕捉语音特征数据中的模式和规则，提高模型的预测性能。同时 TCN 在并行计算方面有着显著的优势，可以很好的利用硬件的并行计算特性，适合部署在低功耗嵌入式设备上。本文建立的 TCN-GRU 噪声谱估计模型中的 TCN 具体网络结构如图3-11所示。

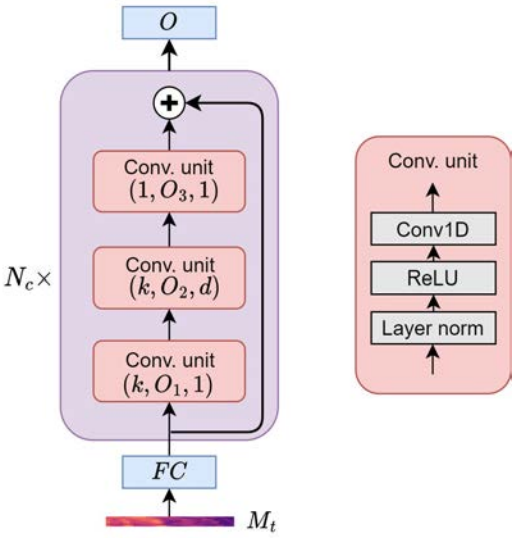


图 3-11 TCN 具体网络结构图

建立的 TCN 网络模块包括一个全连接输入层 FC，一个全连接输出层 O。和 N_C 个级联的时序卷积残差块。其中 k,O,d 分别表示时序卷积层的卷积核大小，输出大小和膨胀系数。第一和第三卷积单元具有为 1 的膨胀系数，作为保障输入输出大小的 1x1 的卷积层。而第二卷积单元采用为 d 的膨胀系数，从而在先前的时间步长上提供更大的感受野，实现了空洞卷积。

根据 TCN 网络的提出者的建议，应最少采用三层膨胀卷积，且膨胀系数按照 2 的幂次方递增，且考虑到 TCN 接受的是带噪语音梅尔谱特征，已经有一定的信息压缩作用，所以本章为第二层卷积单元设置的膨胀系数最大为 4。那么理论上在卷积核大小为 3 的情况下，膨胀系数达到最大时，残差块最多可以感受前 7 帧的信息。对于第 b 个残差块，决定采用循环设置膨胀系数的方式，第 n 个残差块使用的膨胀系数为 $2^{B\%3}$，图3-12所示的例子 (N=7) 可以清楚的表示膨胀系数与残差块序号的关系。

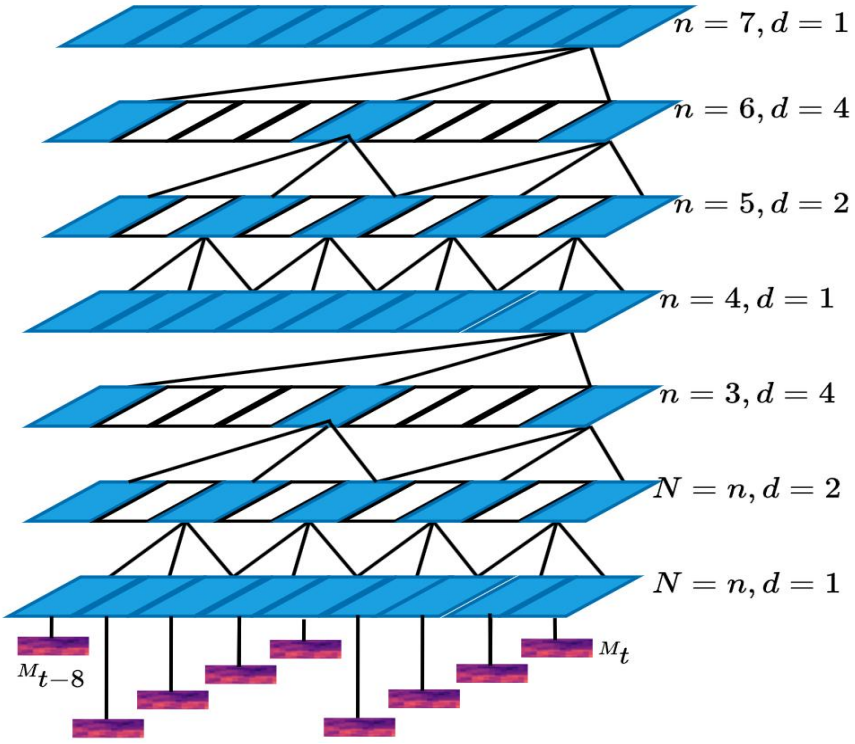


图 3-12 TCN 级联示意图

级联 N_C 个残差块的原因是随时序卷积块级联个数的增加，整个网络可以利用追溯的时间段就越大。作为训练时间、模型大小和性能之间的折衷，本文决定使用 $N_C = 24$(共 24 个时序卷积残差块)，$k = 3$，$O_2 = 64$，$O_1 = 64$,$O_3 = 64$ 作为本文的 TCN 网络参数。

随后 TCN 网络的输出会送到 GRU 网络中。虽然 GRU 网络无法进行并行加

速，但是 GRU 网络具有很强的处理时间序列数据的能力，它可以帮助预测模型学习时间序列数据中的短期依赖关系，防止处理长序列时出现梯度消失或爆炸问题。GRU 网络的方程如下所示。

$$z_t = \sigma\left(\boldsymbol{w}_{xz}^T \boldsymbol{x}_t + \boldsymbol{w}_{hz}^T \boldsymbol{h}_{t-1} + b_z\right) \tag{3-22}$$

$$r_t = \sigma\left(\boldsymbol{w}_{xr}^T \boldsymbol{x}_t + \boldsymbol{w}_{hr}^T \boldsymbol{h}_{t-1} + \boldsymbol{b}_r\right) \tag{3-23}$$

$$g_t = \tanh\left(\boldsymbol{w}_{xg}^T \boldsymbol{x}_t + \boldsymbol{w}_{hg}^T (\boldsymbol{r}_t \otimes \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h\right) \tag{3-24}$$

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes g_t \tag{3-25}$$

式中，x_t 为 t 时刻的输入数据，h_t 为 t 时刻的输出或状态，w 为 GRU 的权值，σ 为 sigmoid 激活函数，$\otimes$ 为 Hardmard 积，*tanh* 为 *tanh* 激活函数。作为训练时间和性能之间的折衷，本文建立的 TCN-GRU 噪声谱估计模型中的 GRU 具体网络结构为三层 GRU 网络模块，每一层有 64 个门控循环单元，输入输出都为 64。之后，GRU 网络模块输出估计的噪声梅尔频谱。

3.2.4 损失函数

回归问题的目标是预测一个连续的数值，而均方误差在度量连续输出与真实值之间的误差时非常直观。MSE 通过对预测误差的平方求平均，能够有效衡量预测值与真实值之间的差距。均方误差是一个平滑且可微的函数，这对于深度学习中的梯度下降优化算法非常重要。通过反向传播算法，我们可以计算出每一层的梯度并进行参数更新，在本章节提出的模型中使用均方误差 MSE 作为损失函数，均方误差的计算公式如式3-26所示。

$$\mathcal{L}_{\text{TCN-GRU}} = \left\| \hat{M} - M \right\|_F^2 \tag{3-26}$$

其中 M 是噪声和干净语音融合之前所添加的噪声的梅尔频谱，$\hat{M}$ 是模型预测的噪声梅尔频谱。

3.3 实验验证分析

3.3.1 实验环境

实验中所使用的软硬件设备参数见表3-1。

表 3-1 实验平台软硬件环境配置

开发环境	参数
GPU	NVIDIA GeForce RTX 3060Ti 8G
CPU	Intel Core i5-12400
操作系统	Ubuntu-20.04 LTS
RAM	32G
编程语言	Python-3.8
深度学习框架	PyTorch-2.0.0

3.3.2 数据集

(1) 纯净语音开源数据集

希尔贝壳中文普通话语音数据库 AISHELL-2 是全球最大的开源中文语音数据集[60]，语音时长为 1000 小时，其中 718 小时来自 AISHELL-ASR0009-[ZH-CN]，282 小时来自 AISHELL-ASR0010-[ZH-CN]。录音文本涉及唤醒词、语音控制词、智能家居、无人驾驶、工业生产等 12 个领域。录制过程在安静室内环境中，采样频率为 16KHZ。AVSpeech[61] 是由由谷歌发布的干净语音数据集，每条数据长度在 3 10s，共计 4700h，包括不同人种、不同语言的 150,000 个说话人。每条数据只出现一个说话人的声音。

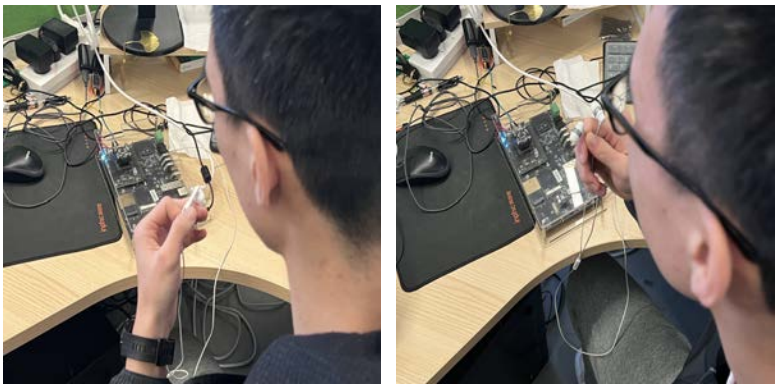
(2) 噪声集

NOISE-92 噪声库是一个专为语音信号处理领域设计的噪声数据集[62]，旨在为研究人员和开发者提供丰富的环境噪声样本。该数据集包含了 92 种不同类型的环境噪声该噪声集的采样频率为 19.98KHZ, 需下采样到 16KHZ 使用。WHAM 噪声数据集全称为 WSJ0 Hipster Ambient Mixtures[63]，这些噪声录制于旧金山湾区的咖啡馆、餐厅、酒吧、办公楼、公园等城市环境中，弥补了 NOISE-92 噪声部分噪声类型缺失的问题。此噪声集为双通道数据集，分离为单通道使用。

(3) 纯净语音自制数据集

上述开源纯净语音数据集采样设备较为单一，而噪声数据类型不够全面。为增加模型的鲁棒性，真实模拟低功耗设备环境，本次实验自制部分数据集，在多

种环境下使用海思平台 SS928 和多种双声道耳机，让多个志愿者完成部分人声纯净语音以及多种类型噪音的录制，噪声类型主要录制室内非稳态噪声如敲击键盘噪音、写字噪音、会议桌上物品移动噪音等。录制设备和环境如图3-13所示，其中 SS928 平台核心板内置 Audio 编解码芯片，可以支持 16bit 语音输入和输出。采样率为双通道 (左右声道)16KHZ，16 位深，最终通过声道分离提取为单通道噪声信号，录制的每一条噪声时长在 15s 左右。



(a) (b)

图 3-13 自制数据集录制环境图

(4)VoiceBank+Demand 数据集 [64]

该数据集专门用于语音降噪任务，尤其是在有噪声环境下的语音恢复。被广泛应用于语音增强研究，是语音增强研究中的标准数据集之一，提供了真实环境噪声下的带噪语音和干净语音配对，许多深度学习语音增强模型都使用了该数据集作为基准测试数据。

(5) 实验数据集

选用 AISHELL-2、AVSpeech 和自制数据集共 20000 条纯净语音作为训练集，2000 条纯净语音作为测试集，采样频率均为 16kHZ。具体配置如表3-2所示。

表 3-2 数据集配置

数据集来源	训练集 (条)	测试集 (条)
AISHELL-2	14000	1400
AVSpeech	5000	500
自制数据集	1000	100

使用 NOISE-92 噪声库的 Babble、White、Destroyerops、Destroyerengine、machinegun、Factory1 噪声、自制噪声以及 WHAM 噪声库作为训练集混合噪声，使用 NOISE-92 噪声库的 F16、Factory1、Pink 和 Factory2 作测试集混合噪声。将上述纯净语音数据与噪音数据随机落在 0dB、5dB、10dB、15dB 信噪比下进行混合，时间长度截取为 6s-12s，从而得到本次实验所用的带噪语音数据集 DeNoiseBank 数据集，后续使用该数据集的测试集用来在 0dB、5dB、10dB 和 15dB 的信噪比下进行测试。

VoiceBank+Demand 测试集与训练集没有重叠，使用 VoiceBank+Demand 测试集将本文算法与其他先进的语音增强算法进行比较。

3.3.3 实验设置

OMLSA-TCN-GRU 降噪模型在训练阶段的超参数设置如表3-3所示。

表 3-3 超参数配置

超参数类型	参数
学习率	0.0005
优化器	Adam
Batch_size	64
Epochs	200
Dropout	0.2
最大频率	16000
帧长	32ms
帧移	16ms
窗口类型	汉明窗
FFT 点数	512
梅尔滤波器组数	64

3.3.4 语音质量评价指标

语音作为交流工具，对其质量的评估就显尤为重要。语音质量有三种评价指标，即清晰度、可懂度以及自然度。语音清晰度指的是对说话人语音理解度的一个衡量。而语音可懂度为一定条件下的语音理解能力，自然度注重的是语音的保真性。目前，对于自然度的评价一直没有统一的评判标准。而对于语音可懂度和清晰度的评判标准有多种方法。语音的评价指标又分为主观评价和客观评价两方面。

主观评价方法在评价过程中受到评价者主观感受的影响，不同的评价者可能会对同一段语音信号给出不同的评分结果，而且主观评价方法成本较高且比较费时。所以研究人员也提出了很多客观的评价方法。

客观评价方法是指基于信号处理和数学模型分析的方法，通过对增强后的语音与目标语音作对比，然后进行评价。客观评价方法不需要人员参与，而是选取一些参数作为统一标准。下面对几种常用的客观评价指标进行简单介绍。

（1）PESQ [65]

PESQ 方法是一种客观语音质量评价方法，利用数学模型计算原始信号和增强信号之间的差异。该模型建立了语音和人类的心理感知的对应关系，可提供一 MOS 预测值，同时将其映射到 MOS 刻度范围。PESQ 的分数范围在-0.5 4.5，分数越高表示语音质量越高。PESQ 要求两个输入的信号在时间域是相同的，利用线性滤波、增益变化补偿和均衡听觉变换得到对称干扰和非对称干扰两个失真参数，通过累加时间和频率上的参数来计算语音质量, 计算方法如式3-27所示，其中 $a0 = 4.5$，$a1 = 0.1$，$a2 = 0.309$，通常使用使用 ITU-T P.862 标准提供的工具进行计算。

$$PESQ = a_0 - a_1 d_{SYM} - a_2 d_{ASYM} \tag{3-27}$$

（2）STOI [66]

