学校代码：10270　　学号：202502895

上海师范大学

硕士专业学位论文

基于深度神经网络的单通道语音增强方法研究

学　　院：信息与机电工程学院

专业学位类别：工程硕士

专 业 领 域：电子信息

研 究 生 姓 名：戈晓枫

指 导 教 师：龙艳花、易辉跃

完 成 日 期：2023 年 5 月 28 日

摘要

语音增强（Speech enhancement，SE），是目前智能语音领域的研究热点之一，其是实时通信，智能家居，可穿戴医疗设备等应用领域中的关键性技术。随着深度学习技术的创新和发展，基于深度神经网络的语音增强技术由于其卓越的性能，逐渐取代传统的基于信号处理的语音增强技术，成为该领域研究者们的研究重点，同时被广泛地应用。

虽然语音增强技术近年来有了重大的进展与明显的进步，然而以下问题仍在很大程度上限制了语音增强系统的性能和其在现实场景中的应用：（1）在实时通信等许多应用任务中，对语音增强系统的实时性有很高的要求，这使得系统的参数量和延迟有了很严苛的限制，如何兼顾实时性和语音增强性能，设计出一个低延迟，高性能的语音增强系统是目前的挑战之一；（2）过度抑制是语音增强领域常见的现象，这会给语音带来不可逆的失真，严重的过度抑制会使得语音的可懂度下降，这显然与语音增强的初衷相悖，如何改善过度抑制现象也是目前语音增强领域的热点之一；（3）无论是传统的语音增强方法还是基于深度神经网络的语音增强技术，都无法准确地去除信号中可能包含的干扰说话人语音，这限制了语音增强系统在现实生活场景中的应用。如何做到只保留目标说话人语音，去除干扰说话人语音和噪声，实现个性化的语音增强（Personalized speech enhancement, PSE）近年来逐渐受到了关注。然而关于该任务的研究目前仍相对较少，其中存在的问题与挑战也仍待发现与解决。本论文针对以上语音增强领域中的难点，进行深入研究，主要包括以下创新点：

(1) 针对单通道实时语音增强任务，复现 PercepNet 基线系统，并从模型和声学特征入手，提出一种相位感知的结构，在不影响系统实时性的前提下，提升语音增强性能。

(2) 基于 PercepNet 系统，通过提出一种新的多任务学习策略和基于信噪比估计的后处理技术，改善单通道实时语音增强任务中的过度抑制问题。

(3) 针对个性化语音增强任务，以 sDPCCN 为基线系统，提出一种动态声学补偿方法来改善测试语音和注册语音声学环境不匹配的问题，并通过自适应焦点训练机制提升困难样本性能，提高了系统性能与鲁棒性。

本文使用多个已开源数据集进行实验，其中选用 McGill TSP speech database，NTT Multi-Lingual Speech Database for Telephonometry 和 VCTK 数据集验证所提出的相位感知结构，多任务学习策略以及基于信噪比的后处理技术的有效性；选用 4th Deep Noise Suppression (DNS) Challenge track2 数据集验证

动态声学补偿方法和自适应焦点训练机制在单通道个性化语音增强任务中改善声学环境不匹配和困难样本问题的有效性。实验结果表明，与基线系统相比，本文所提出的创新点均可较大程度地提升系统性能与系统鲁棒性，为未来单通道实时语音增强技术以及单通道个性化语音增强技术的发展与落地提供了重要参考。

关键词：语音增强，相位感知，多任务学习，动态声学补偿，自适应焦点训练

Abstract

Speech enhancement (SE) is one of the research hotspots in the field of intelligent speech at present. It is a key technology in the application fields of real-time communication, intelligent furniture, wearable medical devices, etc. With the innovation and development of deep learning technology, speech enhancement technology based on deep neural network has gradually replaced the traditional speech enhancement technology based on signal processing due to its excellent performance, becoming the research focus of researchers in this field and being widely used.

Although speech enhancement technology has made significant and obvious progress in recent years, the following problems still limit the performance of speech enhancement system and its application in real scenes to a large extent: (1) In many application tasks such as real-time communication, there is a strict requirement for real-time performance of speech enhancement system, which makes the system's parameters and computation amount severely limited. How to design a speech enhancement system with low delay and high performance is one of the current challenges; (2) Over attenuation is a common phenomenon in the field of speech enhancement, which will bring irreversible distortion to speech. Severe over attenuation will reduce the intelligibility of speech, which is obviously contrary to the original intention of speech enhancement. How to solve the problem of over attenuation is also one of the hotspots in the field of speech enhancement at present; (3) Neither the traditional speech enhancement methods nor the speech enhancement technology based on deep neural network can accurately remove the interference speaker's speech that may be contained in the signal, which limits the application of the speech enhancement system in real life scenes. How to realize personalized speech enhancement (PSE), retaining only the target speaker's speech, removing the interference speaker's speech and noise has gradually attracted attention in recent years. However, the research on this task is still relatively few, and the problems and challenges still need to be found and solved. This paper focuses on the above difficulties in the field of speech enhancement and conducts in-depth research, mainly including the following innovations:

(1) For the single channel real-time speech enhancement task, the baseline system PercepNet is reproduced, and a phase-aware structure including the deep neural network model and acoustic features is proposed to improve the speech enhancement

performance without affecting the real-time performance of the system.

(2) Based on PercepNet system, a creative multi-task learning strategy and a post-processing technology based on signal-to-noise ratio (SNR) estimation are proposed to alleviate the over attenuation problem in single channel real-time speech enhancement task.

(3) For the personalized speech enhancement task, a dynamic acoustic compensation is proposed to alleviate the acoustic environment mismatch between the test speech and the enrollment speech based on the baseline system sDPCCN. The adaptive focal training mechanism is used to improve the performance of hard samples and improve the system performance and robustness.

In this paper, several open source datasets are used for experiments, among which McGill TSP speech database, NTT Multi Lingual Speech Database for Telephony and VCTK datasets are selected to verify the effectiveness of the proposed phase-aware structure, multi-task learning strategy and post-processing technology based on signal-to-noise ratio estimation; The 4th Deep Noise Suppression (DNS) Challenge track2 dataset was selected to verify the effectiveness of dynamic acoustic compensation mechanism and adaptive focal training to solve the acoustic environment mismatch and hard sample problems in personalized speech enhancement task. The experimental results show that compared with the baseline system, the innovation proposed in this paper can greatly improve the speech enhancement performance and system robustness. It provide an important reference for the development and implementation of single channel real-time speech enhancement technology and single channel personalized speech enhancement technology in the future.

Key Words: Speech enhancement, Phase-aware, Multi-task learning, Dynamic acoustic compensation, Adaptive focal training

目　录

第 1 章 绪论

1.1 研究背景及意义

语言是实现人与人之间便捷交流的信息工具，而语音信号是实现这种工具功能的介质。然而在现实生活中，语音信号经常会被各种类型的背景噪声或其他干扰说话人语音，甚至混响、回声等干扰信号所污染。这样的语音，由于添加了干扰成分，一方面会导致人类主观听觉感受质量的下降，另一方面严重影响目标语音内容的可懂度。

语音增强的主要目的就是消除语音信号中可能存在的干扰分量，恢复出干净的目标语音信号从而提高语音的质量和可懂度。首先语音增强在通信领域有重要意义，普通用户的电话通信或视频通信，通常都含有现实生活场景中的各种噪声或干扰说话人语音，此时就需要语音增强技术消除干扰分量，从而保证说话者所说的内容能被远端的听者准确无误地理解。另外，基于语音识别的输入法和语义理解被越来越广泛地应用于智能手机[1]，车载设备[2]，智能家具[3]等电子设备上，这些设备通常都在复杂的声学环境中被使用，语音增强技术可以提升语音识别的准确性，使得电子设备更准确地接受用户的指令，提升用户的使用感受。

由于深度学习理论实现以及硬件技术的逐渐成熟，有监督的基于深度学习的语音增强方法[4][5]已成为目前主流的研究方向，利用既有的语音数据和噪声数据，使系统学习语音和噪声的特性，以此指导将干净语音从带噪语音信号中分离出来。虽然传统的语音增强方法，在处理非平稳噪声方面存在局限性，但很多研究者已经使用传统信号处理方法和深度学习技术相结合的混合模型[6][7]，并取得了优秀的效果。总体而言，语音增强研究已经取得了一些进展，但其仍旧是一个新兴的未成熟的研究课题，依然存在许多问题与挑战，比如以下几个方面：（1）基于深度神经网络的语音增强系统虽然性能优秀，但网络参数量大，计算复杂，无法应用于实时通信场景，因此如何权衡兼顾系统实时性和语音增强性能是语音增强领域的一项重要挑战。（2）语音增强经常会有过度抑制的情况发生[8]，即过度估计噪声分量，使得部分语音分量也被去除，使语音产生失真，语音可懂度随之明显降低，因此如何改善过度抑制情况就尤为重要。（3）在现实生活场景中，语音信号中的干扰分量可能同时包含噪声和除目标说话人以外的其他干扰说话人语音，而目前已有的研究中，无论是基于深度神经网络的语音增强系统还是传统的语音增强方法，都无法将干扰说话人语音准确去除。如何实现个性化语音增强，即使得语音增强系统能同时去除干扰说话人语音和

噪声，近年来受到了广泛关注。

1.2 国内外研究现状

随着通信质量引起人们的重视和语音识别技术的广泛应用，国内外专家、学者对于语音增强的研究逐渐深入，同时取得一定的进展。根据语音采集设备的通道数量，语音增强可分为单通道语音增强与多通道语音增强，本文将重点研究单通道的语音增强方法，下面对近年来传统技术和深度学习技术在单通道语音增强领域的研究现状做出介绍。

传统的单通道语音增强算法研究至今已有几十年的历史，传统的单通道语音增强方法基本上都是无监督的，包括谱减法[9]，维纳滤波[10]，幅度谱估计[11]，信号子空间方法等。其中最早被提出的是谱减法，谱减法的原理是将带噪语音信号的前若干帧信号的平均能量作为估计的噪声谱，而后从带噪语音谱中减去估计的噪声谱，恢复出干净的语音信号。谱减法是基于噪声信号是平稳的假设，即噪声信号的特性不会随时间变化而变化，但这样的假设在现实中显然不总是成立的。当语音信号中的噪声是非平稳噪声时，谱减法估计的噪声谱会出现过估计或欠估计，若噪声谱过估计，则会造成语音失真，若噪声谱欠估计，则会导致有噪声残留，且容易产生“音乐噪声”。随后，维纳滤波法被提出，维纳滤波法的残留噪声是一种类似于白噪声的信号，这让增强后的语音听上去比较舒适，但维纳滤波是对信号平稳条件下的最小均方误差估计，因此在处理含有非平稳噪声的带噪语音信号时效果也并不理想。信号子空间法，是将带噪语音信号分解为语音子空间和噪声子空间，将噪声子空间去除，进而得到干净语音，但也有和其他传统方法相同的问题存在，即在处理非平稳噪声时表现不佳。总体而言，由于传统信号处理的语音增强方法都是基于对噪声信号或语音信号的统计特性估计，而当信号非平稳时，估计得到的统计特性便不再准确，传统的语音增强方法的性能也因此受到了极大限制。

随着深度学习技术热潮的兴起，神经网络方法开始应用于单通道语音增强，包括卷积神经网络（Convolutional Neural Networks，CNN）[12]，循环神经网（Recurrent Neural Network，RNN）[13]，残差网络（Residual Networks，ResNet）[14]等。神经网络结构克服许多传统方法的缺点：基于深度神经网络的语音增强方法不再局限于处理平稳噪声，通过大数据学习不同类型噪声的特性，这种语音增强方法可以对含有非平稳噪声的语音信号进行去噪。近年来基于卷积神经网络和循环神经网络结合的语音增强系统被广泛应用于该领域，如 Strake.M 等人提出的 FCRN 结构[15]，相较之前的单一网络结构系统，FCRN 有了很大的性能提升。基于神经网络的语音增强方法的关键除了网络结构的设计，神经网络