短时客观可懂度取值范围在 0 1 中，数值为 1 表示语音能够被充分理解，对于一段语音中的一个字或单词，可懂度是二值的，只有能或不能被听懂。其计算方法可由式3-28表示，其中 m 为帧号，k 为一帧中的频率点，T 与 F 分别表示频带与帧的数量，r 为局部相关系数。

$$STOI = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} r_{m,k}(t) \tag{3-28}$$

(3)COVL [67]

COVL 提供了对整体听觉质量的综合评价，得分在 1 到 5 之间。更高的 COVL 分数表示增强后的语音具有更好的整体听觉质量。COVL 是一个全面的评估标准，考虑了语音的清晰度、纯净度和可理解性。COVL 计算公式如式3-29所示，式中 PESQ,LLR,WSS 分别表示语音质量感知评估 (Perceptual Evaluation of Speech Quality，PESQ)、对数似然比测度 (Log-Likelihood Ratio，LLR)、加权谱倾斜测度 (Weighted Spectral Slope，WSS)。

$$COVL = 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - 0.007 \cdot WSS \tag{3-29}$$

3.3.5 OMLSA-TCN-GRU 降噪模型性能对比实验结果

为验证本章提出的 OMLSA-TCN-GRU 语音降噪模型的性能，本小节将其与 OMLSA-IMCRA 语音降噪模型和实时语音降噪模型 DCCRN [68]、RNNoise 模型 [30] 以及 NSNet2 [69] 在 VoiceBank-Demand 测试集的实验结果进行了对比，实验结果如表3-4所示。

表 3-4 VoiceBank-Demand 测试集对比实验结果

	参数量 (M)	PESQ	STOI(%)	COVL
OMLSA-IMCRA	-	2.29	88.2	2.69
RNNNoise* [30]	0.06	2.33	92.2	2.84
NSNet2* [69]	6.17	2.47	90.3	**2.90**
DCCRN* [68]	3.7	2.54	93.8	2.75
OMLSA-TCN-GRU	0.56	**2.57**	**93.9**	2.85

* 数据来源于论文 [35]

本章节利用三个客观指标来评估不同模型的性能，即 PESQ 、STOI 和 COVL，它们与人类的感知质量、语音可懂度和语音整体听觉质量密切相关。从表3-4中可以看出，本章节提出的降噪模型有着优越的性能，在使用更少的网络参数的情况下，PESQ 和 STOI 得分方面都超过了 NSNet2 和 DCCRN 实时语音降噪模型，虽然参数量比 RNNNoise 模型更多，但是在语音评价指标得分上大幅领先 RNNNoise 模型。在 COVL 得分上本章提出的模型没有达到最优，这可能是本章提出的模型会对原始语音造成一定的语音损伤。

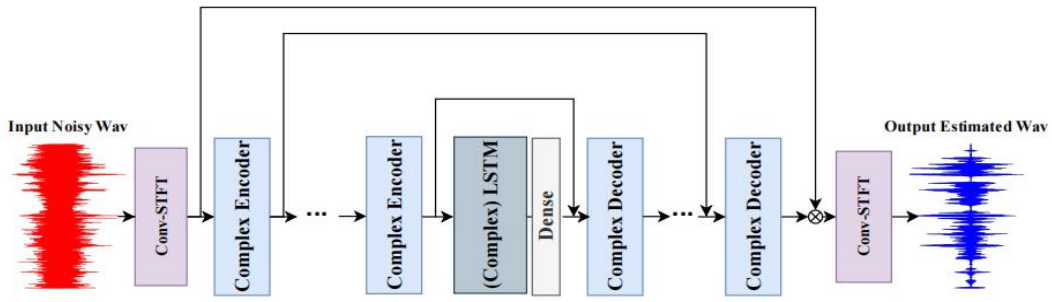


图 3-14 DCCRN 网络结构图

为了更细致的分析模型的性能，本章提出的方法与基线方法与 DCCRN 模型在 DeNoiseBank 测试集上在不同信噪比下进行了性能对比实验，DCCRN 模型是一个结合了 DCUNET 和 CRN 的网络模型，曾在 Interspeech 2020 深度噪声抑制（DNS）挑战赛实时降噪赛道排名第二。图3-14对 DCCRN 的网络结构进行了简要展示。实验结果如表3-5 、表3-6、表3-7所示，分别表示本章方法与其他模型在 PESQ、STOI、COVL 客观语音质量评价指标上的得分。

表 3-5 不同信噪比下降噪算法 PESQ 得分表

	参数量 (M)	**0dB**	**5dB**	**10dB**	**15dB**	**Avg.**
OMLSA-IMCRA	-	2.16	2.66	3.00	3.34	2.79
DCCRN [68]	3.7	**2.63**	**3.03**	3.35	3.63	3.16
OMLSA-TCN-GRU	0.56	2.50	2.99	**3.38**	**3.80**	**3.17**

表 3-6 不同信噪比下降噪算法 STOI(%) 得分表

	参数量 (M)	**0dB**	**5dB**	**10dB**	**15dB**	**Avg.**
OMLSA-IMCRA	-	62.50	74.77	83.82	90.14	77.81
DCCRN [68]	3.7	**73.29**	**83.13**	89.43	93.43	84.82
OMLSA-TCN-GRU	0.56	72.26	82.95	**90.02**	**94.17**	**84.85**

表 3-7 不同信噪比下降噪算法 COVL 得分表

	参数量 (M)	**0dB**	**5dB**	**10dB**	**15dB**	**Avg.**
OMLSA-IMCRA	-	2.73	2.89	3.01	3.09	2.93
DCCRN [68]	3.7	3.03	3.15	3.26	3.37	3.20
OMLSA-TCN-GRU	0.56	**3.04**	**3.25**	**3.47**	**3.61**	**3.34**

DCCRN 模型可获取:https://github.com/huyanxin/DeepComplexCRN

本章节利用三个客观指标来评估不同模型的性能，即 PESQ 、STOI 和 COVL，它们与人类的感知质量、语音可懂度和语音整体听觉质量密切相关。从表3-5、表3-6、表3-7中结果来看，可以观察到以下现象。首先，本节所提出的 OMLSA-TCN-GRU 降噪模型在 PESQ、STOI 和 COVL 方面都明显优于基线算法 OMLSA-IMCRA。与 OMLSA-IMCRA 相比，在 PESQ、STOI 被 COVL 方面，OMLSA-TCN-GRU 降噪

模型平均高出 12.5%、7.3% 和 7.8%，在低信噪比和高信噪比的情况下的评价指标相比原算法都有较大的改善。这表明所提出改进方法比原方法有更好的性能。

其次，与 DCCRN 降噪模型相比，本章节提出的 OMLSA-TCN-GRU 降噪模型在使用更少的模型参数的情况下有着相似的降噪性能，在 PESQ、STOI 和 COVL 得分方面相差不超过 2.5%，且在信噪比较高的条件下，本章提出的模型有着更好的降噪效果。但是对于相对较低的信噪比，DCCRN 似乎相对更有利。例如，在 0dB 下，DCCRN 比 OMLSA-TCN-GRU 实现了高约 5.2% 的改善、高约 2.8% 的 STOI 改善以及相近的 COVL 得分。这可能是因为映射能力有限、参数量较少的单级降噪模型往往无法很好地完成相对困难的任务，当训练后的模型应用于更复杂的真实环境时，可能会对纯净语音本身带来一定的损伤，引入一些失真，从而降低主观质量。这个问题会在第四章节解决。

同时，本章节对三个语音降噪模型产生的降噪后语谱图进行评估，图3-15表示 9s 纯净语音的语音信号波形图和语谱图。

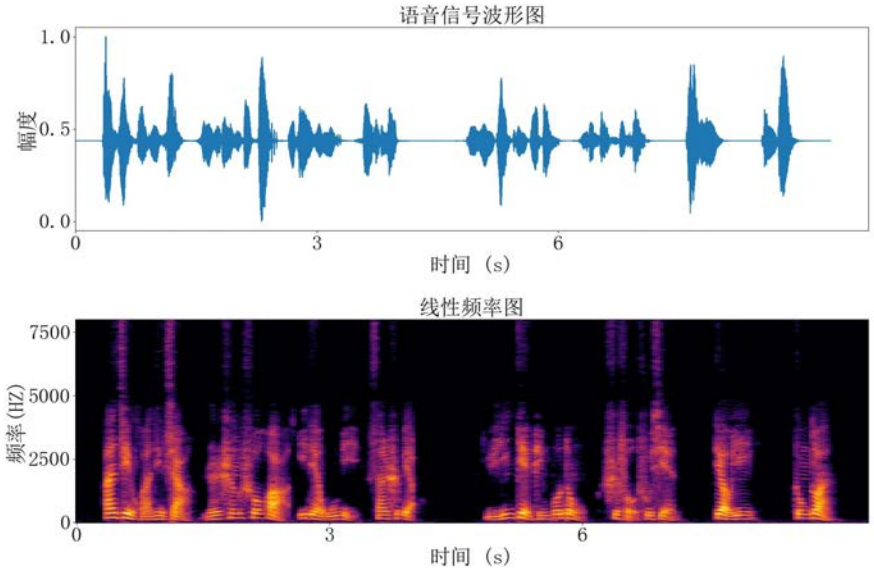


图 3-15 9s 纯净语音波形-语谱图

图3-16(a) 表示 9s 纯净语音和 Noise-92 噪声集中的 factory2 噪声以 15dB 的信噪比条件混合之后的噪声图，图3-17(b)-(d) 表示使用 OMLSA-IMCRA 降噪算法、DCCRN 降噪算法和本章提出的 OMLSA-TCN-GRU 改进算法分别进行降噪得到的降噪后语音语谱图。图3-16(a) 表示相同的 9s 纯净语音和 Noise-92 噪声集中的 factory2 噪声以 0dB 的信噪比条件混合之后，图3-17(b)-(d) 表示使用 OMLSA-IMCRA 降噪算法、DCCRN 降噪算法和本章提出的 OMLSA-TCN-GRU 改进算法分别进行降噪得到的降噪后语音语谱图。

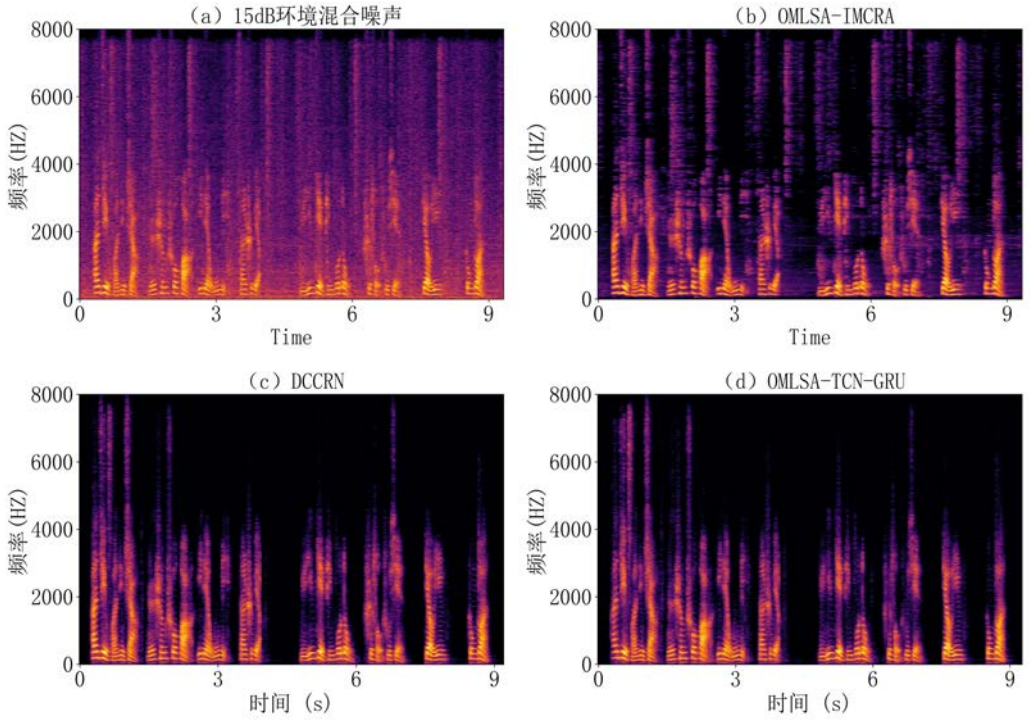


图 3-16 15db 信噪比环境下不同模型降噪效果语谱图

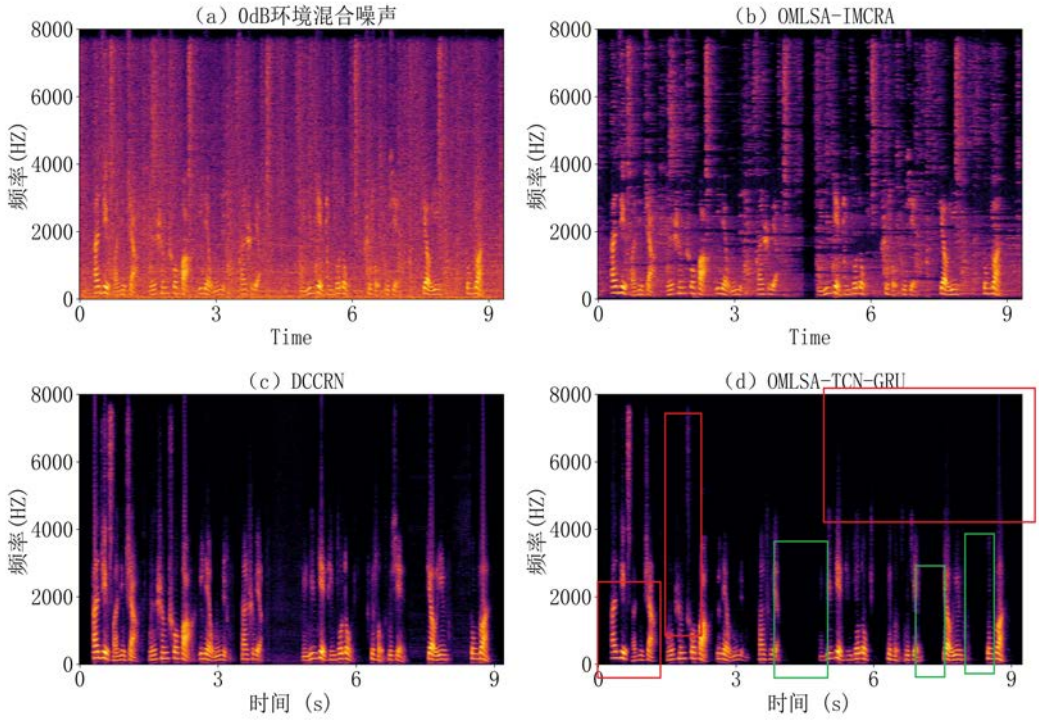


图 3-17 0db 信噪比环境下不同模型降噪效果语谱图

从图3-16中可以更直观的看出，在 15dB 信噪比的噪声环境下，由 OMLSA-

IMCRA 方法产生的降噪语音虽然一定程度上消除了部分稳态噪声，但是仍然表现出较多的的残余噪声，而经由 OMLSA-TCN-GRU 模型产生的降噪后语音表现出比 OMLSA-IMCRA 模型更少的残留噪声，基本上可以将噪声去除，红色矩形框位置展示了原算法效果不如本章提出的改进算法的地方。同时从语谱图上直观的对比可以看到与 DCCRN 模型有着大致相同的降噪效果。

从图3-17中可以直观的看出，在 0dB 信噪比的噪声环境下，由 OMLSA-IMCRA 方法产生的降噪语音表现出大量的残余噪声，几乎无法消除背景中的噪声。而经由 OMLSA-TCN-GRU 模型产生的降噪后语音仍然表现出比 OMLSA-IMCRA 模型更少的残留噪声，能够有效的去除大部分噪声。并且从语谱图中直观的看到，本章提出的改进算法在消除稳态噪声方面似乎有着更好的效果，如绿色矩形框所示。但有些地方的噪声消除能力不如 DCCRN 模型，并且会在一定程度上对原始信号产生损伤，这可能会导致降噪后的语音出现小幅的机械音、掉音、失真等情况，如红色矩形框所示。

通过以上分析。可以看出 OMLSA-TCN-GRU 模型在语音降噪任务上的优越性，能够在信噪比较高的环境下很好的完成降噪任务，效果远高于原算法并且能够将在参数量和计算量较小的情况下与最新的降噪算法效果一致。在低信噪比的环境下表现优秀，但是仍然有改进的空间，低信噪比环境下的降噪算法改进策略将在第四章中提出。

3.3.6 消融实验

接下来，为验证联合 TCN-GRU 模块中的 TCN 网络和 GRU 网络的在模型中的作用，本小节设计了消融实验，在信噪比 0dB，5dB，10dB，15dB 的噪声环境下进行了实验，实验结果见表3-8。

表 3-8 消融实验结果

模型编号	模块名		参数量	评价指标		
	TCN	GRU		PESQ Avg.	STOI Avg.(%)	COVL Avg.
Model1	✓		0.49	3.08	80.27	2.95
Model2		✓	0.08	2.71	75.94	2.94
Model3	✓	✓	0.57	3.13	83.75	3.15

表中 TCN 和 GRU 分别表示 TCN 网络模块，GRU 网络模块。Model1 表示不使用 GRU 模块，只设计 TCN 模块的语音降噪噪声谱估计模型；Model2 表示不使用 TCN 模块，只设计 GRU 模块的语音降噪噪声谱估计模型；Model3 表示同时使

用 TCN 网络模块和 GRU 网络模块的完整 TCN-GRU 语音降噪噪声谱估计模型。从表中的实验结果分析可知：

(1) 对比 Model1 和 Model2 的实验结果，可以看出 TCN-GRU 中的 TCN 模块降噪性能明显优于 GRU 模块，这可能是因为在设计网络的过程中使用了参数更多更深层次的 TCN 网络，而因为 GRU 网络无法进行并行计算，所以想在低功耗嵌入式设备上部署无法使用参数太多的 GRU 网络。

(2) 对比 Model2 和 Model3 的实验结果，可以看出设计的 TCN-GRU 联合网络性能远高于单独使用设计的 GRU 网络，比单一使用 GRU 网络在 PESQ、STOI 和 COVL 上分别完成了 15.4%、10.2%、7.1% 的改善，这说明了 TCN-GRU 联合网络中 TCN 网络的重要性，可能是作为联合网络中主要的特征学习网络。

(3) 对比 Model1 和 Model3 的实验结果，可以看出 TCN-GRU 联合网络比单独使用 TCN 网络在 PESQ、STOI 上分别完成了 1.6%、4.3% 的小幅度改善，但是在 COVL 上完成了 6.7% 的改善，这可能是因为 TCN 网络在降噪过程中虽然能够很好的抑制噪声，但是也对语音信号进行了一定程度的损伤。而加入了 GRU 模块后减小了对语音信号的损伤程度，这证明了 TCN-GRU 联合网络中 GRU 网络的重要性，使用非常少的参数量来对原始语音信号进行一定的保护是可以接受的。

3.4 本章小结

本章首先对常部署在低功耗嵌入式设备上的语音降噪算法 OMLSA-IMCRA 算法的降噪原理进行了说明和解释，并且通过仿真实验对该算法在应对稳态噪声和非稳态噪声时的表现进行了分析，实验结果该算法虽然有着计算量小的优点，在处理稳态噪声时有着不错的效果，但是在处理非稳态噪声时降噪效果较差，尤其是在处理低信噪比环境下的非稳态噪声时几乎无法处理噪声。

针对这一问题，本章分析了 OMLSA-IMCRA 算法的不足之处，发现该算法在处理非稳态噪声时效果不好的原因在于传统噪声谱估计方法 IMCRA 无法有效估计非稳态噪声环境下的噪声谱。所以本章在 OMLSA-IMCRA 的基础上进行了改进，结合最近流行的深度学习技术，提出了一种基于 TCN-GRU 联合网络的改进语音降噪模型。

最后，为验证所提出方法的有效性，基于公开和自制混合数据集开展了语音降噪仿真实验与消融实验。实验结果表明：

（1）TCN-GRU 联合网络中的 TCN 模块能够有效捕捉语音特征信息中的长期依赖性，GRU 模块帮助模型理解语音特征信息的动态变化，联合网络的降噪性能都优于单独的 TCN 网络和 GRU 网络

（2）本章提出的方法在 PESQ、STOI、COVL 这三个客观语音质量评分上对于原算法都分别有 12.5%、7.3% 和 7.8% 的改善，这表明该模型拥有较好的语音降噪能力。

（3）与其他基于深度学习的方法对比，本章节提出的算法在高信噪比条件下能以更少的参数达到同样的效果，但是在低信噪比条件下效果弱于对比方法。

第四章 基于 STDCT 变换的语音降噪后处理模型

传统的降噪算法在某些情况下达到不错的性能，但因为无法在非稳态噪声环境中无法跟踪快速变化的噪声从而无法精确对噪声谱进行建模，导致其在非稳态噪声环境中降噪效果不佳，其他深度学习的方法虽然在一定程度上提高了算法的抗噪性和鲁棒性，但大多数研究是在较为理想情况下进行的，并没有考虑到外部复杂环境的干扰。在第三章中，本文提出的模型在相对理想条件下展现了不错的降噪性能，获得了不错的语音质量评分，但是由于映射能力有限，单级降噪模型往往无法很好地完成相对困难的任务。而且由于由于模型使用的是纯净语音和噪声语音的合成带噪语音进行训练，当训练后的模型应用于更复杂的真实环境时，可能会对纯净语音本身带来一定的损伤，引入一些失真，从而大大降低主观质量，导致降噪后的语音变得非常不自然。

针对这个问题，本章在第三章提出的模型的基础上引入了低复杂度的 STDCT-DNet(Short Time Discrete Cosine Transform Denoise Net) 网络模型作为后处理模块来进一步抑制第三章模型输出中的残余噪声，提高模型的抗噪性能和降噪后语音的主观质量。该后处理模块基于 STDCT 特征学习网络，能够通过恢复带噪音的相位信息以及更加精确的恢复带噪语音的幅度信息从而更好的抑制残余噪声。通过对所提出算法与现有的多阶段语音降噪算法进行仿真实验和比较分析，验证了本章提出的噪声后处理模块在实时语音降噪方面的有效性。

4.1 多阶段语音降噪系统

主流的 SE 方法在时频（T-F）域处理语音信号 ancement。具体来说，首先通过 T-F 变换将含噪语音波形转换为 T-F 频谱。然后使用频域特征或压缩后的频域特征进行降噪，或者基于深度学习的方法会将其输入深度神经网络（DNN），训练该网络以预测目标语音的频谱或相应的频谱掩模。最后，增强后的语音通过 T-F 逆变换重建，在现有的工作中，短时傅立叶变换（STFT）是最常用的 T-F 变换方法。

长期以来，TF 域方法仅关注恢复目标幅度谱而不保留相位信息[70]。然而，后来证明精确的相位恢复可以进一步提高 SE 性能[71]。由于语音的相位谱不像幅度谱那样具有明显的结构特征，许多方法试图间接预测复域中的相位。然而，构建单个深度学习网络以在一个阶段中准确预测目标复频谱仍然具有挑战性。为了缓解这个问题，多阶段语音降噪算法提出了一种新的优化算法，将原始的单级优化任务分解为多个简单的、渐进的子任务，常见的是二级任务，第一级主要负责幅度谱

增强，从而实现对噪声的粗去除。第二阶段进一步预测复频谱的残余分量，从而抑制残余噪声，修复语音相位，已有实验证明两阶段算法优于传统的单阶段算法[48]。

4.2 基于短时离散余弦变换的 STDCT-DNet 噪声后处理模型

在语音信号处理领域，常用的语音特征有时域信号特征，语音频谱特征，梅尔频谱特征，梅尔倒谱特征等，这些特征要么是时域特征，要么是经过短时傅立叶变换得到的频域特征，但是可能丢失一些隐式信息[72]。而短时离散余弦变换（STDCT）可以将波形变换到实值的 T-F 域，是一种实值 TF 变换，变换后的频谱同时包含语音的幅度信息和相位信息，是另外一种频域特征。

某些领域中，基于 STDCT 的深度学习方法已经由实验证明有更好的效果。比如在语音分离领域[72]、语音识别领域[57]，被证明在参数大小相当的情况下，基于 STDCT 频域特征的模型比基于 STFT 特征变化的模型的性能有着更好的效果。在语音分类领域，基于 STDCT 频域特征的模型能在参数大小相同的情况下有着更高的分类准确率。除此之外，在语音降噪领域，2023 年 Wu 等人[73] 通过大量的实验证明，多通道信号间的相位信息是重要的特征，值得使用复值网络来恢复那些被显式表达的相位信息，但是在单声道语音增强领域中使用复值网络来恢复那些被显式表达的相位信息会增加更多的计算量，且并没有比实值深度神经网络有更好的性能。并且当模型大小相对较小时，使用复值运算甚至会对增强性能产生不利影响。

受此启发，本章节提出了 STDCT-DNet 降噪模型作为语音降噪后处理模块，通过 STDCT 变换域中包含的隐式相位信息，而非短时傅里叶变换得到的 T-F 域特征来恢复带噪语音的相位信息，同时进一步更精确的修复带噪语音的幅度信息。

4.2.1 STDCT-DNet 噪声后处理模块架构与两阶段降噪模型架构

由于 STDCT 谱是与梅尔频谱一样的实值谱，因此本章提出的 STDCT-DNet 网络结构域上一章提出的 TCN-GRU 噪声谱预测网络结构几乎相同。但是考虑到 STDCT 谱同时包含幅度信息和相位信息，所以需要一个效率更高的网络来完成 STDCT 谱的恢复，同时精细修复带噪语音的幅度信息并且修复带噪语音的相位信息。

所以本章提出的 STDCT-DNet 网络在上一章的 TCN-GRU 网络的基础上，添加了自注意力机制，从而更好的去捕捉远距离依赖，使网络能够更加聚焦于有用的时序特征。

经过实验验证最合适的多头自注意力模块参数后，最终设计的 STDCT-DNet

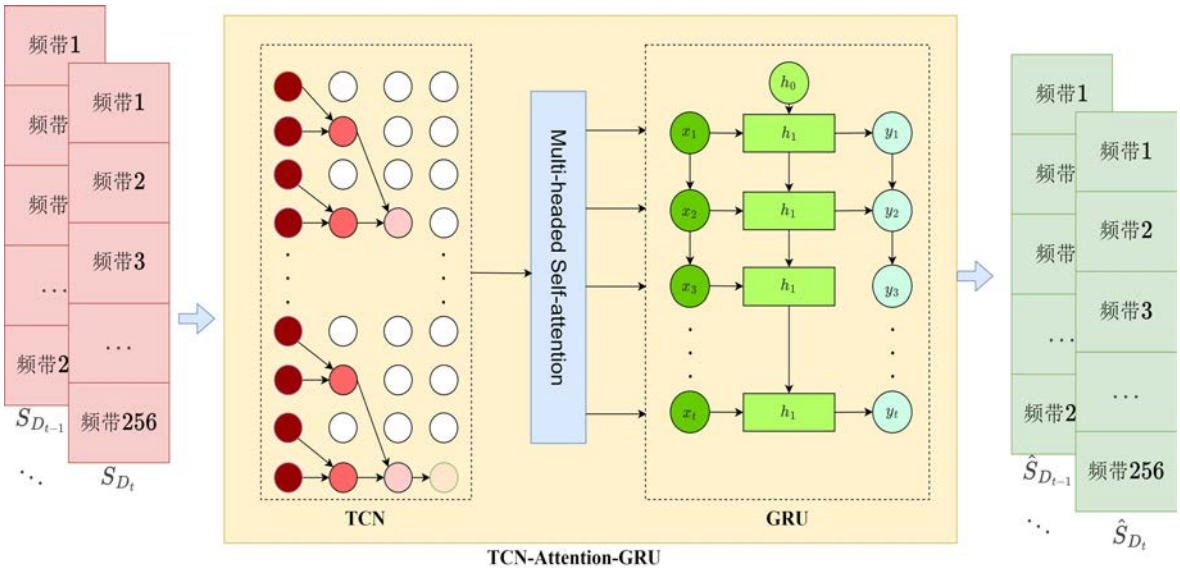


图 4-1 STDCT-DNet 噪声后处理模块框架图

噪声后处理模块结构如图4-1所示。对多头自注意力模块的参数设置的实验验证将在后文给出。

最终在第三章提出的 OMLSA-TCN-GRU 模型的基础上，引入本章节提出的 STDCT-DNet 噪声后处理模块，得到最终的两阶段降噪模型，降噪流程框架图如图4-2所示。

如流程图所示，两阶段降噪模型首先对带噪语音进行 STFT 变换和 STDCT 变换，从而获得带噪语音的的 STFT 频谱 $|X_F|$ 和 STDCT 频谱 S_{D_t}。然后 STFT 频谱 $|X_F|$ 的幅度信息作为第三章提出的 OMLSA-TCN-GRU 模型的输入，完成第一阶段的带噪语音幅度恢复任务，随后结合相位信息完成 ISTFT，恢复回时域信号。

对第一阶段已经完成过幅度恢复的降噪后语音 $\hat{x}_1$ 进行 STDCT 变换，然后结合带噪语音进行 STDCT 变换后得到的频谱 $S_{\hat{D}_t}$ 作为第二阶段 STDCT-DNet 噪声后处理模块的输入，用于对第一阶段的结果进行进一步细化，恢复相位信息并且精修复幅度信息。