输入的声学特征与训练目标的设置也至关重要。主流语音增强算法所使用的声学特征包含原始时域波形，STFT 谱，幅度谱，梅尔倒谱系数等，目前关于哪一种声学特征或哪几种声学特征的组合最适合语音增强任务的研究与讨论依旧是该领域的热点之一。目前的语音增强算法按照训练目标分类主要可分为基于时频掩蔽（Time-Frequency Mask）的算法和基于特征映射（Feature Mapping）的算法。基于时频掩蔽的算法通过神经网络学习代表语音和噪声之间相互关系的掩蔽值，其中最常用的掩蔽包括理想二值掩蔽（Ideal Binary Mask, IBM）[16]，理想比值掩蔽（Ideal Ratio Mask, IRM）[17][18]，复数比值掩蔽（Complex Ratio Mask, CRM）[19]。基于特征映射的算法通过神经网络学习带噪语音声学特征与干净语音声学特征之间的非线性关系，最常用的特征映射是幅度谱映射[20]与 STFT 时频谱映射[21]。

由于在实时通信任务中，语音增强系统需要在低延迟的情况下预测干净语音信号，如何在保持语音增强性能的基础上降低系统参数量和计算量成为语音增强领域最大的难点，因此设计高性能且实时的单通道语音增强方法成为目前的研究热点。近年来，越来越多的研究者提出了实时的语音增强系统，如 L.Xie 等人提出的 DCCRN[22]和 S-DCCRN[62]系统，X.Le 等人提出的 DPCRN 系统[24]等通过使用 U-Net[25]结构的神经网络，并严格限制系统参数量和时延，实现实时语音增强。J-M.Valin 等人在 2018 年提出的 RNNoise 系统[26]和 2020 年基于此改进的 PercepNet 系统[27]，以深度神经网络为基础，将语音信号划分频带处理，并创新性地将深度神经网络方法结合传统信号处理方法，在系统计算量和时延极低的情况下完成高性能的语音增强，获得了研究者们的广泛关注。本文后续章节将对 RNNoise 和 PercepNet 的原理进行详细介绍，并选择其作为研究单通道实时语音增强的基线系统，在此基础上展开后续研究。

在现实生活场景中，目标说话人的语音可能不只受到噪声干扰，其他说话人的语音也会对语音整体听觉感受和目标语音的可懂度造成严重影响。因此近年来，个性化的语音增强，即同时去除干扰说话人语音和干扰噪声，增强目标说话人语音的任务，由于其更符合现实生活场景的应用需要，也逐渐受到了语音增强领域研究者们的广泛关注。由于传统的语音增强方法无法区分不同说话人之间的声学特性差异，因此目前有关个性化语音增强的研究均基于深度神经网络[28][29]。个性化语音增强系统通常需要额外的目标说话人的一条或多条注册语音，如 Eskimez 提出的 pDCCRN 系统[30]，Q.Wang 提出的 Voicefilter 系统[31][32]等，均在深度神经网络中引入带有目标说话人信息的嵌入向量，指导网络能区分不同说话人，从而增强目标说话人语音并去除其他干扰分量。J.Han 等人提出的 sDPCCN[33]在目标说话人提取任务中取得了优秀的性能，本课题将 sDPCCN

应用到个性化语音增强任务，作为后续研究的基线系统，在此基础上对单通道个性化语音增强任务中目前仍存在的问题展开研究。

综上所述，随着大数据和深度学习技术的不断发展，基于深度神经网络的单通道语音增强方法逐渐取代了传统的方法，其中数据，声学特征，网络结构扮演着至关重要的角色。语音增强在现实场景中的应用越来越广泛，对于语音增强系统的要求也变得更多样化，其中实时语音增强和个性化语音增强由于应用的需要逐渐成为单通道语音增强领域的重要研究方向，本文后续章节也将对上述两个任务进行详细阐述和着重研究。

1.3 本文研究内容及组织结构

本文主要研究基于深度神经网络的单通道实时语音增强与单通道个性化语音增强。针对单通道实时语音增强任务，首先搭建基线系统 PercepNet，并从网络结构与声学特征角度，提出一种相位感知结构，使系统能更好地学习语音的相位信息，并恢复出干净语音的相位，进一步提升语音增强性能。而后本文提出多任务学习策略和基于信噪比估计的后处理方法，在很大程度上改善了基线系统以及其他语音增强系统中常见的过度抑制问题，尤其在高信噪比情况下明显提升了语音质量与可懂度。针对单通道个性化语音增强任务，本文基于基线系统 sDPCCN，提出了动态声学补偿方法与自适应焦点训练，前者改善测试语音与目标说话人注册语音之前的声学环境不匹配问题，后者提升了数据集中困难样本的性能，实现了鲁棒性的个性化语音增强。

本文的组织结构如下：

第一章：介绍语音增强的研究背景与意义，简述单通道语音增强的发展历史、目前已有的技术与研究热点，分析目前仍存在的问题与挑战，并给出本文的研究内容与组织架构；

第二章：介绍 RNNoise 和基于此改进的 PercepNet，复现 PercepNet 作为单通道实时语音增强任务研究的基线系统，并完成相关实验和结果分析。

第三章：介绍基于 PercepNet 系统所提出的相位感知结构，详细阐述其声学特征，实虚部增益与网络结构改进，并完成相关实验和对应结果分析。

第四章：介绍过度抑制现象，详细阐述为改善此现象提出的多任务学习策略与基于信噪比估计的后处理方法，并完成相关实验和对应结果分析。

第五章：介绍个性化语音增强原理与基线系统 sDPCCN，分析声学环境不匹配问题与困难样本问题，详细阐述为改善上述问题所分别提出的动态升学补偿技术和自适应焦点训练机制，并完成相关实验和对应结果分析。

第六章：对本文中所提及的技术以及做出的创新和改进进行总结与分析，

并对后续的基于深度神经网络的单通道语音增强研究进行展望。

第 2 章 单通道实时语音增强基线系统

为满足实时通信的要求，有关单通道的实时语音增强的研究越来越多，其中 RNNoise 和 PercepNet 语音增强算法由于其出色的性能得到了广泛使用，本文也将 PercepNet 作为研究单通道实时语音增强的基线系统。

本章节的组织结构如下：第 2.1 节介绍单通道语音增强的原理；第 2.2 节阐述 RNNoise 系统的原理；第 2.3 节介绍基于 RNNoise 系统改进的 PercepNet 系统；第 2.4 节主要介绍在 VCTK 测试集[34]上有关复现 PercepNet 系统的相关实验。

2.1 单通道语音增强原理

语音增强的目的是去除信号中的干扰分量，增强目标语音，从而提高语音质量和可懂度。按噪声和语音的相互关系将噪声分类，可分为加性噪声和乘性噪声。由于现实场景中的噪声多为加性噪声[35][36]，即环境噪声，因此语音增强领域以及本文着重研究对加性噪声的处理方法，带有加性噪声的语音信号的时域表达式如下：

$$y(k) = x(k) + n(k) \tag{2-1}$$

其中 y 表示带噪语音信号，x 表示干净语音信号，n 表示加性噪声，k 表示时间索引。带噪语音的频域表示通过对式 $(2-1)$ 两边进行短时傅里叶变换得到：

$$Y(f,t) = X(f,t) + N(f,t) \tag{2-2}$$

其中 $Y(f,t)$，$X(f,t)$，$N(f,t)$ 分别表示第 t 帧，第 f 个频点的带噪语音信号，干净语音信号和噪声信号。本文研究的单通道语音增强处理的语音信号是单通道设备录制的，单通道语音增强只通过带噪语音信号的时域或频域信息恢复出干净语音，而不考虑语音的空域信息，即声源信息。

图 2-1 展示了一个基本的基于深层神经网络的语音增强系统。基于深度学习的单通道语音增强系统的建模主要分为四个步骤：数据合成，特征提取，深度神经网络增强，信号处理输出预测的干净语音。首先将干净语音和噪声数据合成带噪语音数据，提取不同的声学特征，而后将特征输入深层神经网络预测干净语音的声学特征，最后通过信号处理将增强后的声学特征转换为语音信号，并与实际的干净语音计算误差评估性能。

图 2-1 基于深层神经网络的语音增强系统建模流程

搭建一个基于深层神经网络的语音增强系统主要包含两个阶段：训练阶段和测试阶段。

训练阶段：基于深层神经网络的语音增强系统训练依赖于大量的数据，带有干净语音标签的带噪语音数据在现实场景中是很难获得的，而单独的干净语音和噪声数据的获取相对容易，因此常使用已有的干净语音数据和噪声数据以不同信噪比合成带噪语音数据。将带噪语音数据通过信号处理方法或编码器网络提取声学特征后再通过深层神经网络建模带噪声学特征和干净声学特征之间的相关性，经增强后的声学特征再转换为语音信号。通过构建干净声学特征和增强后声学特征之间的损失函数，或干净语音信号与增强后语音信号之间的损失函数，经反向传播算法更新神经网络参数，得到一个能够较好地恢复出干净语音信号的语音增强系统。

目前常用于单通道语音增强的声学特征有梅尔频率倒谱系数（Mel-Frequency Cepstral Coefficients，MFCCs），Bark 频率倒谱系数（Bark Frequency Cepstral Coefficients，BFCCs），幅度谱，复数频谱等。而训练目标的选择主要有两种：基于映射的目标和基于掩蔽的目标。基于映射的目标直接描述干净语音特征，如常用的频谱映射直接预测干净语音的频谱。基于掩蔽的训练目标描述带噪语音特征和干净语音特征之间的关系，通过将预测的掩蔽值与语音特征相乘完成语音增强，常用的掩蔽有：

（1）理想二值掩蔽

基于深度学习的语音增强第一个使用的训练目标就是理想二值掩蔽（Ideal Binary Mask，IBM），理想二值掩蔽是基于判断时频点中噪声或语音占主导的想法提出的，IBM 定义如下：

$$IBM(f,t)=\begin{cases}0, & SNR(f,t)\le LC\\ 1, & SNR(f,t)>LC\end{cases} \qquad (2-3)$$

若一个时频点的信噪比超过设定的阈值，即语音主导时，IBM 则为 1，否则为 0。

（2）理想比值掩蔽

理想比值掩蔽（Ideal Ratio Mask，IRM）可以看作是理想二值掩蔽的软判决，IRM 是该时频点的干净语音能量与带噪语音能量的比值，其定义表达式如下：

$$IRM(f,t)=\left(\frac{|X(f,t)|^2}{|X(f,t)|^2+|N(f,t)|^2}\right)^{\gamma} \quad (2-4)$$

其中 γ 控制掩蔽的大小，通常情况下 γ 取为 0.5。

（3）频谱幅度掩蔽

频谱幅度掩蔽（Spectral Magnitude Mask，SMM)定义为时频点内干净语音幅度与带噪语音幅度的比值，其值也在 0 至 1 范围内，用于预测干净语音幅度，其定义表达式如下：

$$SMM(f,t)=\frac{|X(f,t)|}{|Y(f,t)|} \quad (2-5)$$

式中 $|X(f,t)|$ 和 $|Y(f,t)|$ 分别表示干净语音与带噪语音在第 t 帧，第 f 频点的幅度。

（4）复数比值掩蔽

复数比值掩蔽（Complex Ratio Mask，CRM）与上述几种掩蔽不同的是，CRM 考虑了语音的相位信息，通过分别重构时频点内语音频谱的实部和虚部，实现恢复干净语音的幅度与相位，其定义表达式如下：

$$CRM=\frac{X_rY_r+X_iY_i}{{Y_r}^2+{Y_i}^2}+\frac{X_rY_i+X_iY_r}{{Y_r}^2+{Y_i}^2}j \quad (2-6)$$

式中 X_r，X_i，Y_r，Y_i 分别表示干净语音与带噪语音频谱的实虚部，为使得表达式更简洁，式 $(2-6)$ 省略了 (f,t)，但其仍然是对语音每一帧的每一个频点的信号进行处理

测试阶段：对于给定的带噪语音数据，提取与训练阶段相同的声学特征后送入已训练完成的语音增强系统，在获得预测的干净语音特征后，通过信号处理恢复出干净语音信号，完成整个语音增强过程。