4.2.2 TCN-Attention-GRU 模块

本章在第三章提出的 TCN-GRU 模块的基础上引入多头自注意力机制，关于多头注意力机制的相关原理已在第二章解释。提出的 STDCT-DNet 网络的具体网络结构如图4-3所示。

为了兼顾模型大小和模型的性能，本章设计的 STDCT-DNet 模型中加入的多头自注意力机制时，对 Transformer 的前馈神经网络层进行了修改，放弃了隐藏层，只使用一个全连接层作为前馈神经网络。同时调整了 TCN 网络的参数，调整膨胀

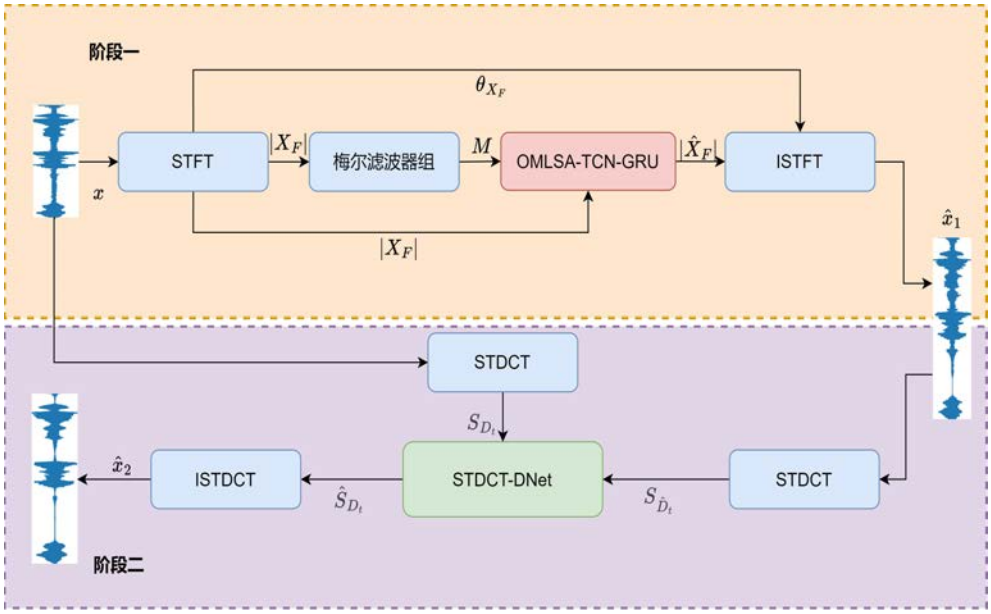


图 4-2 两阶段降噪模型流程图

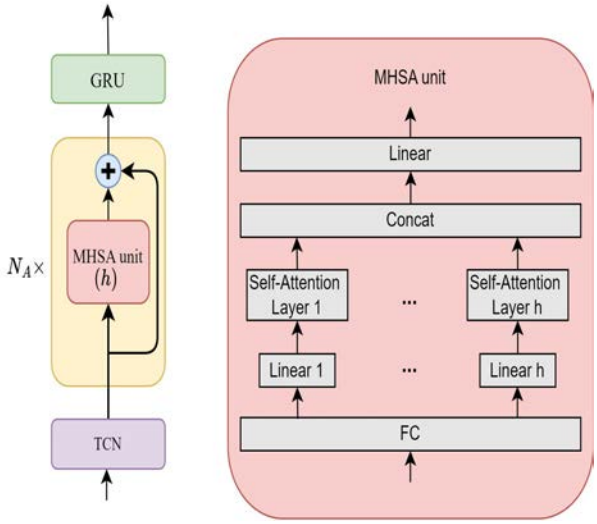


图 4-3 TCN-GRU 模块中引入的多头自注意力机制结构图

系数最大为 16，并减小时域卷积残差块级联层数 N_C，设置 N_C 为 20，并调整每个残差块的第一层和二层输出大小 O_1 和 O_2 为 32。这样在没有增加较多参数的情况下可以有更大的感受野，获取到更多的前文语音信息。GRU 网络结构保持不变。

4.2.3 训练策略及损失函数

类似于以前的两阶段工作[48]，本章在经过实验后最终选择的 STDCT-DNet 网络的训练策略是与第一阶段的 OMLSA-TCN-GRU 降噪模型进行联合训练，使用的损失函数如式4-1所示。

$$\mathcal{L}_{\text{STDCT-DNet}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}_{\text{stage2}} = \left\| \hat{M} - M \right\|_F^2 + \left\| \hat{S}_{\hat{D}_t} - S_{D_t mix} \right\|_F^2 \tag{4-1}$$

式中 $\mathcal{L}_{\text{stage1}}$ 表示 OMLSA-TCN-GRU 降噪模型 TCN-GRU 噪声谱预测模型的 MSE 损失函数，$\mathcal{L}_{\text{stage2}}$ 表示 STDCT-DNet 网络的 MSE 损失函数。$\hat{S}_{\hat{D}_t}$ 表示预测的 STDCT 谱值，$S_{D_t mix}$ 表示干净语音和第一阶段得到的降噪语音 STDCT 变换后的混合谱值。

4.3 实验验证分析

4.3.1 实验环境

实验环境与第三章实验环境一致。

4.3.2 实验设置

OTG-STDCT-DNet 两阶段降噪模型在训练阶段的超参数设置如表4-1所示。

表 4-1 超参数配置

超参数类型	参数
学习率	0.0005
优化器	Adam
Batch_size	64
Epochs	200
Dropout	0.2
最大频率	16000
帧长	32ms
帧移	16ms
窗口类型	汉明窗
DCT/STFT 点数	512

4.3.3 数据集

训练集和测试集与第三章所采用数据集一致。

4.3.4 训练策略与注意力参数设置有效性实验

以前的两阶段语音降噪模型，往往单独训练第一阶段的模型，并且其第二阶段的相位恢复模型选择的训练策略多为联合第一阶段模型进行训练，这是经过验证的最优的训练方式。但是这基于第一阶段模型是一个基于深度学习的端到端的带噪语音幅度谱修复模型这一前提，而本文在第三章提出的 OMLSA-TCN-GRU 带噪语音幅度谱修复模型并不是一个完全的端到端深度学习网络，所以在此对单独训练 STDCT-DNet 模型和 OMLSA-TCN-GRU 模型、OMLSA-TCN-GRU 与 STDCT-DNet 联合训练这三种不同的训练策略进行实验，以选择对于第二阶段 STDCT-DNet 网络的最优训练策略。

同时因为 STDCT-DNet 网络在第三章设计的 TCN-GRU 联合网络的基础上加入了自注意力机制，所以同时对自注意力机制的参数设置进行实验，以选择最优的参数设置。设置单层自注意层计算量低，模型简单，可能无法充分学习长时间依赖，而设置四层或以上自注意力层数更适合处理长时依赖最佳，但是推理计算成本很高，不适用于实时任务，更适用于 Transformer-based 语音增强。考虑到最终算法要部署在低功耗嵌入式设备上，所以本节分别设置 B=2 和 B=3 进行实验，详细实验设置如下。

(1) 实验一

单独训练 STDCT-DNet 模型和 OMLSA-TCN-GRU 模型，多头自注意力层数设置为两层。OMLSA-TCN-GRU 模型训练方法已在第三章说明。STDCT-DNet 模型选用的损失函数如式4-2所示：

$$\mathcal{L}_{\text{STDCT-DNet}} = \left\| \hat{S}_{\hat{D}_t} - S_{D_t mix} \right\|_F^2 \tag{4-2}$$

式中 $\hat{S}_{\hat{D}_t}$ 表示预测的 STDCT 谱值，$S_{D_t mix}$ 表示干净语音和第一阶段得到的降噪语音 STDCT 变换后的混合谱值。

(2) 实验二

单独训练 STDCT-DNet 模型和 OMLSA-TCN-GRU 模型，多头自注意力层数设置为三层。本次实验选用的损失函数如实验一式4-2所示。

(3) 实验三

与 OMLSA-TCN-GRU 降噪模型联合训练 STDCT-DNet 模型，自多头自注意力

层数设置为两层，本次实验选用的损失函数如式4-3所示。

$$\mathcal{L}_{\text{STDCT-DNet}} = \mathcal{L}_{\text{stage1}} + \mathcal{L}_{\text{stage2}} = \left\|\hat{M} - M\right\|_F^2 + \left\|\hat{S}_{\hat{D}_t} - S_{D_t mix}\right\|_F^2 \tag{4-3}$$

式中 $\mathcal{L}_{\text{stage1}}$ 表示 OMLSA-TCN-GRU 降噪模型 TCN-GRU 噪声谱预测模型的 MSE 损失函数，$\mathcal{L}_{\text{stage2}}$ 表示 STDCT-DNet 网络的 MSE 损失函数。$\hat{S}_{\hat{D}_t}$ 表示预测的 STDCT 谱值，$S_{D_t mix}$ 表示干净语音和第一阶段得到的降噪语音 STDCT 变换后的混合谱值。

(4) 实验四

与 OMLSA-TCN-GRU 降噪模型联合训练 STDCT-DNet 模型，多头自注意力层数设置为三层。本次实验选用的损失函数如实验三式4-3所示。

最终对实验提出的两种训练策略和两种多头自注意力层数设置进行实验，共进行四次实验。使用第三章提出的的测试数据集对最终得到的两阶段语音降噪模型进行测试，最终得到的客观语音质量评价指标 PESQ、STOI、COVL 结果如表4-2所示。

表 4-2 不同训练策略和不同参数设置下评价指标得分表

	参数量 (M)	**PESQ Avg.**	**STOI Avg.(%)**	**COVL Avg.**
实验一	1.87	3.27	84.85	3.62
实验二	2.26	3.09	85.12	3.63
实验三	1.87	3.31	85.21	3.65
实验四	2.26	3.32	85.43	3.65

从表中的实验结果分析可知：

(1) 对比实验三和实验一，实验四和实验二，发现当多头自注意力层数相同的情况下，通过联合训练的方法训练 STDCT-DNet 模型比单独训练 STDCT-DNet 模型得到的两阶段模型降噪效果更好。其中实验三训练出的模型在 PESQ、STOI、COVL 得分上比实验一训练出的模型改善了 0.9%、0.1% 和 0.8%，实验四训练出的模型在 PESQ、STOI 和 COVL 得分中比实验二训练出的模型改善了 0.9%、0.4%、0.5%。这可能是因为每个子模型单独优化自己的目标，容易造成信息损失，导致最终模型的效果受到限制，早期阶段的错误可能被后续阶段放大，最终影响整体性能。

(2) 对比实验四和实验三，可以看到在同时采用联合训练的条件下，实验四训

练出的模型在 PESQ、STOI 得分上比实验三训练出的模型改善了 0.3%、0.2%，在 COVL 得分上没有明显的改进。这可能是实验使用的训练数据有限，当使用的注意力机制过于复杂，可能导致过拟合，丢失了部分真实语音细节。

基于此，本章节选择联合训练作为 STDCT-DNet 网络的训练方式，同时考虑到低功耗嵌入式设备上的模型部署问题，三层多头自主力机制的使用并没有显著的提高两阶段降噪模型的性能，反而增加了 0.39M 左右的参数量，所以最终选择使用两层多头自注意力机制，设置 $N_A = 2$。

4.3.5 OTG-STDCT-DNet 降噪模型性能对比实验

验证本章提出的基于 STDCT 的语音降噪后处理模块 STDCT-DNet 的能够为语音降噪任务带来更加优异的性能，本小节将其与第三章提出的 OMLSA-TCN-GRU 降噪模型、RNNNoise 模型[30]、DCCRN 模型[68]、以及两阶段实时降噪模型 DeepfilterNet2[35] 以及两阶段实时降噪模型 CCFNet+(Lite)[74] 在 VoiceBank-Demand 测试集上行了对比实验，实验结果如表4-3所示。

表 4-3 VoiceBank-Demand 测试集对比实验结果

	参数量 (M)	**PESQ**	**STOI(%)**	**COVL**
OMLSA-IMCRA	-	2.29	88.2	2.69
RNNNoise* [30]	0.06	2.33	92.2	2.84
NSNet2* [69]	6.17	2.47	90.3	**2.90**
DCCRN* [68]	3.7	2.54	93.8	2.75
OMLSA-TCN-GRU	0.56	**2.57**	**93.9**	2.85
DeepfilterNet2*	2.31	**3.08**	**94.3**	3.70
CCFNet+(Lite)† [74]	0.16	2.94	-	-
OTG-STDCT-DNet	1.87	3.06	**94.3**	**3.72**

* 数据来源于论文 [35]，† 数据来源于论文 [74]

本章利用三个客观指标来评估不同模型的性能，即 PESQ、STOI 和 COVL，已在第三章对这三个指标进行过简要描述。从表4-3中可以看出，在第三章提出的 OMLSA-TCN-GRU 模型基础上添加了本章节提出的降噪后处理模块之后得到的两阶段降噪模型有着优越的性能，在 COVL 评价指标上得到了最高分，并在使用了更少的参数的情况下在 STOI 评价指标上得到了与 DeepfilterNet2 模型相同的得分。与第三章提出的 OMLSA-TCN-GRU 模型相比，在 PESQ、STOI 和 COVL

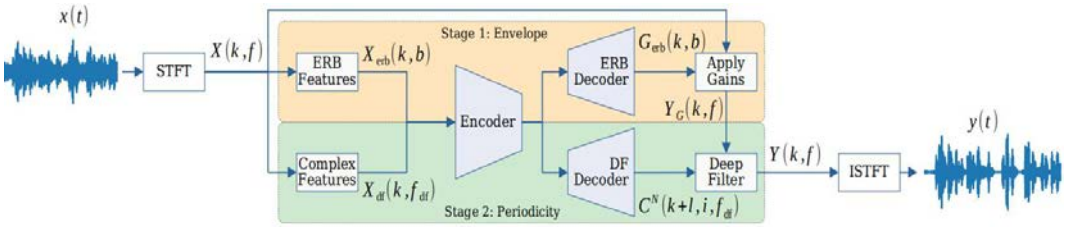


图 4-4 Deepfilternet2 网络结构图

得分上分别改善了 18.2%、0.4%、30%。和其他单网络模型 (RNNNoise、DCCRN) 相比，在 PESQ、STOI 和 COVL 得分指标上也有着较大的改善。与其他两阶段降噪模型相比提出的 OTG-STDCT-DNet 降噪模型也有着优越的性能，在 STOI 指标和 COVL 和指标得分上均优于或等于 DeepfilterNet2 模型和 CCFNet+(Lite) 模型，并在使用更少的参数的情况下在 PESQ 指标得分上仅落后于 DeepfilterNet2 模型 0.6%。虽然 OTG-STDCT-DNet 降噪模型的参数比 CCFNet+(Lite) 模型更多，但是 OTG-STDCT-DNet 模型中的可并行化计算量较多，仍然能够满足嵌入式设备实时降噪的部署要求，所以在降噪效果优于 CCFNet+(Lite) 模型的情况下有更好的应用价值。