2.2 RNNoise

随着语音增强技术的发展，许多实时通信和智能家居系统开始将语音增强作为系统的前端技术，然而这些应用对于系统实时性的要求非常高。基于深层神经网络的语音增强方法为了性能的考虑，往往有非常庞大的系统参数量和计算量，这使得其能应用的场景受到限制，因此实时的单通道语音增强成为了语音增强领域的研究热点。RNNoise 是 Jean-Marc Valin 等人于 2018 年提出的一种单通道实时语音增强系统。其利用简单的语音增强网络和传统语音增强方法相结合的技术，在低参数量低延迟的条件下完成了高性能的语音增强，在工业界被广泛地应用。下面对 RNNoise 从整体系统流程，声学特征，神经网络结构，

所结合的传统信号处理方法四个方面进行详细介绍。

2.2.1 RNNoise 系统流程

RNNoise 系统的整体流程如图 2-2 所示。RNNoise 是一种用于全频带信号（48kHz 采样率）的算法，其在时频域进行语音增强。首先，由于语音的短时平稳性，RNNoise 将时域的信号进行分帧，其中帧长为 20ms，帧移为 10ms，而后加 Vorbis 窗函数[37]改善帧首位的连续性，Vorbis 窗函数的计算方式如(2 − 7)所示：

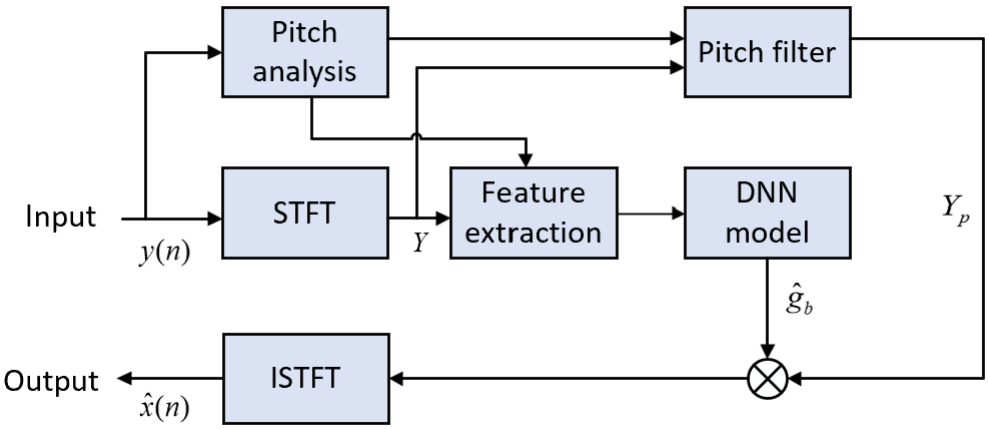


图 2-2 RNNoise 系统流程

$$\omega(k) = sin\left[\frac{\pi}{2}sin^2(\frac{\pi k}{N})\right] \qquad (2-7)$$

式中 N 为窗长。通过快速傅里叶变换（Fast Fourier Transform，FFT）将时域信号变换到时频域后提取不同频带的声学特征，并送入基于循环神经网络（RNN）的网络模型预测掩蔽值。除此之外，RNNoise 将进行基音分析（Pitch Analysis），并根据计算得到的基音周期对带噪语音信号进行滤波（Pitch Filtering），之后将滤波后的信号与网络模型估计得到的掩蔽值相乘得到预测的时频域干净语音，最后通过逆快速傅里叶变换（Inverse Fast Fourier Transform，IFFT）将时频域信号转回时域，经去窗函数，重叠相加操作后得到最终输出的增强后语音。综上所述，RNNoise 的整体流程和主流的基于深度学习的语音增强方法比较类似，不同的是将信号划分频带处理并加入了基音滤波器用于滤除谐波间噪声。

2.2.2 声学特征

为降低系统的复杂度和计算量，RNNoise 降低频率分辨率，对语音信号进行分频带处理，频带的划分依据 Bark 尺度，总共分为 22 个频带。RNNoise 对每

一帧语音信号提取 42 维特征，其中 22 维是每一个频带的 BFCC，其计算方式与 MFCC 类似：首先将 FFT 变换后的频谱特征变换至 Bark 频谱[38]，而后计算对数能量并将其转换为倒谱，通过对频带中频点能量进行累加得到该频带的特征，最后对其进行 DCT 变换得到 BFCC。前 6 维 BFCC 的一阶差分和二阶差分，共 12 维，也被作为声学特征使用。除此之外，RNNoise 对基音周期进行估计，并根据估计得到的基音周期计算信号的基音相关性，其计算方法如下：

$$p_b = \frac{\sum_f \omega_b(f)\, R[|Y(f)||P^*(f)|]}{\sqrt{\sum_f \omega_b(f)|Y(f)|^2 * \sum_f \omega_b(f)|P(f)|^2}} \qquad (2-8)$$

式中 $P(f)$ 为延迟一个基音周期的信号 $y(k-T)$的 DFT 变换，$R[]$表示复数值的实部，.* 表示复共轭。

前 6 个频带的基音相关性的 DCT 变换也被作为声学特征使用。除上述 40 维声学特征外，RNNoise 还使用基音周期 T 和频谱非稳定性度量，组成了共 42 维的声学特征。不难发现，RNNoise 的声学特征更关注语音的低频部分，其中有 18 维特征都只涉及 Bark 谱的前 6 个频带，这样的做法是考虑到人耳相对而言对低频信号更敏感，所以更多地使用低频特征使神经网络学习更多的低频信息，从而更好地预测干净语音的低频部分，使增强后语音的听感质量更好。

2.2.3 神经网络结构

RNNoise 使用的神经网络模型主要基于循环神经网络 RNN,其结构如图所示：

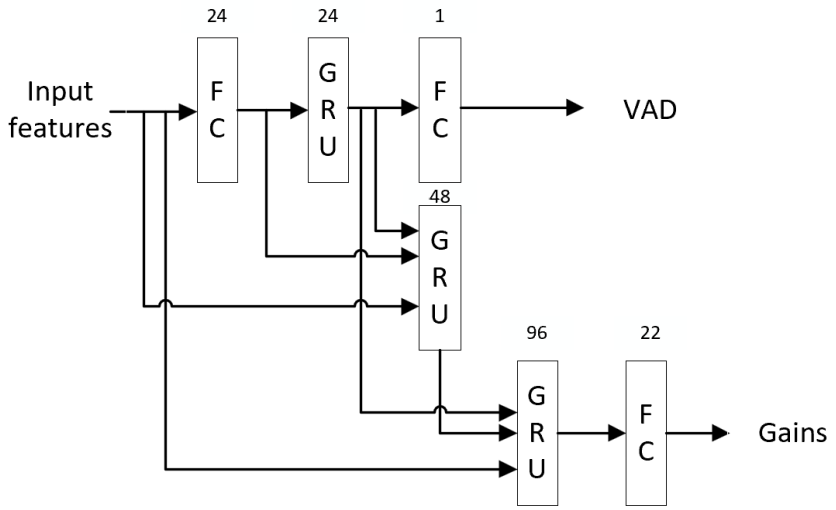


图 2-3 RNNoise 神经网络结构

主要包含三层地 GRU 网络层[39]和三层全连接层，图中每一层网络的括号中数值

表示该层节点数，即该层输出是多少维，其中全连接层的计算只需要当前帧的输入，而 GRU 层的计算只和当前帧与先前帧的信息相关，这是系统能够进行实时语音增强的基础。神经网络的输出由两个部分组成：语音活动检测（Voice Active Detection，VAD）和理想比值掩蔽（在 RNNoise 原论文中称为增益），VAD 值为 0 或 1，当时频点的信号能量大于阈值时为 1，表示该时频点有语音存在，反之则为 0，即该时频点不包含语音。由于 VAD 信息和掩蔽值存在一定的相关性，如 VAD 为 0 时，掩蔽值往往更小，为 1 时掩蔽值往往更大，RNNoise 通过学习 VAD 信息，辅助掩蔽值的学习，使预测更准确。预测的 VAD 值和真实的 VAD 值之间的误差通过交叉熵损失函数计算，其计算方法如下：

$$L_{(V,\hat{V})} = -\left[V \log \hat{V} + (1-V)\log\left(1-\hat{V}\right)\right] \tag{2-9}$$

使用 V 和 $\hat{V}$ 分别表示实际的 VAD 值和预测的 VAD 值。

由于 RNNoise 是分频带处理信号的，其认为同一个频带内的频谱相对平滑，因此网络预测的理想比值掩蔽为 22 维，即同一频带的所有频点共用一个掩蔽值，其计算如下所示：

$$g_b = \sqrt{\frac{E_X(b)}{E_Y(b)}} \tag{2-10}$$

其中 $E_X(b)$ 和 $E_Y(b)$ 分别表示干净语音和带噪语音频带 b 的能量。

这样的做法是为了减小网络中间层的节点数和参数量，从而减少了整个系统的参数量和计算量，使系统实时性更高。预测的掩蔽值与真实的掩蔽值之间的误差通过以下方式计算：

$$L_{(g_b,\hat{g}_b)} = ({g_b}^{\gamma} - {\hat{g}_b}^{\gamma})^2 \tag{2-11}$$

式中 γ 取为 0.5。

2.2.4 基音滤波器

由于 RNNoise 分频带处理信号降低了系统的频率分辨率，且所使用的网络结构相对简单且参数量小，这导致系统虽然实时性非常高，但经神经网络增强后的语音质量相对其他基于神层神经网络的语音增强方法而言较差。为解决这一问题，RNNoise 将传统信号处理方法融入整个语音增强系统，RNNoise 使用一个基音滤波器滤除谐波间的噪声，作用类似于梳状滤波器，其在时频域进行操作，首先根据相关性计算得到基音周期[40] T，将信号延迟时间 T 并转换至频域得到基音延迟信号 P，滤波操作通过式 $Y(f) + \alpha_b P(f)$ 进行计算，其中滤波强度 α_b 通过式$(2-12)$计算：

$$\alpha_b = \min\left(1, \sqrt{\frac{{p_b}^2(1-{g_b}^2)}{{g_b}^2(1-{p_b}^2)}}\right) \qquad (2-12)$$

在式$(2-12)$中，当网络预测的掩蔽值为 1 时，即预测信号中不包含噪声时，滤波强度将为 0，这样信号将不会因为滤波操作而产生失真。当基音相关性为 0 时，滤波强度也将为 0。而后将得到的信号进行调整，使每个频带具有与原始信号相同的能量。这样的滤波操作使得语音中的基音部分得到了增强，谐波间的噪声得到了滤除。之后将滤波后信号与神经网络预测的掩蔽值相乘得到最终增强后的时频域信号。

2.3 PercepNet

2020 年，Jean-Marc Valin 等人为进一步提升单通道实时语音增强性能提出了基于 RNNoise 改进的 PercepNet 系统。相比 RNNoise 系统，PercepNet 使用深度更深参数量更大的神经网络模型，并对使用的声学特征，网络学习目标，基音滤波器做出改进。相比 RNNoise 系统，PercepNet 在测试集上的语音增强性能有了明显提升，且 PercepNet 仍严格限制系统的时延在 40 毫秒之内，符合实时语音增强的需求。PercepNet 的整体流程如图 2-4 所示，可以发现其和 RNNoise 的整体流程基本相同，因此下面将只对 PercepNet 相较 RNNoise 做出的改进，从声学特征，神经网络模型，基因滤波器以及新添加的后处理模块这四方面进行介绍。

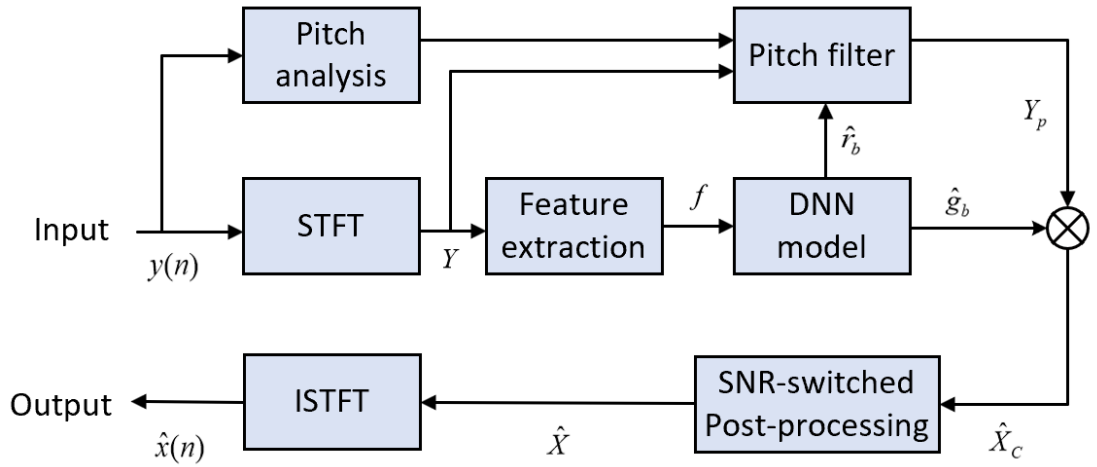


图 2-4 PercepNet 系统流程

2.3.1 声学特征

PercepNet 依然按照划分频带的方式处理语音信号，不同的是 PercepNet 使用更贴合人耳听觉感知的等效矩形带宽[41]（Equivalent Rectangular Bandwidth，ERB）将信号划分为了 34 个频带，并提取了 70 维的声学特征。其中有 34 维是每个频带的能量特征，另有 34 维是每个频带的基音相干性（Pitch Coherence），其是语音信号的复数谱和其周期部分复数谱的余弦距离，计算方法如下所示：

$$q_y = \frac{R[P^H Y]}{\|P\| * \|Y\|} \qquad (2-13)$$

式中 P 表示带噪语音信号中的周期部分，通过使用后续章节介绍的梳状滤波器计算获得，P^H 表示其厄密共轭。

除此之外，PercepNet 将基音周期和基音相关性（Pitch correlation）也作为声学特征，其中基音相关性与 RNNoise 中的计算方式类似，但这里使用的是所有频带基音相关性的均值。

2.3.2 神经网络模型

PercepNet 使用的神经网络模型主要基于循环神经网络，其结构如图所示：

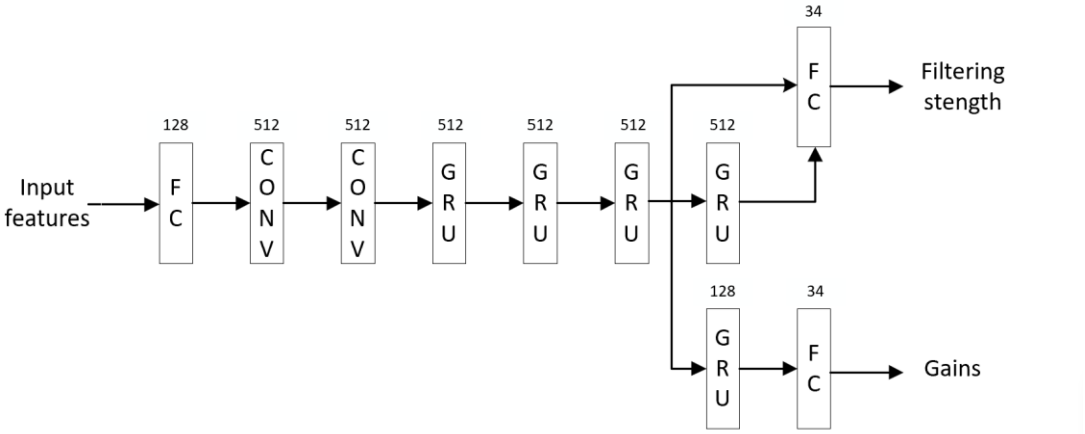


图 2-5 PercepNet 神经网络结构

主要包含 5 层 GRU 网络层，两层一维卷积层以及三层全连接层，无论是神经网络的深度还是中间层的节点数，PercepNet 相比 RNNoise 都有了很大的提升，这使得 PercepNet 系统的参数量是 RNNoise 的数十倍，但由于其严格限制了计算一帧语音信号的声学特征和神经网络计算时用到的未来帧信息，随着硬件设备的发展，参数量的提升并不影响 PercepNet 符合实时语音增强的需求。网络的学习目标由两个部分组成，第一部分与 RNNoise 相似，是 34 维的每个频带的理想比值掩蔽。第二部分是每个频带的基音滤波的滤波强度，RNNoise 中基音

滤波强度由式$(2-12)$直接计算获得，而 PercepNet 系统改为由神经网络预测得到，其真实值的计算方法将在后续介绍基音滤波的章节中详细介绍。两个部分学习目标的损失函数如下所示：

$$L_{(g_b,\hat{g}_b)} = \sum_b (g_b{}^{\gamma} - \hat{g}_b{}^{\gamma})^2 + C_4 \sum_b (g_b{}^{\gamma} - \hat{g}_b{}^{\gamma})^4 \tag{2-14}$$

$$L_{(r_b,\hat{r}_b)} = \sum_b ((1-r_b)^{\gamma} - (1-\hat{r}_b)^{\gamma})^2 \tag{2-15}$$

式 $(2-14)$ 中 C_4 取为 10。

2.3.3 后处理方法

PercepNet 在神经网络模型之后添加了后处理模块。为了进一步增强语音，后处理模块通过以下计算方法获得：

$$\hat{g}_b{}^{\omega} = \hat{g}_b \sin\left(\frac{\pi}{2}\hat{g}_b\right) \tag{2-16}$$

预测得到的较小的掩蔽值通过应用正弦函数使其变得更接近于 0，而较大的掩蔽值将受到较小的影响，以此加大带有严重噪声语音信号的噪声滤除强度。为了防止语音信号整体的过度衰减，PercepNet 引入了一个增益函数 G，其计算方法如下所示:

$$G = \sqrt{\frac{(1+\beta)\frac{E_0}{E_1}}{1+\beta\left(\frac{E_0}{E_1}\right)^2}} \tag{2-17}$$

其中 E_0 表示经过原始掩蔽值$\hat{g}_b$增强后的语音的总能量， E_1 表示经过掩蔽值 $\hat{g}_b{}^{\omega}$增强后的语音的总能量。经过后处理模块的语音信号可由下式表示：

$$\hat{X}_b = G\hat{g}_b{}^{\omega}Y \tag{2-18}$$

为防止增强后的语音不带有任何混响从而听感上非常的不自然，PercepNet 将经过基音滤波，神经网络增强，后处理之后的语音信号再通过下式限制了其最小的能量衰减使语音听感上更自然。

$$\hat{X}_b^{(r)}(t) = min\left(max\left(\hat{X}_b(t), \delta\hat{X}_b(t-1)\right), \hat{Y}_b(t)\right) \tag{2-19}$$

2.3.4 基音滤波

PercepNet 对基音滤波器也做出了改进，首先使用一个五阶的梳状滤波器保留语音中的周期部分，去除非周期部分，其频率响应如图所示：

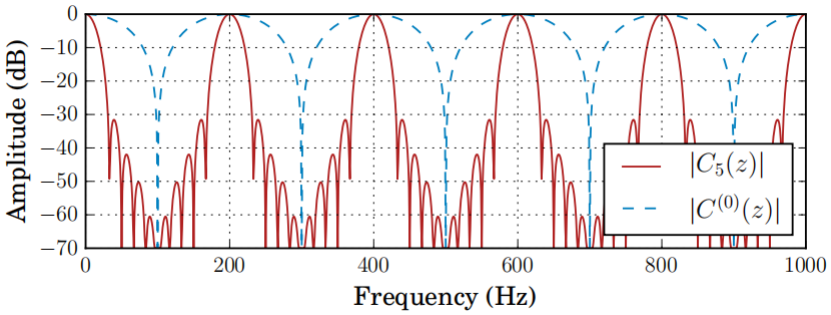


图 2-6 梳状滤波器频率响应

经过基音滤波增强后的信号可用下式表示：

$$Z = (1 - r)Y + r\hat{P} \tag{2-20}$$

其中 $\hat{P}$ 是经梳状滤波器估计的信号周期部分，r 为滤波强度，PercepNet 改用神经网络估计滤波强度，其真实值通过求解基音滤波后信号的基音相干性与干净语音基音相干性相等获得，求解得到的表达式如下所示：

$$r = \frac{\alpha}{1 + \alpha} \tag{2-21}$$

$$\alpha = \frac{\sqrt{b^2 + a\left({q_x}^2 - {q_y}^2\right)} - b}{a} \tag{2-22}$$

式 $(2-22)$ 中$a = {q_{\hat{p}}}^2 - {q_x}^2$，$b = q_{\hat{p}} q_y (1 - {q_x}^2)$。

2.4 实验设计与分析

本章节对上述未开源的 PercepNet 系统进行复现实验，并对实验结果做详细分析。

2.4.1 数据集与实验配置

由于 PercepNet 原论文中所使用的数据集尚未开源，因此本章节实验选用与 RNNoise 论文中相同的训练集，其中干净语音数据来自 McGill TSP speech database[42]和 NTT Multi-Lingual Speech Database for Telephonometry，噪声数据由 RNNoise 论文中的官方网站提供[43]，混响数据来自 RIR_ NOISES set[44]，以上数

据总共包含 6 小时干净语音数据和 4 小时噪声数据。利用以上数据集，共合成了 120 小时的数据用于训练，数据的信噪比范围为-5 至 20dB，每条数据时长为 4 秒。测试数据集选用 PercepNet 原论文中使用的 VCTK 数据集，其包含 824 条测试样本。所有的数据均是 48kHz 采样率采样的，实验中将数据按照 20 毫秒帧长，10 毫秒帧移进行分帧。Batch size 设置为 32，使用 Adam 优化器[45]对网络进行优化，初始学习率设置为 0.001。其余配置按照 PercepNet 论文中的相关描述。

2.4.2 评价指标

本章节实验使用语音增强领域常用的 PESQ[46]和 STOI[47]作为模型性能的评价指标。

PESQ (ITUT P8.62):全称为 perceptual evaluation of speech quality，即语音质量的感知评估，是一种客观的，全参考的评价指标，其计算待测语音和用于参考的干净语音的相关性，得分范围为-0.5 至 4.5，得分越高表示语音质量越高。

STOI：全称为 short-time objective intelligibility，即短时客观可懂度，是一种客观评价指标，与 PESQ 类似，STOI 得分的计算也需要参考语音来计算相似度，但其更侧重于衡量语音的可懂度，得分范围为 0 到 1 之间，为了更直观地表示，本章节实验使用百分比的表示形式，得分越高表示语音可懂度越高。

2.4.3 实验结果分析

表 2-1 展示了复现的 PercepNet 以及 RNNoise 在 VCTK 测试集上的性能，表格的第一行是测试集中原始带噪数据的得分，系统 1 和系统 3 是 PercepNet 论文中使用未开源数据集训练的 RNNoise 与 PercepNet 系统性能，系统 2 和系统 4 是使用合成的 120 小时数据训练的 RNNoise 系统与复现的 PercepNet 系统。对比系统 1 和系统 3 的性能可以发现 PercepNet 系统相比 RNNoise 系统有了明显的提升。同时可以观察到系统 2 的 PESQ 得分比系统 1 低了 0.06，由于 RNNoise 系统是作者已开源的，因此系统 1 与系统 2 的性能差异可以归因于训练集时长与合成所用噪声与语音多样性的差异。PercepNet 原论文中使用了 120 小时的干净语音数据和 80 小时的噪声数据，可以合理推测最终合成的训练集时长与复现所用数据集时长直接有着较大的差异，因此即使系统 4 的性能相比系统 3 较差，PESQ 值低了 0.08，考虑到训练数据量差异，可以认为 PercepNet 系统复现正确。

表 2-1 复现实验结果

ID	Model	PESQ	STOI(%)
/	Noisy	1.97	92.12

1	RNNoise	2.29	/
2	RNNoise	2.23	92.74
3	PercepNet	2.54	/
4	PercepNet	2.46	93.43

2.5 本章小结

本章节主要介绍了单通道实时语音增强的基线系统，首先介绍了基于神经网络的单通道语音增强方法。随后详细阐述了针对单通道实时语音增强所提出的 RNNoise 系统，以及在此基础上进行改进的 PercepNet 系统的技术原理。在 VCTK 测试集上对复现的 PercepNet 系统进行了实验验证，实验结果表明对 PercepNet 系统的复现正确，后续第三和第四章节将在复现的 PercepNet 系统基础上进行进一步的有关单通道实时语音增强的研究。