为了更细致的分析模型的性能以及后处理模块的作用，本章提出的方法与第三章提出的模型、DCCRN 模型和 DeepfilterNet2 降噪模型在 DeNoiseBank 测试集上不同信噪比环境下进行了性能对比实验。DCCRN 模型已在上一章进行了简述，DeepfilterNet2 降噪模型是一个两阶段实时降噪模型，同样设计了二阶段相位恢复模型，最初被提出用于实时语音增强，可以部署在低功耗嵌入式设备中完成降噪任务，图4-4对 DeepfilterNet2 的网络结构进行了简要展示。

在加入本章提出的后处理模块后，两阶段模型整体降噪性能与其他几种模型性能对比实验结果如表4-4、表4-5、表4-6所示，其中 OTG 表示第三章提出的 OMLSA-TCN-GUR 降噪模型的缩写。

表 4-4 不同信噪比下降噪算法 PESQ 得分表

	参数量 (M)	**0dB**	**5dB**	**10dB**	**15dB**	**Avg.**
DCCRN [68]	3.7	2.63	3.03	3.35	3.63	3.16
OTG	0.56	2.50	2.99	3.38	3.80	3.17
DeepfilterNet2 [35]	2.31	2.89	3.22	3.50	3.72	3.33
OTG-STDCT-DNet	1.87	2.85	3.20	3.48	3.74	3.31

表 4-5 不同信噪比下降噪算法 STOI% 得分表

	参数量 (M)	0dB	5dB	10dB	15dB	Avg.
DCCRN [68]	3.7	73.29	83.13	89.43	93.43	84.82
OTG	0.56	72.26	82.95	90.02	94.17	84.85
DeepfilterNet2 [35]	2.31	74.68	84.56	91.01	95.10	86.33
OTG-STDCT-DNet	1.87	74.49	84.43	90.92	95.02	86.21

表 4-6 不同信噪比下降噪算法 COVL 得分表

	参数量 (M)	0dB	5dB	10dB	15dB	Avg.
DCCRN [68]	3.7	3.03	3.15	3.26	3.37	3.20
OTG	0.56	3.04	3.25	3.47	3.61	3.34
DeepfilterNet2 [35]	2.31	3.44	3.58	3.68	3.74	3.60
OTG-STDCT-DNet	1.87	3.41	3.59	3.67	3.80	3.61

DCCRN 模型可获取:https://github.com/huyanxin/DeepComplexCRN

DeepfilterNet2 模型可获取:https://github.com/Rikorose/DeepFilterNet

从表4-4、表4-5、表4-6中结果来看，可以观察到以下现象：

(1) 首先，在第三章提出的 OMLSA-TCN-GRU 降噪模型的基础上添加了本章所提出的 STDCT-DNet 噪声后处理模型后，在 PESQ、STOI 和 COVL 方面都明显有了较大的改善。与 OMLSA-TCN-GRU 降噪模型相比，分别在 PESQ、STOI 被 COVL 得分上改善了约 4.4%、1.6% 和 8.0%，尤其是在低信噪比环境下，比如在 0dB 信噪比噪声环境下比 OMLSA-TCN-GRU 降噪模型改善了约 14.0%、3.0% 和 12.1%。这表明引入的 STDCT-DNet 噪声后处理模型在相位恢复和幅值精修复方面起到了非常好的效果，削弱和修复了第一阶段 OMLSA-TCN-GRU 降噪模型带来的语音损失。这同时也说明了相位信息在语音降噪中有着重要的作用。

(2) 其次，与 DCCRN 降噪模型相比，引入本章所提出的 STDCT-DNet 噪声后处理模型后，无论是在低信噪比还是高信噪比的环境下，在 PESQ、STOI 和 COVL 得分方面同样有着大幅的改善，这说明在对语音信号的不同特征进行处理时，采用单一网络对单一的特征处理可能会有更好的效果。虽然 DCCRN 模型采用了复数网络同时对带噪语音幅值和相位信息进行了修复，但是使用单一的网络来处理单一的信息可能会有更好的效果。

(3) 最后，与 DeepfilterNet2 模型相比，本文提出的模型在不同信噪比环境下，使用了更少的参数量的情况下达到了几乎一样的效果，虽然在第三章节提出的第一阶段模型的传统算法可能会引入额外的计算量，但是由于传统算法为数字信号处理算法，有着很好的并行计算特性，可以使用 cpu 进行计算，同时与使用 GPU/NPU 的深度学习网络同时完成一些计算，所以可以在低功耗嵌入式设备上部署本章与第三章提出的多阶段降噪模型会有着更快的推理时间，足够达到实时性的要求。这一点会在第五章进行证明。

同时，本章节对三个语音降噪模型产生的降噪后语谱图进行评估，图4-5-a 表示纯净语音和 Noise-92 噪声集中的咖啡厅噪声以 0dB 的信噪比条件混合之后的噪声图，图4-5(b-e) 表示使用 DCCRN 降噪算法、第三章提出的 OMLSA-TCN-GRU 降噪算法、Deepfilternet2 降噪算法和本章提出的 OTG-STDCT-DNet 降噪算法分别进行降噪得到的降噪后语音语谱图。

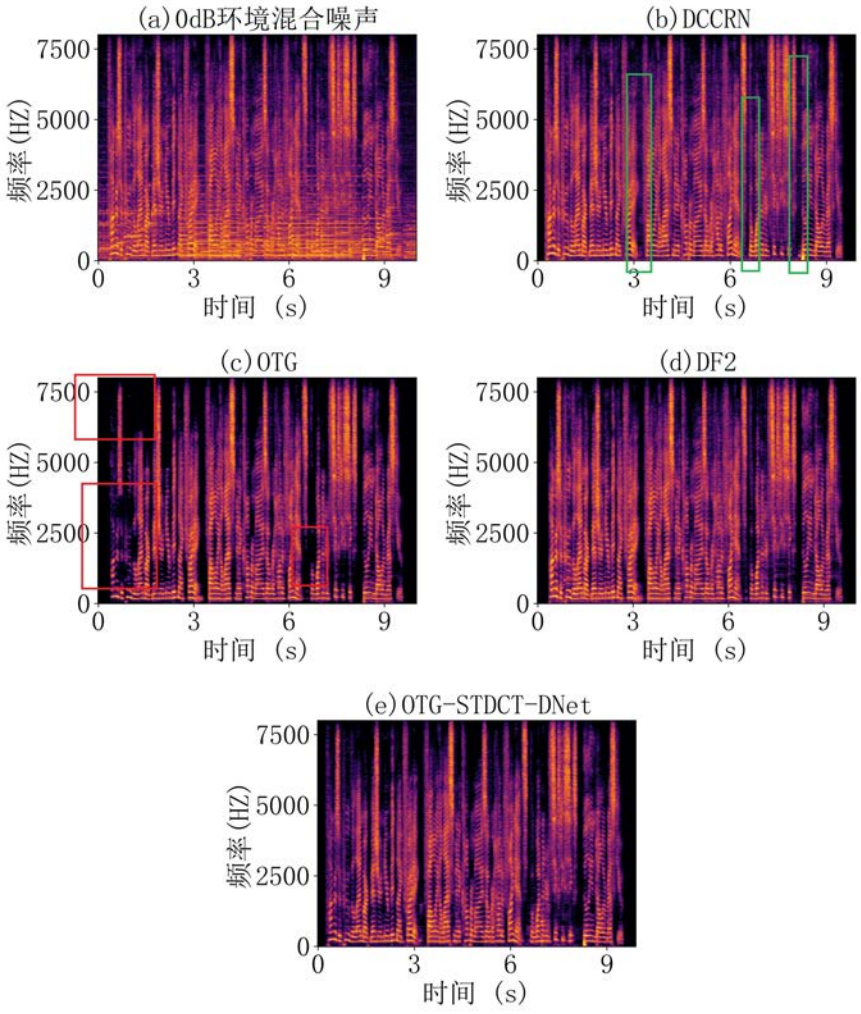


图 4-5 0db 信噪比环境下不同模型降噪效果语谱图

从图中可以看出，在 0dB 低信噪比的噪声环境下，在加入本章节提出的语音降噪后处理模块 STDCT-DNet 之后，解决了第三章提出的 OMLSA-TCN-GRU 模型的缺点，能够更好的消除噪声并且不会对原始语音信号产生较大的损伤，其效果可以从红色矩形框中看出。同时从语谱图中可以看出，最终的 OTG-STDCT-DNet 降噪模型在降噪效果上优于 DCCRN 模型，能够比 DCCRN 模型消除更多的非稳态噪声，这一点从绿色矩形框中可以看出。与 Deepfillternet2 相比，从语谱图上来看在更少的参数下几乎达到了相同的降噪效果。

4.3.6 消融实验

虽然在网络中添加注意力机制在语音领域已经有提升语音任务效果的先例，但同时也有研究表明在语音降噪模型中添加注意力机制并不保证一定能对性能提升有较大帮助，为了证明 STDCT-DNet 模型中添加的多头自注意力机制能帮助 STDCT-DNet 模型以及最终的两阶段降噪模型更好的消除噪声，本小节设计了消融实验，实验结果见表4-7。

表 4-7 消融实验结果

模型编号	模块名			参数量	评价指标		
	OTG	PPTG	Attention		PESQ Avg.	STOI Avg.(%)	COVL Avg.
Model1	✓			0.57	3.13	83.75	3.15
Model2	✓	✓		1.09	3.27	85.43	3.52
Model3	✓		✓	1.35	3.22	84.93	3.21
Model4	✓	✓	✓	1.87	3.31	86.21	3.61

表中 OTG、PPTG 和 Attention 分别表示 OMLSA-TCN-GRU 降噪模型、不添加多头自注意力机制的 STDCT-DNet 噪声后处理模型和添加多头自注意力机制的 STDCT-DNet 噪声后处理模型。Model1 表示只使用第三章提出的 OMLSA-TCN-GRU 降噪模型，Model2 表示使用 OMLSA-TCN-GRU 降噪模型和未引入多头自注意力机制的 STDCT-DNet 噪声后处理模型，Model3 表示只使用 OTG 和 Attention 模块。从表中的实验结果分析可知：

(1) 对比 Model1 和 Model3、Model1 和 Model3 的实验结果，可以看出在 STDCT-DNet 噪声后处理模型中使用 TCN-GRU 网络或者多头自注意力机制，都能有效的改善语音降噪的效果，这说明在 STDCT-DNet 噪声后处理模型中使用 TCN-GRU 网络或者多头自注意力机制都能够对相位信息进行一定的恢复。但是单独使用多头自注意力机制的效果没有单独使用 TCN-GRU 网络的改善效果明显，尤其是在

COVL 得分上，单独使用 TCN-GRU 网络的噪声后处理模型相对单阶段模型改善了约 11.7%，但是单独使用多头自注意力机制的噪声后处理模型相对单阶段模型只改善了约 1.9%，这可能是因为数据量小导致在注意力机制过于复杂时，出现了过拟合现象，导致对第一阶段模型的语音损失修复能力不足。

(2) 对比 Model2 和 Model4 的实验结果，可以看出在 STDCT-DNet 模型中引入多头注意力机制有更好的降噪效果，这可能是因为有些室内环境下纯净语音会受到混响、回声效应等影响，多头自注意力机制可以利用远程上下文信息，帮助去除长时间依赖的噪声，自动分配合适的注意力权重，更好地增强语音信号部分，而不是盲目抑制所有噪声。

4.4 本章小结

本章首先介绍了短时离散余弦变换在语音领域应用以及多头自注意力机制，然后提出了基于 STDCT 变换特征的 STDCT-DNet 噪声后处理模型，并在第三章建立的 TCN-GRU 噪声谱估计模型的基础上引入了多头自注意力机制。

之后为了找到合适的训练策略和设置合理的模型参数，本章开展了训练策略和参数设置实验，并在折中考虑性能和参数量之后选择了合适的训练策略和模型参数。

最后，为验证所提出方法的有效性，在第三章实验的环境和数据集基础上开展了语音降噪仿真实验与消融实验，实验结果表明：

（1）在第三章提出的 OMLSA-TCN-GRU 模型的基础上引入 STDCT-DNet 噪声后处理模型后，OTG-STDCT-DNet 两阶段降噪模型的性能有了较大的提升，在 PESQ、STOI、COVL 这三个客观语音质量评分上对于原算法都分别有 14.0%、4.5% 和 18.8% 的改善，这表明

（2）本章提出的方法在 PESQ、STOI、COVL 这三个客观语音质量评分上对于原算法都分别有 12.5%、7.3% 和 7.8% 的改善，这表明该模型拥有较好的语音降噪能力。

（3）与其他基于深度学习的两阶段模型对比，本章节提出的算法在低信噪比和高信噪比条件下都能以更少的参数量达到同样的效果，这有利于在低功耗嵌入式设备上的使用。

第五章 基于 SS928 的嵌入式单通道实时语音降噪系统设计实现

本章详细介绍了嵌入式实时语音降噪系统的搭建。选择海思 SS928 平台为核心构建系统硬件，软件部分分为上位机软件和下位机软件，上位机软件基于 Windows 平台使用 Python 语言编写效果展示与通信，下位机软件基于嵌入式平台 Linux 系统编写。通过实验验证实时嵌入式语音降噪系统的有效性。

5.1 嵌入式单通道实时语音降噪系统整体框架

本文设计的实时语音降噪系统整体框架如图所示。主要包括海思 SS928 主控模块、ES7210 模数转换模块、模拟麦克风和用于可视化降噪结果的上位机 PC 端等，涵盖了声音信号的获取和转换、外部设备与计算平台的数据双向传输、语音降噪、上位机下位机数据传输的完整流程，整体框架如图5-1所示。