第 3 章 基于 PercepNet 的相位感知结构

许多语音增强系统，包括 PercepNet 在内都只对语音的幅度谱进行增强，并使用带噪语音的相位与增强后的幅度恢复出时域的信号[48][49]，对相位信息的忽略在一定程度上限制了系统的性能。随着语音增强技术的发展，如何准确地恢复出干净语音的相位开始受到了关注[50][51]，本章节将基于 PercepNet 系统，从声学特征，网络学习目标，网络结构入手，提出一种相位感知结构，进一步提升语音增强的性能。最后使用与第二章节实验相同的数据集验证所提出方法的有效性。相关研究内容已发表在语音领域顶级国际会议 INTERSPEECH 2022[52]。

本章节的组织结构如下：第 3.1 节介绍从特征角度提出的相位感知的声学特征；第 2.2 节介绍从学习目标角度提出的相位感知的实虚部掩蔽；第 2.3 节介绍基于 GRU 网络层改进的 TF-GRU 结构；第 3.4 节主要介绍在 VCTK 和 D-NOISE 测试集上对提出的相位感知结构进行验证的相关实验。

3.1 相位感知的声学特征

PercepNet 中使用的声学特征主要包含基于能量的特征和基音相关的特征，这使得神经网络在训练过程中学习到的声学信息不包含语音相位信息。多项工作验证了相位信息对于语音增强任务的重要性[53][54]，因此首先在声学特征方面对 PercepNet 做出改进。直接计算语音的相位是困难的，因此使用语音复数频谱的实部与虚部作为特征，其间接包含了语音的幅度信息和相位信息。出于系统参数量和计算量的考虑，所提出的方法依旧按照划分频带的方式处理语音信号，每个频带计算实部特征和虚部特征，即每个频带中频点实部的均值与虚部的均值。为了使送入神经网络的特征包含更多的声学信息，PercepNet 原有的声学特征依然保留。原有特征和提出的时虚部特征分别通过一层全连接层后进行特征维度的拼接，再输入后续的 GRU 神经网络层，具体操作可用下式描述：

$$GRU_{input} = concat[FC(f_o), FC(f_c)] \qquad (3-1)$$

式中 f_o 表示 PercepNet 中原有的声学特征，f_c 表示增加的相位感知的声学特征。

3.2 相位感知的实虚部掩蔽

PercepNet 系统中使用的掩蔽如式$(2-10)$所示，是频带内的干净语音信号能量与带噪语音信号能量比值的二分之一次方。由于同一频带内信号的平滑性，通过将掩蔽值与带噪语音复数谱的实部虚部分别相乘，调整频带能量与干净语

音频带能量相同，实现语音增强。PercepNet 使用的掩蔽与理想比值掩蔽类似，只增强带噪语音的幅度谱，不对语音的相位进行处理，而直接使用未经增强的带噪语音的相位。为了进一步提升语音增强性能和增强后语音的听感，越来越多的研究会将如何恢复出干净语音的相位考虑在内。本文提出一种相位感知的实虚部掩蔽，分别对信号的实部与虚部进行调整，同时增强语音的幅度与相位，其具体计算方式如下所示：

$$g_b^r(t) = \frac{\|X_b^r(t)\|_2}{\|Y_b^r(t)\|_2} \tag{3-2}$$

$$g_b^i(t) = \frac{\|X_b^i(t)\|_2}{\|Y_b^i(t)\|_2} \tag{3-3}$$

其中 $X_b^r(t)$，$X_b^i(t)$，$Y_b^r(t)$，$Y_b^i(t)$分别表示干净语音与带噪语音频谱的实部与虚部，$\|.\|_2$表示 L2 范数。在神经网络中通过两个分支分别预测实部和虚部的掩蔽，每个分支主要包含一层 GRU 层和一层全连接层，对于两个部分的掩蔽使用同样的损失函数衡量其预测值与真实值之间的误差。其计算方法如下所示：

$$L_g = \sum_b \left(g_b{}^{\lambda} - \hat{g}_b{}^{\lambda}\right)^2 + C_1 \sum_b \left(g_b{}^{\lambda} - \hat{g}_b{}^{\lambda}\right)^4 \tag{3-4}$$

最终训练的损失函数如下式所示：

$$L = C_2 L_g(g_b^r, \hat{g}_b^r) + C_2 L_g(g_b^i, \hat{g}_b^i) + C_3 L_r(r_b, \hat{r}_b) \tag{3-5}$$

式中g_b^r，$\hat{g}_b^r$，g_b^i，$\hat{g}_b^i$，r_b，$\hat{r}_b$分别为实部掩蔽，虚部掩蔽，滤波强度的真实值与预测值，式中C_2为 4，C_3为 1，L_r为由式$(2-15)$计算。

3.3 TF-GRU 网络结构

本章节提出一种 TF-GRU 网络结构用于更好地对已提取包含基音信息，能量信息和相位信息的声学特征的时域相关性和频域相关性建模，其结构如图 3-1 所示，其中 T-GRU 模块即常用的 GRU 网络层用于对时域相关性建模，F-GRU 用于对频域相关性建模，下面对 T-GRU 网络和 F-GRU 网络分别进行介绍。

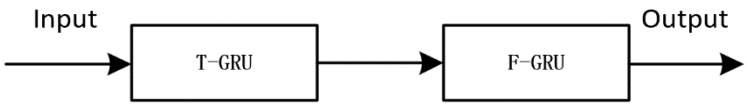


图 3-1 TF-GRU 网络结构

3.3.1 T-GRU 网络结构

T-GRU 即 PercepNet 中使用的 GRU 网络层，其是循环神经网络的一种变种，是语音领域常用的用于前后帧之间的时间相关性建模的网络结构,其具体结构如图所示：

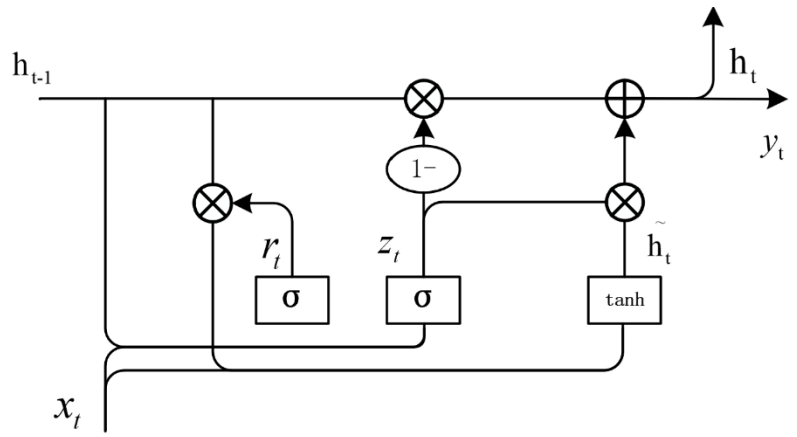


图 3-2 GRU 网络结构

在 t 时刻，GRU 层输入 t 时刻的输入 x_t 以及 $t-1$时刻的隐藏状态 h_{t-1}，输出 y_t 与传递给 $t+1$ 时刻的隐藏状态 h_t。GRU 的内部通过重置门和更新门完成时间相关性的建模，其中重置门决定当前时刻的输入信息如何与前一时刻的记忆信息结合，具体计算方法如下：

$$r_t = \sigma(x_t W_{xr} + h_{t-1} W_{hr} + b_r) \tag{3-6}$$

其中 σ 为 sigmoid 函数，更新门用于控制前一时刻的状态信息被保存到当前状态的程度，其计算方法如下：

$$z_t = \sigma(x_t W_{xz} + h_{t-1} W_{hz} + b_z) \tag{3-7}$$

通过重置门和更新门，GRU 控制记忆信息与当前输入信息在计算当前帧的输出时的重要性，并传递下一时刻计算用到的新的记忆信息，计算方式如下所示：

$$\widetilde{h_t} = tanh(x_t W_{hz} + r_t * h_{t-1} W_{hh} + b_h) \tag{3-8}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h_t} \tag{3-9}$$

$$y_t = \sigma(W_o h_t + b_o) \tag{3-10}$$

其解决了 RNN 不能建模长时相关性的缺点，受到了广泛使用。

3.3.2 F-GRU 网络结构

GRU 已被许多先前的工作证明有着良好的时域相关性建模能力[55][56]，而语音的声学特征除了有时域相关性外，不同频率之间特征的相关性也值得被重视。因此本文提出了 F-GRU 网络结构，其具体结构如下图所示：

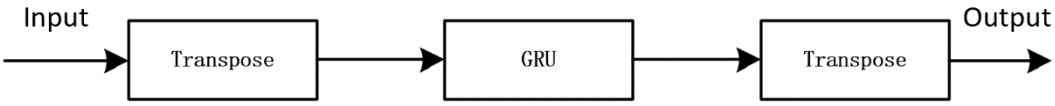


图 3-3 F-GRU 网络结构

由于神经网络的计算过程本质上是矩阵间的运算，巧妙地将输入 GRU 的矩阵进行转置操作，时间轴与频率轴的转换使得原先用于建模时间相关性的 GRU 网络能够对不同频率间的相关性建模，之后将输出的矩阵再次转置将时间轴与频率轴转换到与输入时一致。

TF-GRU 模块先后使用 T-GRU 和 F-GRU 对时间和频率相关性建模，更好地发掘并利用了各类声学特征所包含的声学信息。由于 TF-GRU 包含两个 GRU 层，为了不过多增加整体系统的参数量，保持系统的实时性，设计对其中包含的 GRU 层的节点数进行调整，使得 TF-GRU 模块的参数量与原 PercepNet 中的 GRU 层基本保持一致。

3.4 实验设计与分析

本章节上述基于 PercepNet 提出的相位感知结构进行实验验证，并对实验结果进行分析。

3.4.1 数据集与实验配置

训练数据集沿用第二章节复现 PercepNet 实验所合成的数据，为了更全面地验证提出方法的有效性，避免特定测试集带来的影响。本章节除了 VCTK 测试集外，另合成了一个测试集，命名为 D-NOISE，其包含 108 条样本，其中合成所用干净语音数据从 WSJ0 数据集[57]中选取，噪声数据选取自 RNNoise 官方网站，合成的配置与合成训练集时相同。模型的参数配置与第 2 章节实验相同

3.4.2 评价指标

为了更直观地与基线系统进行性能比较，验证相位感知结构的有效性，本章节实验使用与复现 PercepNet 系统实验时相同的评价指标：PESQ 和 STOI。

3.4.3 实验结果分析

表格 3-1 和 3-2 展示了在 VCTK 测试集以及 D-NOISE 测试集上，本章节介绍的相位感知的声学特征（Complex features），实虚部掩蔽（Complex gains），TF-GRU 网络结构的实验结果，同时对比基线系统 RNNoise 和 PercepNet 的系统性能。（表格中的“+”号均表示在前一个系统的基础上应用一项新的技术，例如系统 4 是在系统 3 的基础上添加了相位感知的实虚部掩蔽）。

表 3-1 VCTK 测试集上相位感知结构的实验结果

ID	Model	PESQ	STOI(%)
/	Noisy	1.97	92.12
1	RNNoise	2.23	92.74
2	PerepNet	2.46	93.43
3	+Complex features	2.51	93.88
4	+Complex gains	2.55	94.54
5	+TF-GRU	2.58	94.87

表 3-2 D-NOISE 测试集上相位感知结构的实验结果

ID	Model	PESQ	STOI(%)
/	Noisy	2.10	86.53
1	RNNoise	2.33	88.27
2	PerepNet	2.57	90.53
3	+Complex features	2.60	91.28
4	+ Complex gains	2.63	92.19
5	+TF-GRU	2.65	92.41

首先通过对比表格 3-2 中系统 1 与系统 2 在 D-NOISE 测试集上的性能可再次验证 PercepNet 系统相较 RNNoise 系统的优势。对比系统 2，3，4，5 的在两个测试集上的性能，可以发现将相位感知的声学特征，实虚部掩蔽和 TF-GRU 结构添加到 PercepNet 基线系统后，系统性能有了连续的提升。

对比系统 5 与系统 2 的性能，相位感知结构总体上为系统带来了明显的性能提升，在 VCTK 测试集上，PESQ 和 STOI 分别提升了 0.12 和 1.44%，在 D-NOISE 测试集上，两项评价指标分别提升 0.08 和 1.88%。接下来对相位感知结构中的每项技术展开详细分析。

系统 3 与系统 2 相比，在 VCTK 测试集上 PESQ 得分提升了 0.05，STOI 得分提升了 0.45%，在 D-NOISE 测试集上的性能提升幅度也基本一致，这验证了

间接包含相位信息的实虚部特征对提升单通道语音增强性能的帮助，通过学习语音的相位信息，能使得系统更好地去除噪声分量，恢复出干净语音信号。对比系统 4 与系统 3，使用实虚部掩蔽在 VCTK 测试集上带来了 0.04 的 PESQ 以及 0.66%的 STOI 的性能提升，在 D-NOISE 测试集上两项评价指标分别提升了 0.03 和 0.91%，这充分说明了语音相位对于语音整体质量的重要性，如何准确地预测干净语音的相位具有重大研究意义。TF-GRU 结构的有效性可以通过比较系统 5 与系统 4 的性能得出，TF-GRU 模块在不增加系统参数量的条件下，使得语音增强性能有了进一步提升，这意味着建模声学特征频率之间的相关性对于准确地恢复出干净语音信号有重要影响。基线系统 PercepNet 的总参数量为 8M，而添加了相位感知系统后参数量为 8.2M。综上所述，基于 PercepNet 提出的相位感知结构能够在几乎不增加系统参数量的条件下显著地提升了单通道实时语音增强的性能，其中相位感知的声学特征，实虚部掩蔽，TF-GRU 结构的有效性均被充分验证。

3.5 本章小结

本章节主要介绍了本文提出的基于 PercepNet 的相位感知结构。首先介绍了相位感知的声学特征，在 PercepNet 原有特征的基础上添加了语音的实虚部特征使神经网络学习语音的相位信息。然后对相位感知的实虚部掩蔽进行详细介绍，将原本只对语音幅度进行增强的 PercepNet 系统改进为对语音的幅度和相位分别增强。接着阐述了 TF-GRU 结构的原理，通过矩阵转置的方式对语音特征的时间相关性和频率相关性建模。最后实验结果表明，提出的相位感知结构能够明显地提升单通道实时语音增强性能，且相位感知的声学特征，实虚部掩蔽，TF-GRU 结构均被证明其有效性。

第 4 章 基于 PercepNet 的多任务学习策略与后处理方法

过度抑制会使语音产生不可逆的失真，降低语音的质量和可懂度[58]。当语音增强系统作为语音识别系统的前端系统时，过度抑制现象相比残留噪声而言会造成更严重的语音识别性能下降，因此本章节为了改善单通道实时语音增强任务中的过度抑制现象，提出了一种多任务学习策略和一种基于信噪比估计的后处理方法。本文将应用了第三章提出的相位感知结构以及本章节提出的多任务学习策略和基于信噪比的后处理方法的系统命名为 PercepNet+，相关研究内容已发表在语音领域顶级国际会议 INTERSPEECH 2022。

本章节的组织架构如下：第 4.1 节介绍 PercepNet+的系统结构；第 4.2 节介绍过度抑制现象与语音增强领域过度抑制现象产生的原因；第 4.3 节介绍为改善过度抑制现象提出的多任务学习策略；第 4.4 节介绍为改善高信噪比语音过度抑制现象提出的基于信噪比估计的后处理方法；第 4.5 节对本章节提出的两种方法在 VCTK 测试集和 D-NOISE 测试集上进行实验验证。

4.1 系统结构

PercepNet+的系统流程和神经网络结构如图 4-1 和 4-2 所示，图 4-1 中红色部分为 PercepNet+系统相较基线系统 PercepNet 所做出改进的模块，其中第 3 章节介绍的相位感知结构主要作用于 Feature extraction 模块和 DNN model 模块。而本章节提出的多任务学习策略对 DNN model 做出进一步改进，基于信噪比估计的后处理方法将网络估计得到的信噪比作为“开关”，使后处理模块被灵活地运用，其具体的技术原理将在第 4.3 章节和第 4.4 章节进行介绍。

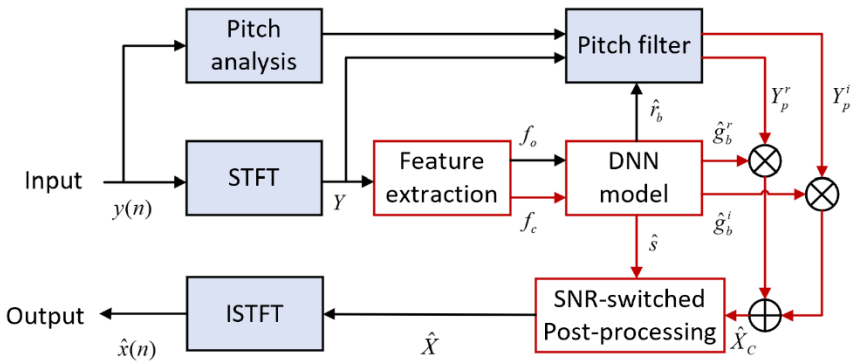
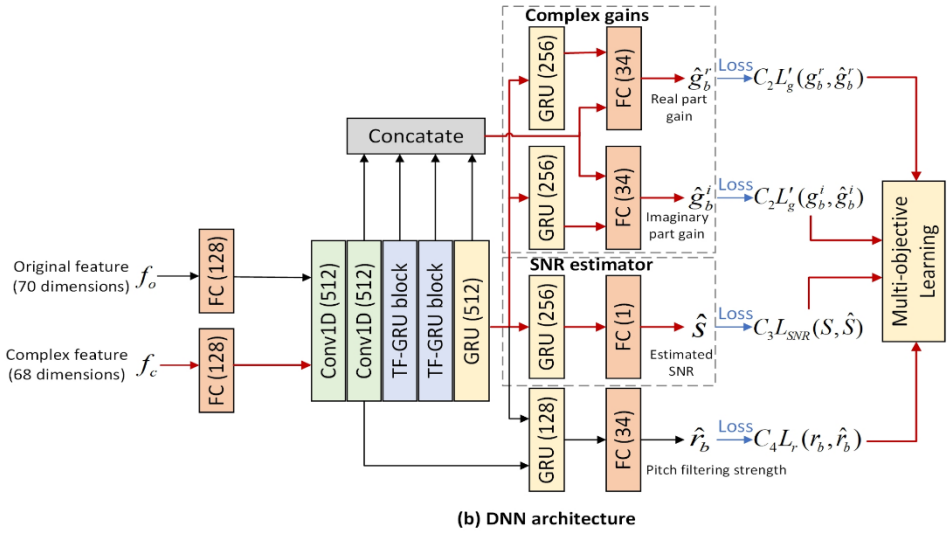


图 4-1 PercepNet+系统流程



(b) DNN architecture

图 4-2 PercepNet+神经网络结构

4.2 过度抑制现象

过度抑制，也称为过度衰减（Over Attenuation），是语音增强领域的一种常见失真，其并不局限于单通道的实时语音增强，所有的语音增强系统均有可能产生该现象[59]。轻度的过度抑制从听感上很难被察觉到，但语音识别系统对其非常敏感；而严重的过度抑制无论是听感上还是用于语音识别系统都是难以接受的。过度抑制现象的产生是“随机”的，多条语音通过同一个语音增强系统，可能只有少部分会产生过度抑制；同一条语音经过不同的语音增强系统，可能同样只有部分系统会使其产生过度抑制。过度抑制的产生与否是由语音增强系统和该语音本身共同决定的，这样的特性使得改善过度抑制变得相对困难。从听感上说，产生过度抑制的语音听起来会有失真，严重的过度抑制甚至导致部分语音内容的缺失，使语音的可懂度受到影响。从数学角度而言，过度抑制的本质是增强后语音的幅度小于对应的干净语音的幅度，即在去除噪声的过程中，有部分语音内容也被去除了。

过度抑制的产生原因主要有以下两点：（1）神经网络的不准确估计。无论是基于时频掩蔽的语音增强系统还是基于特征映射的语音增强系统，都有可能产生过度抑制现象。对于时频掩蔽方法而言，当预测的掩蔽值小于真实的掩蔽值时，语音信号的部分内容被去除，将产生过度抑制现象。对于基于特征映射的方法，如最常见的频谱映射而言，当预测的频谱幅值小于真实干净语音幅值时，过度抑制随之产生。（2）后处理模块进行二次语音增强。近年来，许多语音增强系统除了使用神经网络进行增强之外，还会使用额外的后处理模块用于去除可能存在的残留噪声。后处理模块主要包含以下三种形式：直接对神经网

络预测的掩蔽或特征进行数值上的调整；一种传统的语音增强方法；一个额外的小型的语音增强神经网络[60][61]。无论是哪一种形式的后处理，其目的都在于将噪声去除得更彻底，而这也使得语音部分更有可能被去除，导致过度抑制现象的产生。

4.3 多任务学习策略下的损失函数

多任务学习策略是目前图像、语音领域常用的学习策略，通过为模型增加除预测主要学习目标之外的额外任务，使模型学习到额外的信息，通常添加的额外任务的学习目标与主要学习目标之间存在一定程度的相关性。为改善基线系统 PercepNet 由于网络的不准确估计造成的过度抑制现象，本文提出一种多任务学习策略，调整训练所用的损失函数。首先，本文为模型添加了信噪比估计的额外任务，信噪比与掩蔽值的计算存在一定相关性，通过预测每一帧的归一化后的语音信噪比，使模型学习到语音的信噪比信息，这在先前的诸多工作中被证实对语音增强有益[62][63]。

本文使用额外的一层 GRU 网络层和一层全连接层对归一化后的信噪比进行预测，其实际值的计算如下式：

$$S(t) = \frac{Q(t) - \mu}{\sigma} \tag{4-1}$$

$$Q(t) = 20\log_{10}\left(\frac{X_m(t)}{N_m(t)}\right) \tag{4-2}$$

式$(4-1)$中 μ 和 σ 分别代表信噪比 $Q(t)$ 的均值与方差，式 $(4-2)$ 中 $X_m(t)$ 和 $N_m(t)$ 分别代表干净语音与噪声的幅度谱。对于该任务的预测值与真实值之间的误差，通过均方误差损失函数进行计算。

$$L_{SNR} = \left|S(t) - \hat{S}(t)\right|^2 \tag{4-3}$$

为了使训练得到的模型尽可能地不预测出比真实值更小的掩蔽值，本文使用一种针对性的损失函数对其进行改善。对于第三章节提出的相位感知的实虚部掩蔽，原先都使用式 $(3-4)$ 计算损失。由于增强后语音的幅值与实虚部掩蔽均相关，本文对这两部分损失函数增加额外的惩罚项，称为过度抑制损失函数（Over Attenuation Loss，OA Loss），具体计算方法如下式：

$$L_g' = \delta L_g + (1-\delta)L_{OA} \tag{4-4}$$

其中 L_{OA} 可通过式 $(4-5)$ 和 $(4-6)$ 计算：

$$h(x)=\begin{cases}0 & if\ x\le 0\\ x & if\ x>0\end{cases} \tag{4-5}$$

$$L_{OA}=|h(g_b-\hat{g}_b)|^2 \tag{4-6}$$

当预测的掩蔽值大于等于真实值时，这部分惩罚项将为零，当预测的掩蔽值小于真实值时，额外的惩罚项会使损失函数值更大，进而使通过反向传播优化完成的模型更倾向于预测出一个较大的掩蔽值。