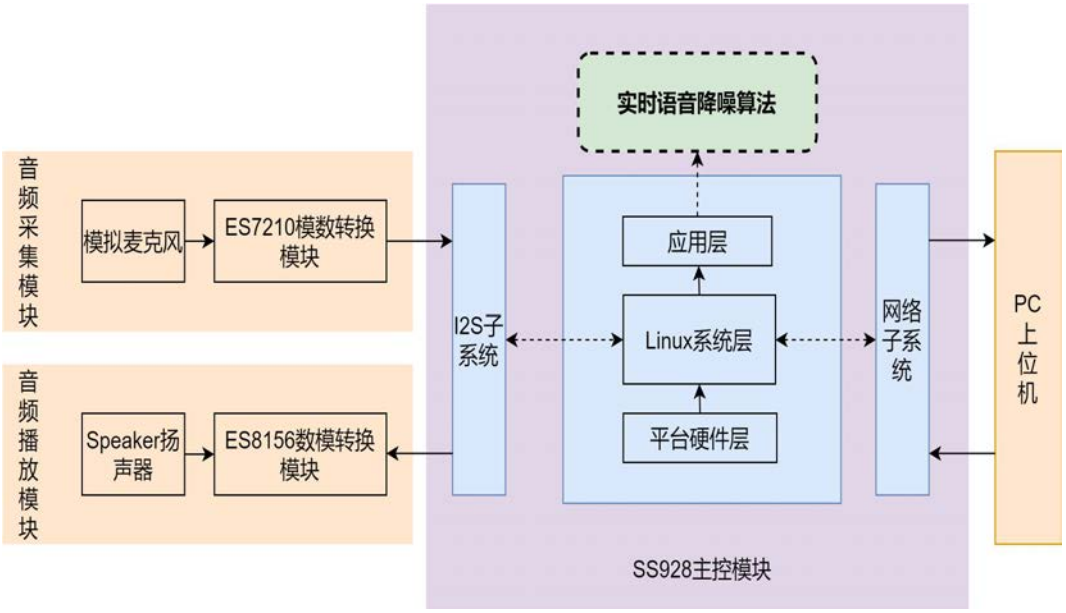


图 5-1 单通道语音降噪系统整体框架图

系统首先使用模拟麦克风将接收到的声音信号转换为电信号，并通过 ES7210 模数转换模块将模拟信号转换为 SS928 芯片可以处理的的数字信号。

随后通过 I/O 端口与 SS928 平台完成基于 I2S 协议的数字音频信号数据传输，并由部署在 SS928 主控模块台 linux 系统上的实时语音降噪算法完成语音降噪。

最后，由 SS928 主控模块将降噪后语音数据通过 TCP/IP 协议发送给上位机完成降噪效果可视化任务，并可选的发送给 ES8156 数模转换模块完成模数转换，传送给 speaker 播放语音。

5.2 嵌入式单通道实时语音降噪系统硬件设计

5.2.1 硬件开发平台

本文使用如图5-2所示的华为海思 SS928v100 平台作为主控芯片，在此基础上开发外围设备和软件。华为海思 SS928 平台是一个支持多 sensor 输入，支持最高 4K60 的 ISP 图像处理能力，支持多通道音频输入，支持多种音视频增强和处理算法的开发平台，为用户提供了卓越的多媒体处理能力。

图 5-2 SS928 核心板示意图

SS928v100 平台内置四核 ARM Cortex-A55，提供高效且丰富和灵活的 CPU 资源，以满足客户计算和控制需求，并支持 ARM-NEON 加速指令。集成单核 MCU，可以满足某些高实时性低延时要求较高场景。同时 SS928v100 平台集成了高效的分析推理单元 NPU，最高 4TopsINT8，并支持业界主流的工具框架。并内置双核 Vision DSP，以满足个人开发者的一些差异化的深度学习模型计算需求，同时有海思提供的配套 SDK，适合快速上手开发。SS928v100 平台可用来完成语音降噪的系统资源和计算资源可从表5-1中看到。

表 5-1 SS928v100 平台资源表

可用资源类型	描述
CPU-1	四核 ARM-CortexA55@1.4GHz，支持 NEON 加速和 FPU 浮点处理单元
CPU-2	单核 32bit MCU@500MHz，支持低延时处理
NPU	4Tops INT8 算力加速引擎
DSP	双核 Vision Q6 DSP
视频接口	支持 8-Lane image sensor 串行输入
音频接口	支持 I2S 接口 16bit 语音输入和输出支持 G.711/AAC/等音频编码格式
网络接口	提供双千兆以太网接口

5.2.2 麦克风和 Speaker 设计

从成本角度说，数字麦克风的价格往往是模拟麦克风价格的 2-3 倍，所以对于非专业音频领域的家庭消费类的嵌入式设备往往采用引线式模拟单端 MIC。带高端 XLR 电容麦的麦克风拾音效果更好但是价格也更高，考虑到成本问题应选择普通麦克风，选型指标如表5-2所示。

表 5-2 MIC 选型指标表

参数	指标
SNR(信噪比)	大于等于 58dB
Sensitivity(灵敏度)	-26dB
拾音范围类型	全向型
拾音距离	1m-3m 左右

综合成本和麦克性能，最终选择使用如图5-3所示的 SPU0410HR5H 作为语音降噪系统中使用的单通道 MIC。

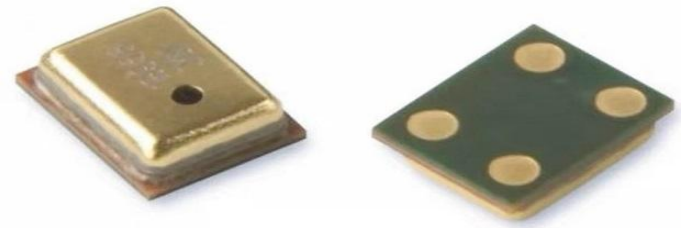


图 5-3 SPU0410HR5H 示意图

为搭建一个更完善的音频系统，本系统在硬件层面加入了扬声器，并可由软件控制是否使用。由于 Speaker 并不影响实时语音降噪算法的降噪性能，所以尽量从节约成本层面考虑，选型指标如表5-3所示。

表 5-3 speaker 选型指标表

参数	指标
SPL(声压级)	大于等于 89dB
单体基频	1KHz
灵敏度	94dB/1W/0.1m
失真率	1W 功率下失真率不超过 10%

5.2.3 模数转换芯片设计

常用的音频信号获取有两种方案，一种是使用数字 MIC，数字 MIC 内置的 Codec 会将模拟信号转为数字信号，但是此种方案的 MIC 成本较高。另一种方案是选择模拟 MIC 搭配外置 Codec，此方案在使用多 MIC 时成本优势明显，在只使用单声道 MIC 时没有明显的成本呢优势，但是外置 Codec 可以支持多种采样率。鉴于此，本文搭建的实时语音降噪系统采用模拟 MIC 配合外置 Coedc 的方式，一是为了可以自由的选择采样率，二是保留了以后使用多通道 MIC 的可能。选用 ES7210 芯片作为 ADC 模数转换模块，该芯片是一款支持 4 路模拟 MIC 的 ADC 芯片，支持 8-100kHz 采样率。，可以通过 I2S 协议与主控模块进行数据传输。

同样为了搭建完善的音频系统，在硬件层面加入了 DAC 转换模块用来连接 Speaker，采用 ES8156 作为 DAC 芯片，该芯片支持 I2S 通信以及 8-96 kHz 采样率。最终得到的主控模块和音频模块的连接方式如图5-4所示。

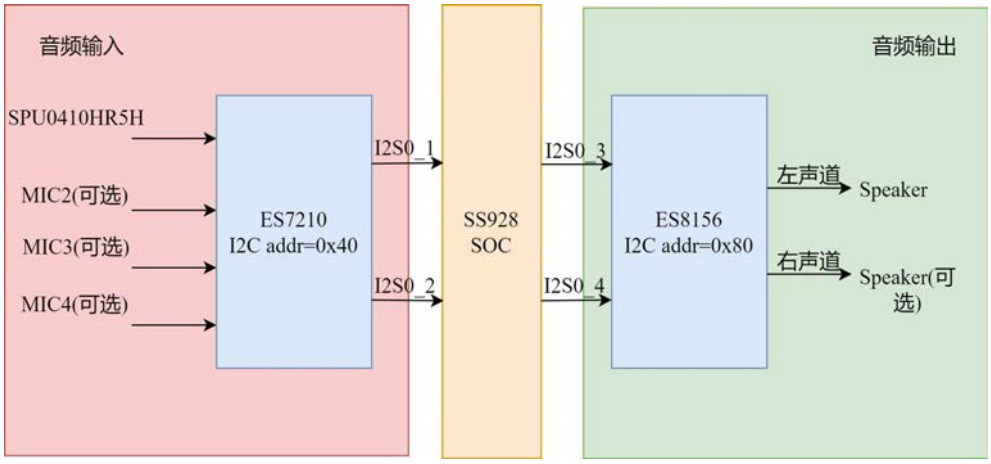


图 5-4 音频主控模块连接示意图

5.3 实时语音降噪系统软件设计

本文所设计的实时语音降噪系统软件层包含两部分：下位机软件部分和上位机软件部分。

5.3.1 下位机嵌入式端软件设计

本文所设计的实时语音降噪系统软件层包含两部分：下位机 SS928 主控芯片部分代码和上位机个人电脑 PC 端部分。

实时语音降噪系统的下位机部分的软件层次如图 5-4 所示，软件部分主要分为四层，自底向上可以分为 Boot 层、Linux 内核层、中间件层以及应用层，具体设计架构如图5-5所示。

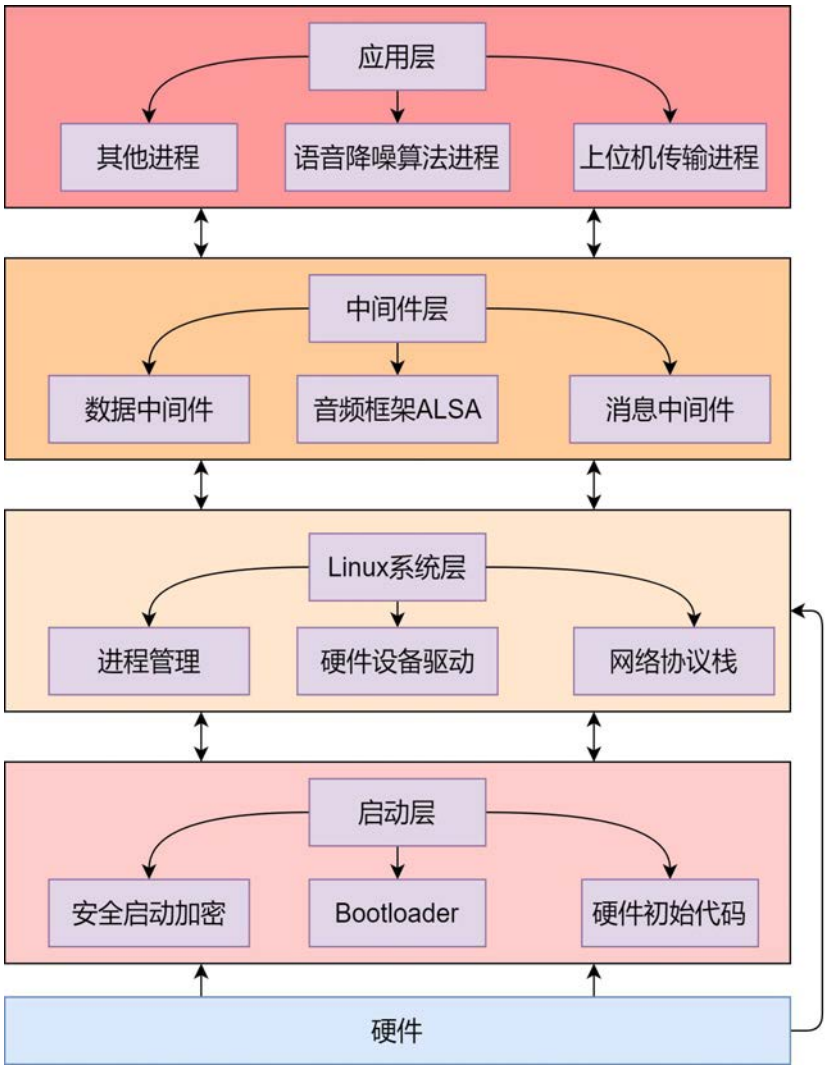


图 5-5 下位机软件架构图

第一层是启动层，在本章设计的语音降噪平台下位机 Soc 中负责在系统启动时初始化初始化 CPU、时钟、缓存、内存控制器、串口等关键硬件外设、从 Flash、ROM、eMMC、SD 卡或网络等介质加载系统加载操作系统。

第二层是 Linux 系统底层软件，包括硬件驱动、文件系统、进程调度、内存管理、用户态调用、网络协议栈等部分，其中本章设计的语音降噪系统需要通过设计硬件驱动软件来适配硬件层面接入的数模转换模块等硬件，并且通过网络协议栈与上位机建立连接以传输数据。

第三层主要为中间件部分，在本章设计中起到系统层与应用层之间的桥梁作用，通过设计中间件层可以减少上层应用代码的复杂的，减小上层代码的出错率。同时通过在本章设计的平台上移植基于 Linux 系统的三方中间件如音频框架 ALSA 可以减少开发量，避免重复工作，更轻松的控制音频流。

第四层为应用层，是本章的设计重点，主要实现了单通道语音降噪系统的具体功能以及与上位机软件的交互逻辑。基于 Linux 中间件层完成实时音频流的获取工作、实时音频流数据的降噪算法处理工作、实时音频流数据的保存工作以及与上位机的数据传输工作，执行流程如图5-6所示。

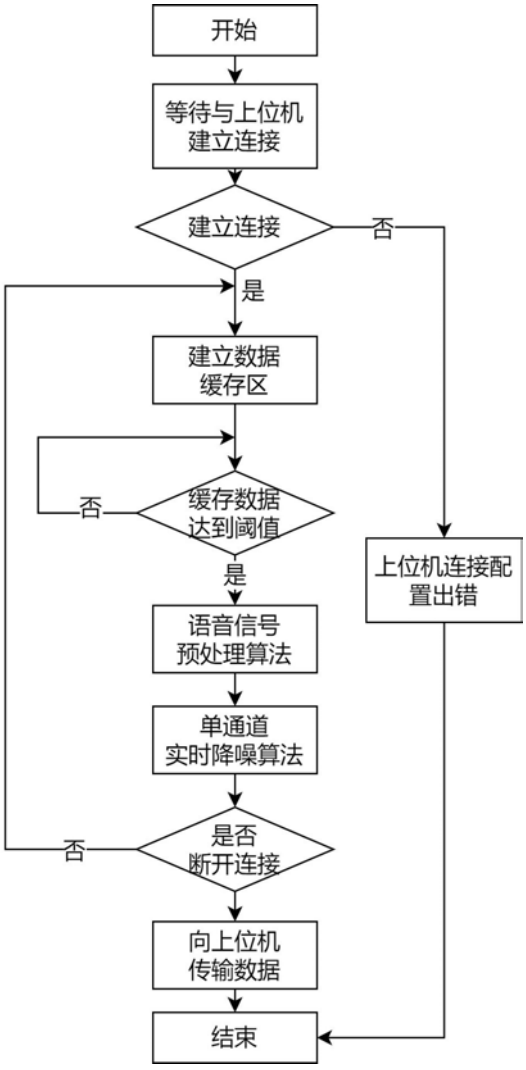


图 5-6 下位机应用层工作流程图

5.3.2 上位机 PC 端软件设计

上位机 PC 端软件设计较为简单，主要负责与下位机的数据传输工作以及降噪效果展示工作，噪声效果展示对于大于 5s 的音频信号，只展示最后 5s。工作流程图如图5-7所示。