基于以上的改进，模型共有四个部分的输出：实部掩蔽，虚部掩蔽，信噪比以及滤波强度。本文将模型作为一个整体训练，将以上四个部分的损失函数通过一定比例进行融合得到完整的训练损失函数，其具体计算如下：

$$L=C_2L'_g(g_b^r,\hat{g}_b^r)+C_2L'_g(g_b^i,\hat{g}_b^i)+C_3L_r(r_b,\hat{r}_b)+C_4L_{SNR}(S,\hat{S}) \tag{4-7}$$

其中 $C_2=4$，$C_3=1$，$C_4=1$。

4.4 基于信噪比估计的后处理方法

PercepNet 系统产生过度抑制现象的另一原因是后处理模块可能使得预测的掩蔽值变得更小。目前已有的后处理方法为了尽可能地去除残留噪声，对增强后的语音再次进行语音增强，这无疑使过度抑制发生的概率大大提升，也使得原本轻微的，可接受的过度抑制变得更严重。

为解决此类由后处理方法导致的过度抑制，本文提出一种基于信噪比估计的后处理方法，其以传统语音增强方法最小均方误差准则下的对数谱幅度估计（MMSE-LSA）为基础，并添加 SNR switch（信噪比“开关”）对其进行更灵活的运用，下面分别对 MMSE-LSA 的基本原理以及提出的 SNR Switch 进行详细介绍。

4.4.1 MMSE-LSA

MMSE-LSA 方法[64]假定干净语音信号与噪声是互相独立的，且符合高斯分布，由于人耳对声强的感知是非线性的，而在对数域上的感知几乎可以认为是线性的，因此 MMSE-LSA 将对干净语音幅度谱的估计问题转换为对于对数幅度谱的估计，其目标是使得估计得到的语音对数幅度与实际对数幅度之间的均方误差最小，且不对干净语音的相位进行估计，即求下式最小：

$$L=E\left\{\left(\ln M_f-\ln \widehat{M}_f\right)^2\right\} \tag{4-8}$$

其中 M 和 $\widehat{M}$ 分别表示干净语音的幅值和其估计值，$\widehat{M}$ 可以写成如下形式：

$$\widehat{M}_f = exp\left\{E\left(\ln M_f \,|Y_f\right)^2\right\} \tag{4-9}$$

基于高斯模型的假设，可以推导得：

$$\widehat{M}_f = \frac{\xi_f}{1+\xi_f} exp\left\{\int_{v_f}^{\infty} \frac{e^{-t}}{t} dt\right\} R_f \tag{4-10}$$

$$v_f = \frac{\xi_f}{1+\xi_f} \gamma_f \tag{4-11}$$

式中 ξ 为先验信噪比，γ 为后验信噪比，R 为带噪语音的幅值，可以看出其计算只依赖于先验信噪比和后验信噪比。先验信噪比和后验信噪比通过下式估计：

$$\xi(t) = \frac{CDF\left(\sigma_{\hat{x}}^2(t)\right)}{CDF\left(\sigma_N^2(t)\right)} \tag{4-12}$$

$$\gamma(t) = \frac{CDF\left(\sigma_Y^2(t)\right)}{CDF\left(\sigma_N^2(t)\right)} \tag{4-13}$$

其中 CDF 为累积分布函数，$\sigma_{\hat{x}}^2(t)$ 是神经网络预测的干净语音信号的方差，$\sigma_y^2(t)$ 和 $\sigma_n^2(t)$ 分别是带噪语音的方差和噪声的方差，噪声通过将带噪语音与增强后语音相减获得。MMSE-LSA 方法与其他传统的语音增强方法有类似的缺点，比如假设噪声与语音的平稳性，噪声和语音相互独立等，这限制了其作为独立的语音增强系统的性能，但其作为基于深层神经网络的语音增强系统的后处理模块用于去除残留噪声在一些已有的研究[62]中被证实切实可行，因此本文将 PercepNet 系统原有的后处理模块替换为 MMSE-LSA 模块。

4.4.2 SNR Switch

后处理的确将噪声去除的更干净，但也导致了过度抑制的产生。要既完全地将残留噪声去除，同时不对语音部分带来损失是非常困难的，无论是传统的语音增强方法还是小型的网络模型都很难做到这一点。实际上，并不是所有的语音都需要后处理方法进行二次的语音增强，尤其是高信噪比的语音，其包含的噪声相对较弱，即使未经语音增强也能清晰地理解语义内容，听感上相对较好，因此只需要神经网络对其进行一次语音增强就能恢复出非常干净的语音信号，后处理方法反而很有可能降低了其质量和可懂度。而信噪比较低的信号由于包含强烈的噪声，即使通过神经网络进行了第一次语音增强，增强后的语音可能仍包含较明显的残留噪声，听感上相对较差，这种情况下运用后处理方法

再次增强就非常必要。因此本文提出 SNR Switch（信噪比“开关”）用于判断当前输入的语音是否需要后处理模块进行第二次语音增强，结构如图 4-3 所示。

具体来说，该机制利用第 4.2 章节提出的信噪比估计网络估计的信噪比作为判断的标准，当估计得到的信噪比大于设定的阈值时，后处理模块将不被执行，系统直接使用经神经网络模块增强的语音作为最终输出的语音信号，而当估计的信噪比小于或等于设定的阈值时，系统执行后处理模块去除残留噪声，将经过两次增强的信号作为最终的输出。

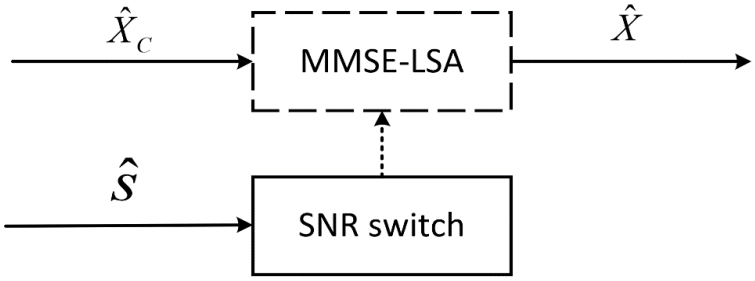


图 4-3 基于信噪比估计的后处理方法示意图

4.5 实验设计与分析

本章节对上述为改善过度抑制现象提出的多任务学习策略和基于信噪比估计的后处理方法进行实验验证，并对实验结果进行分析。

4.5.1 数据集与实验配置

本章节实验沿用第三章实验所用训练集与测试集，并且在第三章系统的基础之上验证本章节提出方法的有效性。模型训练所用参数配置与第二、第三章节实验相同。

4.5.2 评价指标

本章节实验同样主要使用 PESQ 和 STOI 进行评估，通过两项指标的性能差异验证方法的有效性。

4.5.3 实验结果分析

首先对基线系统 PercepNet 系统存在的过度抑制现象进行分析，由于轻微的过度抑制无论是对主观听感还是对客观性能指标都只造成很小的影响，在完全可接受的范围，因此主要针对严重的过度抑制，尤其是经过整个 PercepNet 系统

增强后的语音质量低于原始带噪语音质量的样本进行分析。图 4-4 展示了 VCTK 测试集中不同信噪比下原始带噪样本的 PESQ 得分以及经 PercepNet 增强后的 PESQ 得分。通过两条曲线的对比可以发现，上述两种情况语音的 PESQ 得分差异随着信噪比的提升而减小，当信噪比高于 14dB 时，经 PercepNet 增强后语音的性能反而低于原始带噪语音，这很大程度是由于当信噪比大于 14dB 时语音的质量已经较高，有相当一部分样本经过增强后产生了严重的过度抑制，使得分降低。图 4-5 展示了 VCTK 测试集整体信噪比分布（图中粉色部分），以及其中经过 PercepNet 增强后，PESQ 得分降低的样本分布情况，其中横坐标是信噪比，纵坐标为样本条数。可以发现灰色部分，即产生严重过度抑制使得 PESQ 得分低于原始带噪信号的样本主要位于高信噪比区间，其中 14-17dB 的灰色部分占比甚至超过了该区间内所有样本的百分之七十，因此 4.3.2 章节提出的 SNR Switch 的阈值选为 14dB。以上实验结果充分说明了 PercepNet 系统容易产生过度抑制现象，且在高信噪比情况下过度抑制非常严重，使得大部分高信噪比语音增强后的质量低于未经增强之前，因此本章节提出的多任务学习策略以及基于信噪比估计的后处理方法是有极大意义的。

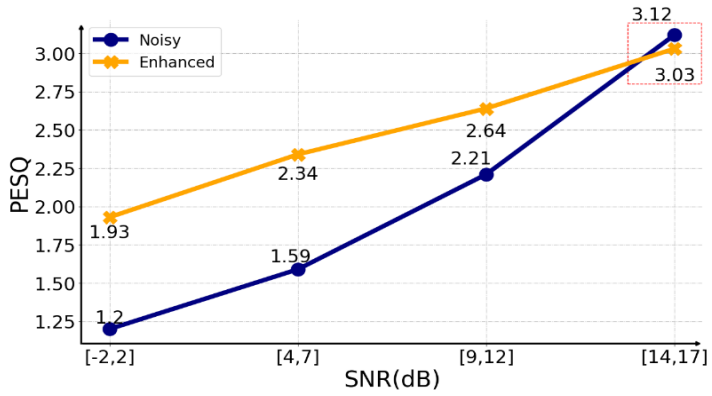


图 4-4 VCTK 测试集上，不同信噪比样本经 PercepNet 增强前后对比

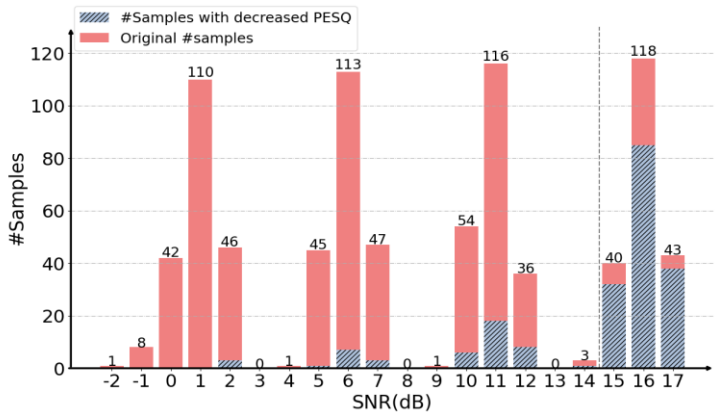


图 4-5 VCTK 测试集原始样本以及增强后 PESQ 得分下降样本的信噪比分布

表 4-1 展示了在 VCTK 测试集和 D-NOISE 测试集上，本章节提出的多任务学习策略以及基于信噪比估计的后处理方法的性能（表格中的“+”号均表示在前一个系统的基础上应用一项新的技术，例如系统 4 是在系统 3 的基础上添加了相位感知的实虚部掩蔽）。系统 6 相比系统 5 在 VCTK 测试集上的 PESQ 得分由 2.58 提升至 2.61，STOI 得分由 94.87%提升至 95.40%，在 D-NOISE 测试集上的性能提升基本相同，验证了添加额外的信噪比估计任务（SNR estimator），使得网络学习到语音的信噪比信息对于预测掩蔽值的有效辅助作用。对比系统 7 和系统 6 的性能，可以发现过度抑制损失函数（OA Loss）对于整个测试集的性能没有太大的影响，这主要是由于其只添加惩罚项使得模型尽可能不预测出一个偏小的掩蔽值，这使得语音分量被保留得更完整，但是噪声分量有部分残留的可能性也相应提高了，然而其对于语音识别这类更注重语音信号完整性的后端任务来说依旧是有意义的，其有效性需要后续针对过度抑制现象的实验进行验证。系统 8 相较系统 7 性能提升明显，其中在 VCTK 测试集上 PESQ 得分提升了 0.03，STOI 提升了 0.30%，在 D-NOISE 测试集上两项指标分别提升了 0.04 和 0.70%，这充分体现了基于信噪比估计的后处理方法（SNR-switched Post-processing, SNR-switched PP）相较原 PercepNet 系统后处理方法的优势，解决了原后处理方法造成过度抑制使得高信噪比语音质量降低的问题。