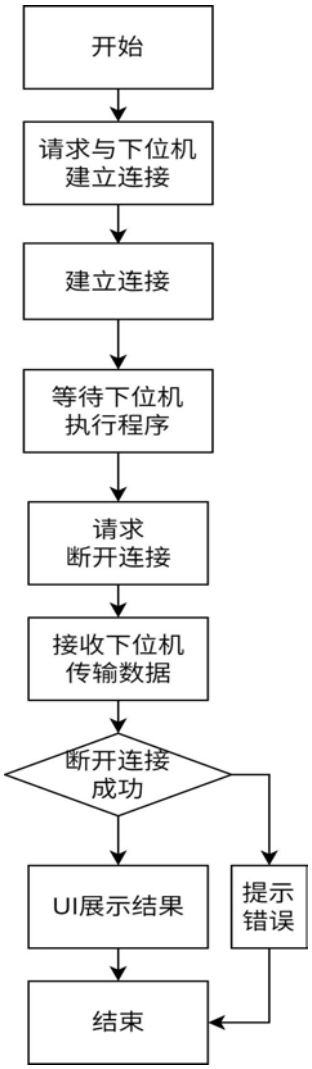


图 5-7 上位机工作流程图

5.4 嵌入式单通道语音降噪系统实时加速设计

5.4.1 深度学习模型部署

为什么充分利用 SS928 平台上的计算资源，本章在 SS928 的 NPU 计算单元上部署 OTG-STDCT-DNet 降噪算法中的两个深度学习网络，这样可以加快推理速度。首先需要将训练好的 PyTorch 模型转化为 ONNX 模型[75]，随后通过昇腾社区提供的工具将 ONNX 模型转化为 OM 模型进行模型优化与加速，从而完成模型在嵌入式设备上的部署。通过海思和昇腾提供的 SDK 在 NPU 上部署深度学习算法的流程可由图5-8表示。

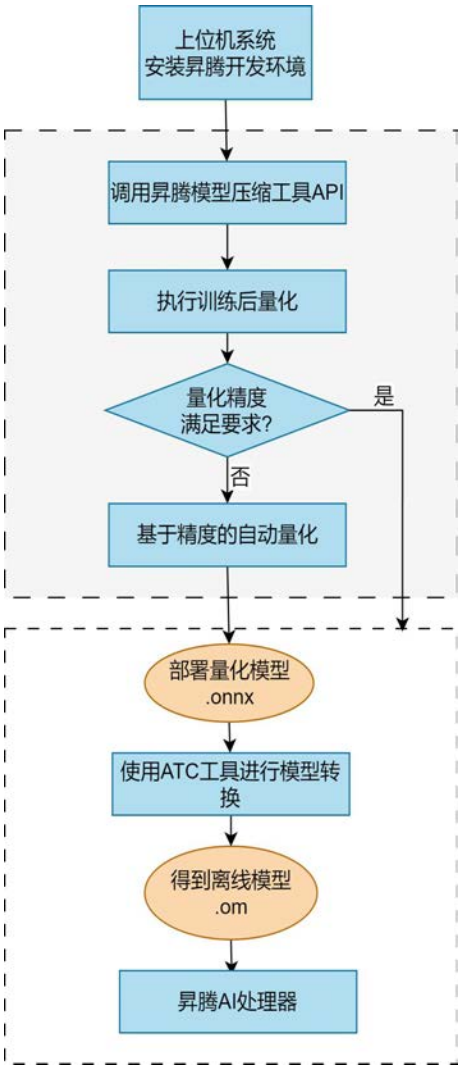


图 5-8 昇腾计算单元部署流程图

第三四章设计的深度学习网络模型所需的所有算子昇腾都提供了支持，对各种算子的支持可以从昇腾社区以及前人对昇腾处理器的应用研究中找到[76]，可以使用昇腾提供的工具进行量化部署。

5.4.2 基于 ARM 架构的信号处理算法加速

NEON 是 ARM-A 架构中的高级 SIMD（Single Instruction Multiple Data）扩展指令集，广泛用于高性能计算任务，旨在提高多媒体、信号处理、机器学习、加密等计算密集型任务的性能。它允许 CPU 在一个时钟周期内同时处理多个数据，从而实现并行计算，加速数据处理，使用 NEON 指令进行并行计算的过程可以由图5-9和图5-10表示，其本质是使用一条指令可以同时处理四个 32 位数据。理论上使用 NEON 指令对所有计算都进行优化的情况下能减少 75% 的计算时间。

图 5-9 NEON 寄存器架构图

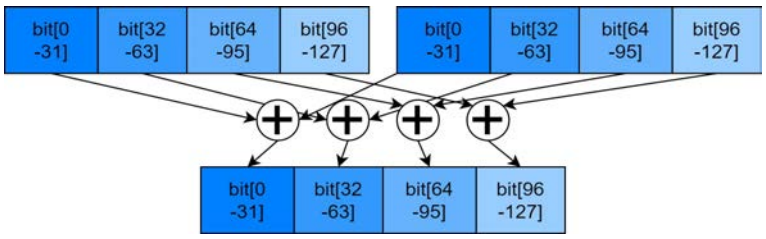


图 5-10 NEON 加法计算示意图

在语音信号预处理算法和 OMLSA 算法中有着大量的数学运算，这些数学运算包括窗函数、快速傅里叶变换、快速傅里叶逆变换、离散余弦变换、指数积分等数学运算，这些数学运算在计算机中都需要使用一定的离散计算方法和数值分析方法来模拟，如快速傅里叶变换在计算机的实现中常使用分裂基快速傅里叶变换算法 [Split Radix FFT Algorithm]。最终计算机中的模拟计算方式可以拆分为一维卷积、矩阵运算、循环运算等运算方式，这些计算方式都是可以执行并行加速的计算方式，所以使用 ARM-A 架构提供的 NEON 并行计算指令可以加速计算速度，减少算法计算时间。

使用 NEON 加速语音信号预处理算法和第三章提出的 OMLSA-TCN-GRU 模型的 OMLSA 算法计算部分，使用一段带噪语音进行测试，加速前后结果对照如表5-4所示。

表 5-4 NEON 加速效果测试结果表

	耗时 (s)
带噪语音时长	6.315
程序实际执行	7.208
预处理算法 +OMLSA 算法 (未使用 NEON 优化)	1.802
预处理算法 +OMLSA 算法 (使用 NEON 优化)	0.720

从表5-4中可以看出，使用 NEON 指令对数字信号处理算法进行优化加速之后，算法解约了 60% 的计算时间。没有达到理论上限的 75% 的优化效果，这是因为在 Linux 系统下的 C/C++ 语言编译器会在编译代码时会在合适的地方自动的转换一部分常规指令变为 NEON 指令，所以未使用 NEON 优化时的算法计算过程实际上已经启用了一部分 NEON 指令。

5.5 系统实现及效果验证

为验证第三章和第四章提出的单通道实时语音降噪算法的效果以及实时加速设计效果，在将模型部署到嵌入式设备上之后，需要对模型的泛化性能以及系统的整体性能进行验证。

(1) 实验环境

实验环境选在安静的会议室，从 VoiceBank+Demand 测试集中随机抽取 30 条音频文件使用手机外放播放，本章设计的实时语音降噪系统作为测试对象执行语音降噪程序，每一条音频文件单独执行一次测试。最后在 PC 端基于 Python 语言对下位机传输的降噪后的语音文件进行分析。测试环境如图5-11所示。

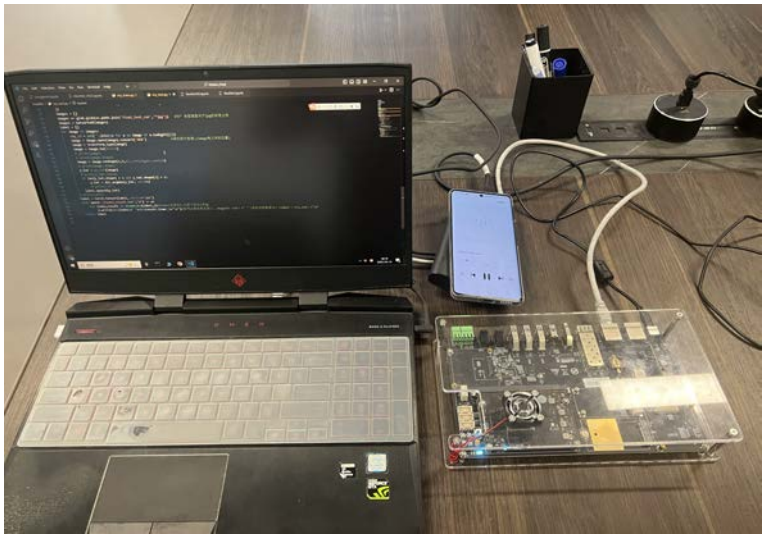


图 5-11 测试环境图

(2) 实验结果及分析

基于上述实验环境对本章搭建的嵌入式单通道实时语音降噪系统的降噪性能进行了验证，实验结果如表5-5所示。

表 5-5 嵌入式单通道实时语音降噪系统验证实验结果表

	PESQ	STOI(%)	COVL	测试语音时长 (s)	实际语音时长 (s)	算法耗时
实验一	2.53	93.6	2.83	130.527	130.527	-
实验二	2.49	93.3	2.75	130.527	130.527	51.637
实验三	2.18	81.9	2.65	130.527	160.937	40.235
实验四	2.47	93.1	2.75	130.527	158.438	61.923

其中实际语音时长由开始录音到下位机收到断开连接请求之间花费的时间，比测试音频文件时间更长是因为测试音频播放之前已经开始录音过程，播放完毕后无法及时结束降噪进程，导致会多录入一段安静环境声音。而算法执行时间的计算由在每一条 VoiceBank+Demand 测试集音频文件的语音降噪过程中的算法执行时间相加得出，每一条音频文件的算法执行时间使用 Linux 系统时间函数计算得到，在执行算法前计算当前时间 S，执行算法后再次计算当前时间 E，通过计算 E 与 S 之间的差值得到算法执行时间，Linux 系统时间函数的误差可以精确到微妙。实际语音时长和算法执行时间的计算方式可由图5-12看出，其中 S_{nt} 表示在第 n 个测试音频的第 t 个时间窗中的算法执行起始时间，E_{nt} 表示在第 n 个测试音频的第 t 个时间窗中的算法执行结束时间，S_n^* 表示第 n 次测试的起始时间，E_n^* 表示第 n 次测试的结束时间。

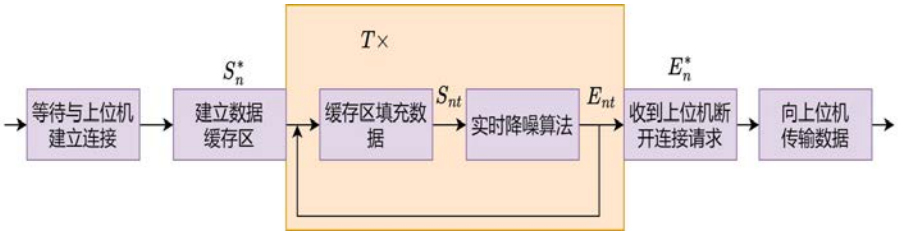


图 5-12 计算消耗时间示意图

表5-5中实验一表示在 PC 端对 OTG-STDCT-DNet 完成的仿真实验结果，实验二表示 OTG-STDCT-DNet 直接通过读取存储到嵌入式端的 30 条测试集数据的实验结果，实验三表示 OMLSA-IMCRA 算法部署在本章设计的系统上的测试结果，实验四表示 OTG-STDCT-DNet 算法部署在本章设计的系统上的测试结果。通过对比分析表中的数据可以观察到以下现象：

首先，对比实验二和实验一，可以看出当模型部署在嵌入式端之后出现了一定程度的性能下降，这是因为模型在 PC 端仿真实验中参数的精度为 32-bit 浮点数（FP32），但是部署到嵌入式设备后参数精度下降到了 8-bit（INT8），这导致了嵌入

式端模型性能不如 PC 端模型的性能。

其次，对比实验四和实验三，可以看出当模型部署在嵌入式端之后性能仍然更加优越，虽然计算时间较长，但是仍然能满足非专业级消费多媒体嵌入式设备的实时降噪要求 (对 10ms 带噪语音执行降噪算法的处理时间不超过 5ms)。

通过上位机可视化界面得到的语音降噪结果如图5-13所示。

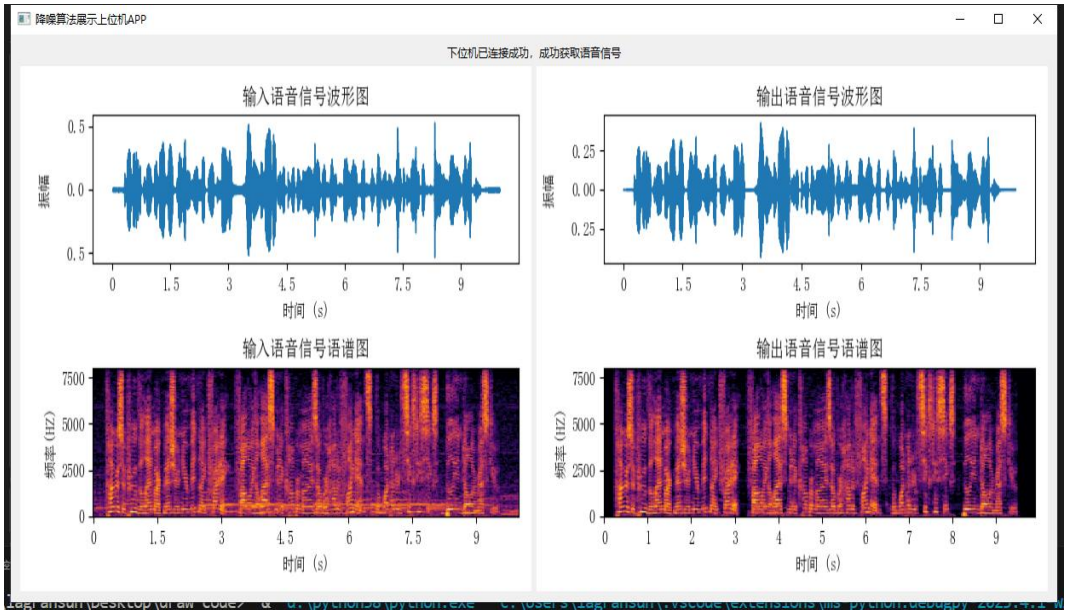


图 5-13 上位机降噪效果展示图

5.6 本章小结

本章主要研究了单通道实时嵌入式语音降噪系统的设计与实现，分别从系统硬件设计、系统软件设计介绍了系统的设计思路。首先介绍了嵌入式实时单通道语音降噪系统的总体框架，其次对系统的硬件组成结构进行了选型设计，包括硬件开发平台选择、麦克风选型、数模转换和模数转换模块选型，然后对实时语音降噪系统的软件流程进行了分析和实现，最后系统实物及测试环境进行了展示，并设置实验验证了设计的嵌入式单通道实时语音降噪系统的有效性，并对实际场景使用嵌入式单通道语音降噪系统而产生的误差来源进行了分析。

第六章 全文总结与展望

6.1 全文总结

本文主要研究了基于深度学习的单通道实时语音降噪技术，设计了嵌入式语音降噪系统，能够在对环境噪音、嵌入式系统运行时的系统设备噪声以及其余突发噪声进行降噪。全文的工作总结如下：

1. 介绍了本课题研究背景和研究意义，详细阐述了单通道实时语音降噪技术的基本原理及其在实际应用中的重要性。同时，系统梳理了国内外嵌入式语音降噪系统的研究现状，分析了当前主流方法的优势与不足，重点讨论了其在计算复杂度、实时性、降噪效果及硬件资源受限等方面的局限性。通过对现有技术瓶颈的深入分析，本研究明确了改进方向和研究重点，为后续工作的开展奠定了理论基础和技术路径。

2. 研究了基于传统方法和基于深度学习的单通道实时语音降噪系统基本原理。研究了现代传统语音降噪方法虽然在某些场景下表现良好，但在应对非稳态噪声时因为依赖于先验假设而无法对进行精确建模，导致降噪效果下降的问题，提出了一种基于 TCN-GRU 网络改进的 OMLSA-IMCRA 算法，该算法基于梅尔频谱特征学习网络，能够更好的估计带噪语音的噪声谱。在 DnoiseBank 和 Voicebank+Demand 数据集上对该方法进行了验证，实验结果证明该方法优于其他方法。

3. 研究了本文提出方法在低信噪比噪声场景的抗噪性能，研究了单级降噪模型往往无法很好地完成相对困难的任务的问题，针对这个问题，在第三章提出的模型的基础上引入了低复杂度的 STDCT-DNet(Short Time Discrete Cosine Transform Denoise Net) 网络模型作为后处理模块来进一步抑制第三章模型输出中的残余噪声，提高模型的抗噪性能和降噪后语音的主观质量。通过对所提出算法与现有的多阶段语音降噪算法进行仿真实验和比较分析，验证了本章提出的噪声后处理模块在实时语音降噪方面的有效性。

4. 搭建了嵌入式单通道实时语音降噪系统，并使用该系统证明了本文提出的方法的有效性，和实际的应用价值。本文搭建了音频采集和模数转换模块采集并储存了可向 SOC 传输的数字语音信号。并将本文将所提出的语音降噪模型部署在下位机嵌入式 SS928 设备上，并开发了上位机可视化软件直观显示下位机系统输出的降噪语音结果。实验结果表明，在本文搭建的嵌入式单通道实时语音降噪系统具有很好的降噪性能，同时能够满足实时性的要求。

6.2 后续工作展望

本文基于深度学习对嵌入式单通道实时语音降噪系统展开了研究，虽然取得了一定的成果，但仍然存在不少不足之处：

1. 本文所设计的嵌入式单通道实时语音降噪系统虽然能满足实时性的要求，但是在特别复杂的噪声环境下效果仍然有待改进，未来在可以通过进一步进一步改进模型结构，优化注意力机制的实现方式提高语音降噪系统的降噪性能。

2. 本文所设计的 OTG-STDCT-DNet 实时降噪模型在嵌入式端的实时性层面和降噪能力层面仍有进步空间，未来，可以基于优化网络结构、减少参数冗余、降低计算复杂度的轻量化模型的设计思路，采用模型轻量化方法（如剪枝、量化、知识蒸馏等）来开发既能保持高准确度，又能满足高推理速度的语音降噪系统。

3. 本文的系统验证中使用的是手机播放的测试数据集，但是在真实会议场景或户外场景下还有其他人说话声等与说话人声音相似的噪声的干扰，实际情况更加复杂，如何获取此类数据集，以及如何在更复杂的应用场景下提升模型的识别性能和泛化能力，仍然是当前需要深入研究的重要问题。

参考文献

[1] Yu D, Deng L. Automatic speech recognition: A deep learning approach[M]. Springer, 2014.

[2] Gay S L, Benesty J. Acoustic signal processing for telecommunication[M]. Springer US, 2000.

[3] Tasell D J V. Hearing loss, speech, and hearing aids[J]. Journal of Speech and Hearing Research, 1993, 36(2): 228-244.

[4] Benesty J, Chen J, Huang Y, et al. On microphone-array beamforming from a mimo acoustic signal processing perspective[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 15(3): 1053-1065.

[5] Taherian H, Wang Z, Chang J, et al. Robust speaker recognition based on single-channel and multi-channel speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1293-1302.

[6] 田斌, 田红心, 易克初. 语音识别中的加性噪声补偿研究 [J]. 西安电子科技大学学报, 2001, 03: 292-295.

[7] Boll S F. Suppression of acoustic noise in speech using spectral subtraction[J]. IEEE Transactions on acoustics, speech, and signal processing, 1979, 27(2): 113-120.

[8] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise[C]. ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, 1979: 208-211.

[9] Berouti M, Schwartz R, Makhoul J. Enhancement of speech corrupted by acoustic noise[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1979: 208-211.

[10] Lockwood P, Boudy J. Experiments with a nonlinear spectral subtractor (nss), hidden markov models and the projection for robust speech recognition in cars[J]. Speech Communication, 1992, 11(2-3): 215-228.

[11] Lim J S, Oppenheim A V. Enhancement and bandwidth compression of noisy speech[C]. Proceedings of the IEEE, 1979: 1586-1604.

[12] Scalart P, Filho J V. Speech enhancement based on a priori signal to noise estimation[C]. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1996: 629-632.

[13] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator[J]. IEEE Transactions on acoustics, speech, and signal processing, 1984, 32(6): 1109-1121.

[14] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator[J]. IEEE transactions on acoustics, speech, and signal processing, 1985, 33(2): 443-445.

[15] Cohen I, Berdugo B. Speech enhancement for non-stationary noise environments[J]. Signal processing, 2001, 81(11): 2403-2418.

[16] Cohen I. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator[J]. IEEE Signal processing letters, 2002, 9(4): 113-116.

[17] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection[J]. IEEE signal processing letters, 1999, 6(1): 1-3.

[18] Kum J M, Park Y S, Chang J H. Improved minima controlled recursive averaging technique using conditional maximum a posteriori criterion for speech enhancement[J]. Digital Signal Processing, 2010, 20(6): 1572-1578.

[19] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics[J]. IEEE Transactions on speech and audio processing, 2001, 9(5): 504-512.

[20] Cohen I, Berdugo B. Noise estimation by minima controlled recursive averaging for robust speech enhancement[J]. IEEE signal processing letters, 2002, 9(1): 12-15.

[21] Cohen I. Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging[J]. IEEE Transactions on speech and audio processing, 2003, 11(5): 466-475.

[22] O'Shaughnessy D. Speech enhancement—a review of modern methods[J]. IEEE Transactions on Human-Machine Systems, 2024, 54(1): 110-120.

[23] Williamson D S, Wang Y, Wang D. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 481-490.

[24] Pascual S, Bonafonte A, Serra J. Segan: Speech enhancement generative adversarial network[J]. arXiv preprint arXiv:1703.09452, 2017.

[25] Stoller D, Ewert S, Dixon S. Wave-u-net: A multi-scale neural network for end-to-end audio source separation[C]. Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR), 2018: 1-8.

[26] Defossez A, Synnaeve G, Adi Y. Real time speech enhancement in the waveform domain[C]. Proceedings of Interspeech 2020, 2020: 1-5.

[27] Tzinis E, Wang Z, Smaragdis P. Sudo rm -rf: Efficient networks for universal audio source separation[C]. Proceedings of IEEE Machine Learning for Signal Processing (MLSP) 2020, 2020: 1-6.

[28] Kavalerov I, Wisdom S, Erdogan H, et al. Universal sound separation[C]. Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2019: 1-5.

[29] Mirsamadi S, Tashev I. Causal speech enhancement combining data-driven learning and suppression rule estimation[C]. Proceeding of the Conference of the International Speech Communication Association, 2016: 2870-2874.

[30] Valin J-M. A hybrid dsp/deep learning approach to real-time full-band speech enhancement[C]. 2018 IEEE 20th international workshop on multimedia signal processing (MMSP), 2018: 1-5.

[31] Yuan J, Bao C. Cyclegan-based speech enhancement for the unpaired training data[C]. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2019: 878-883.

[32] Nicolson A, Paliwal K K. Deep learning for minimum mean-square error approaches to speech enhancement[J]. Speech Communication, 2019, 111: 44-55.

[33] Takahashi N, Goswami N, Mitsufuji Y. Mmdenselstm: An efficient combination of convolutional and recurrent neural networks for audio source separation[C]. 2018 16th International workshop on acoustic signal enhancement (IWAENC), 2018: 106-110.

[34] Pandey A, Wang D. Dense cnn with self-attention for time-domain speech enhancement[J]. IEEE/ACM transactions on audio, speech, and language processing, 2021, 29: 1270-1279.

[35] Schroter H, Maier A, Escalante-b A, et al. Deepfilternet2: Towards real-time speech enhancement on embedded devices for full-band audio[C]. 2022 International Workshop on Acoustic Signal Enhancement (IWAENC), Piscataway, NJ, USA, 2022//: 5pp.-.

[36] 侯正风. 基于二进制小波变换和维纳滤波的语音降噪研究 [J]. 信号处理, 2002, 03: 257-260.

[37] 蔡斌, 郭英, 李宏伟等. 一种改进型 mmse 语音增强方法 [J]. 信号处理, 2004, 20(1): 68-72.

[38] 刘凤增, 李国辉, 李博. Om-lsa 和小波阈值去噪结合的语音增强 [J]. 计算机科学与探索, 2011, 5(6): 547-552.

[39] 张建伟, 陶亮, 周健等. 基于改进谱平滑策略的 imcra 算法及其语音增强 [J]. 计算机工程与应用, 2017, 53(1): 153-157.

[40] Wang Y, Xie J, Wang D. 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: 3123-3132.

[41] Xu R, Wu R, Ishiwaka Y, et al. Listening to sounds of silence for speech denoising[C]. Advances in Neural Information Processing Systems (NeurIPS) 2020, 2020: 1-12.

[42] Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

[43] 贾海蓉, 王栋, 郭欣. 基于 dnn 的子空间语音增强算法 [J]. 太原理工大学学报, 2016, 47(5): 647-650+679.

[44] 阴法明, 唐於烽. 基于深度置信网络的语音增强算法 [J]. 电子器件, 2018, 41(5): 1325-1329.

[45] Fu S-W, Liao C-F, Hsieh T-A, et al. Boosting objective scores of speech enhancement model through metricgan post-processing[J]. arXiv preprint arXiv:1910.05378, 2019.

[46] Kim J, Hahn M. Speech enhancement using a two-stage network for an efficient boosting strategy[J]. IEEE Signal Processing Letters, 2019, 26(5): 770-774.

[47] Fu S-W, Yu C-H, Hsieh T-A, et al. Metricgan+: An improved version of metricgan for speech enhancement[J]. arXiv preprint arXiv:2104.05358, 2021.

[48] Li A, Liu W, Zheng C, et al. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021//, 29: 1829-43.

[49] Wang Y, Narayanan A, Wang D. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

[50] Xu Y, Du J, Dai L-R, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.

[51] Kolbæk M, Tan Z-H, Jensen J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2017, 25(1): 153-167.

[52] Paliwal K K, Wójcicki K K, Shannon B J. The unimportance of phase in speech enhancement[J]. Speech Communication, 2011, 53(5): 634-641.

[53] Paliwal K K, Wójcicki K K, Shannon B J. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(5): 634-641.

[54] Bai S, Kolter J Z, Koltun V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint arXiv:1803.01271, 2018.

[55] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]. Advances in Neural Information Processing Systems, 2017: 5998-6008.

[56] Zhang Q, Zhu H, Song Q, et al. Ripple sparse self-attention for monaural speech enhancement[J]. arXiv preprint arXiv:2305.08541, 2023.

[57] Chen H, Zhang P, Yan Y. An audio scene classification framework with embedded filters and a dct-based temporal module[C]. ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Proceedings, Piscataway, NJ, USA, 2019//: 835-9.

[58] Gray R M, Buzo A, Gray A H J, et al. Distortion measures for speech processing[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28(4): 367-376.

[59] Yao R, Zeng Z, Zhu P. A priori snr estimation and noise estimation for speech enhancement[J]. EURASIP Journal on Advances in Signal Processing, 2016, 2016(1): 101.

[60] Du J, Na X, Lu X, et al. AISHELL-2: Transforming mandarin ASR research into industrial scale[J]. arXiv preprint arXiv:1808.10583, 2018.

[61] Ephrat A, Mosseri I, Lang O, et al. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation[J]. ACM Transactions on Graphics (TOG), 2018, 37(4): 112.

[62] Varga A, Steeneken H J M. Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems[J]. Speech Communication, 1993, 12(3): 247-251.

[63] Wichern G, Antognini J, Flynn M, et al. Wham!: Extending speech separation to noisy environments[J]. arXiv preprint arXiv:1907.01160, 2019.

[64] Valentini-Botinhao C, Wang X, Takaki S, et al. Investigating RNN-based speech enhancement methods for noise-robust text-to-speech[C]. Proc. 9th ISCA Speech Synthesis Workshop, 2016: 146-152.

[65] 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 codec[C]. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), 2001: 749-752.

[66] Taal C H, Hendriks R C, Heusdens R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125-2136.

[67] Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2008, 16(1): 229-238.

[68] Hu Y, Liu Y, Lv S, et al. Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement[J]. arXiv preprint arXiv:2008.00264, 2020.

[69] Braun S, Gamper H, Reddy C K A, et al. Towards efficient models for real-time deep noise suppression[C]. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021: 656-660.

[70] Tan K, Wang D. A convolutional recurrent neural network for real-time speech enhancement[C]. Proceedings of Interspeech 2018, Hyderabad, India, 2018: 3229-3233.

[71] Paliwal K, Wójcicki K, Shannon B. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(4): 465-494.

[72] Wang K, Huang H, Hu Y, et al. End-to-end speech separation using orthogonal representation in complex and real time-frequency domain[C]. Proceedings of Interspeech 2021, 2021.

[73] Wu H, Tan K, Xu B, et al. Rethinking complex-valued deep neural networks for monaural speech enhancement[J]. arXiv preprint arXiv:2301.04320, 2023.

[74] Dang F, Chen H, Hu Q, et al. First coarse, fine afterward: A lightweight two-stage complex approach for monaural speech enhancement[J]. Speech Communication, 2022, 144: 25-37.

[75] Guirguis K, Schorn C, Guntoro A, et al. SELD-TCN: Sound event localization & detection via temporal convolutional networks[C]. Proceedings of the 28th European Signal Processing Conference (EUSIPCO), 2020: 1-5.

[76] 李晓明, 王晓东, 李宏毅, et al. 华为昇腾神经网络加速器性能评测与优化 [J]. 计算机学报, 2022, 45(8): 1618-1637.