表 4-1 VCTK 测试集上多任务学习策略和基于信噪比估计的后处理方法的实验结果

ID	Model	PESQ	STOI(%)
/	Noisy	1.97	92.12
1	RNNoise	2.23	92.74
2	PerepNet	2.46	93.43
3	+Complex features	2.51	93.88
4	+Complex gains	2.55	94.54
5	+TF-GRU	2.58	94.87
6	+SNR Estimator	2.61	95.40
7	+OA Loss	2.62	95.38
8	+SNR-switched PP	2.65	95.68

表 4-2 D-NOISE 测试集上多任务学习策略和基于信噪比估计的后处理方法的实验结果

ID	Model	PESQ	STOI(%)
/	Noisy	2.10	86.53
1	RNNoise	2.33	88.27
2	PerepNet	2.57	90.53
3	+Complex features	2.60	91.28

4	+Complex gains	2.63	92.19
5	+TF-GRU	2.65	92.41
6	+SNR Estimator	2.68	92.75
7	+OA Loss	2.68	92.76
8	+SNR-switched PP	2.72	93.46

表 4-3 展示了 VCTK 的高信噪比（>14dB）和低信噪比（≤14dB）子测试集上，过度抑制损失函数和 SNR Switch 对于改善过度抑制现象的表现，系统 3，4，5 是分别在 PercepNet+基础上去除 SNR Switch 或过度抑制损失函数后的系统，表 4-3 中得分均为 PESQ 得分。分别对比系统 5 和系统 3 以及系统 5 和系统 4，可以得出 OA Loss 对于不同信噪比情况下的过度抑制情况的改善相对均衡，而 SNR Swtich 对于高信噪比情况下的过度抑制改善尤为明显，其将大于等于 14dB 的样本 PESQ 得分提高了 0.05。对比系统 2 和系统 3 以及系统 2 和系统 4，可以发现当两项技术均被应用时，性能比任意应用其中一项更好，过度抑制现象也被进一步改善。

添加了多任务学习策略和基于信噪比估计的后处理方法的系统总参数量为 8.5M，相较应用这两项技术之前增加了 0.3M，可以得出任务学习策略和基于信噪比估计的后处理方法能在不影响系统实时性的基础上解决过度抑制问题，提升系统性能与鲁棒性。

表 4-3 在 VCTK 高低信噪比的子测试集上，过度抑制损失函数和 SNR Switch 的实验结果

ID	Model	≤14dB	>14dB	Overall
/	Noisy	1.59	3.12	1.97
1	PercepNet	2.28	3.03	2.46
2	PercepNet+	2.49	3.14	2.65
3	w/o SNR Swtich	2.49	3.11	2.64
4	w/o OA Loss	2.48	3.13	2.64
5	w/o SNR switch & w/o OA Loss	2.48	3.03	2.62

4.6 本章小结

本章主要介绍了提出的多任务学习策略和基于信噪比估计的后处理方法。首先，对语音增强领域常见的过度抑制现象以及其产生的原因进行详细阐述。其次，介绍了为改善由网络不准确估计造成的过度抑制而提出的多任务学习策略，其主要通过添加信噪比估计任务和过度抑制损失函数使得神经网络学习到更多的声学信息以及更难预测出比实际值小的掩蔽值。然后，介绍了为针对性

解决高信噪比情况下的严重过度抑制问题提出的基于信噪比估计的后处理方法，通过估计得到的信噪比决定 MMSE-LSA 后处理模块是否被执行。最后实验结果表明，所提出的方法在 VCTK 测试集和 D-NOISE 测试集均能带来性能的提升，且过度抑制损失函数和基于信噪比估计的后处理方法可明显地改善过度抑制的情况。

第 5 章 基于 sDPCCN 的鲁棒性个性化语音增强方法

现实场景中的声学环境是非常复杂的，除背景噪声外可能还包含干扰说话人，当存在多个说话人同时说话时，目标语音的可懂度与质量也会大幅下降。然而，由于缺少目标说话人的信息，无论是传统的语音增强方法还是基于深度学习的语音增强系统，在实际使用时都无法准确地辨别干扰说话人与目标说话人，从而不能恢复出干净的目标语音信号。因此，个性化的语音增强任务，或称为特定说话人的语音增强任务成为了目前研究的热点，即同时去除背景噪声与干扰说话人语音，只保留目标说话人的语音信号。本章节基于 sDPCCN 系统对单通道个性化语音增强任务展开研究，针对声学环境不匹配问题提出了动态声学补偿方法，以及针对困难样本问题提出了自适应焦点训练机制。

本章节的组织架构如下：第 5.1 节主要介绍基于深层神经网络的个性化语音增强原理；第 5.2 节介绍基线系统 sDPCCN 原理；第 5.3 节介绍个性化语音增强任务中存在的测试语音与注册语音声学环境不匹配问题，以及为解决此问题提出的动态声学补偿方法；第 5.4 节介绍困难样本问题以及为解决此问题提出的自适应焦点训练机制；最后第 5.5 对本章节提出的两种方法在 DNS-test 测试集上进行实验验证。本章节研究内容已投稿至 ICASSP 2023。

5.1 个性化语音增强原理

基于深层神经网络的单通道个性化语音增强系统与第 2.1 章节介绍的基于深层神经网络的单通道语音增强系统类似，本章节对于两者差别进行介绍。其主要差别在于数据合成阶段以及深层神经网络结构，图 5-1 给出了整个个性化语音增强系统的示意图。由于现实场景中背景噪声和干扰说话人可能同时存在，也可能不同时存在，因此在数据合成阶段，为了使训练得到的模型能应对现实场景中的各类声学环境，一共需要合成三种类型的训练数据：（1）包含目标说话人语音和背景噪声的数据（2）包含目标说话人语音和干扰说话人语音的数据（3）同时包含目标说话人语音，背景噪声和干扰说话人语音的数据。

非个性化的语音增强系统之所以无法去除干扰说话人语音是由于缺少目标说话人信息，无法区别需要保留的目标说话人和其他干扰说话人。因此个性化的语音增强系统通过在深层神经网络中引入额外的目标说话人注册语音信息，指导网络保留与注册语音声学特性最相似的说话人语音。目标说话人注册语音信息通常使用额外的说话人编码器（Speaker encoder）提取能够表征说话人信息的嵌入向量[65]，再将其与神经网络某一层的输出的嵌入向量通过点积相乘，拼

接等方式进行融合再输入下一层语音增强网络，在训练时将语音增强网络与 Speaker encoder 的参数同时更新。除以上使用额外的 Speaker encoder 的方法之外，许多工作[66][67]中使用已训练好的模型提取声纹识别任务中常用的如 d-vector，x-vector 等表征说话人信息的嵌入向量，在个性化语音增强系统的训练过程中只对语音增强部分网络参数进行更新。由于需要额外的目标说话人注册语音，合成训练集所用的数据集需要特别挑选，每个说话人需要至少包含两条语音数据，一条用于合成带有干扰说话人或噪声的训练数据，另一条作为注册语音，通常注册语音的长短需要在 15 秒以上能使个性化语音增强系统有较好的性能。

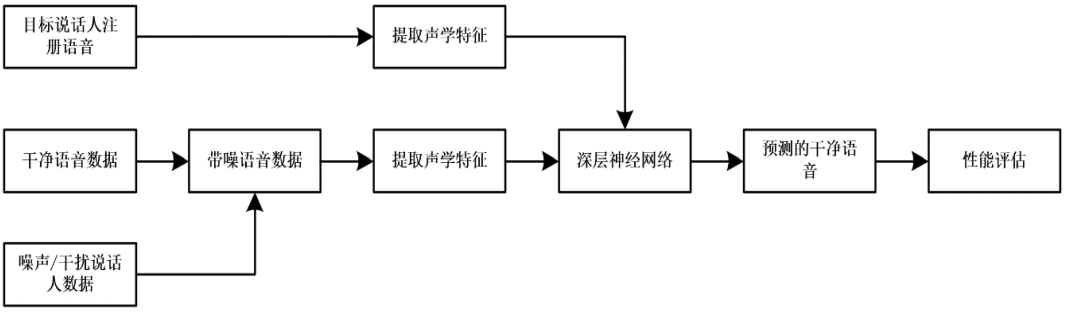


图 5-1 个性化语音增强系统建模流程

5.2 sDPCCN

sDPCCN 系统是由 J.Han 等人提出的用于目标说话人提取的 U-Net 结构网络，其结构如图 5-2 所示。由于其出色的性能，本文将其作为研究单通道个性化语音增强问题的基线系统，下面对其原理进行介绍。

sDPCCN 使用语音的复值频谱作为模型输入，通过编码器（encoder）将输入转变为更高维的表征，再通过解码器（decoder）将其转变为干净语音信号。编码器和解码器主要使用多个 DenseNet[68]模块与二维卷积模块，DenseNet 模块由五个二维卷积模块与残差连接组成。二维卷积模块由二维卷积层，指数线性单元激活函数（ELU）和实例归一化（Instance normalization）组成，而解码器中将二维卷积模块替换为二维解卷积模块。在编码器与解码器之间，sDPCCN 使用两个时域卷积网络（TCN）[69]对语音的长时信息相关性建模，其中每一个 TCN 由 10 组 IN，ELU 以及一维卷积层组成。sDPCCN 在解码器之后使用一个多尺度池化层（Pyramid layer）[70]提升模型对语音时间全局信息建模的能力，其结构如图 5-3 所示。

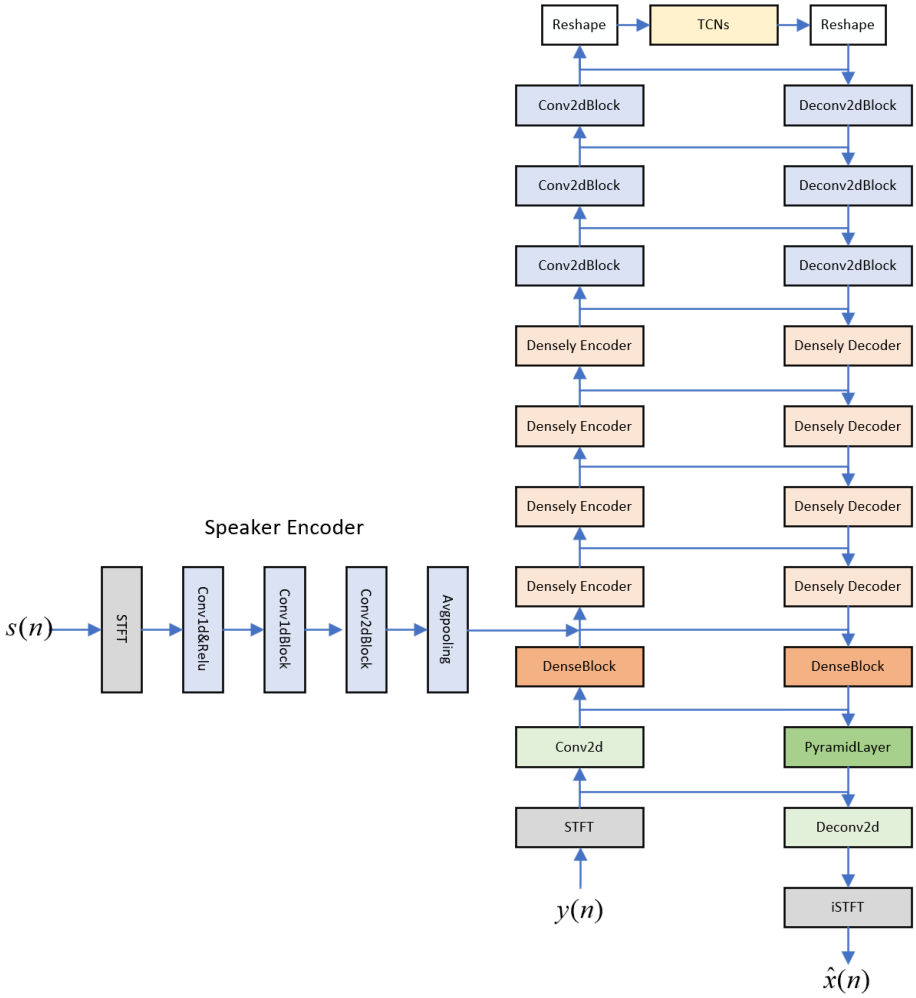


图 5-2 sDPCCN 系统[33]

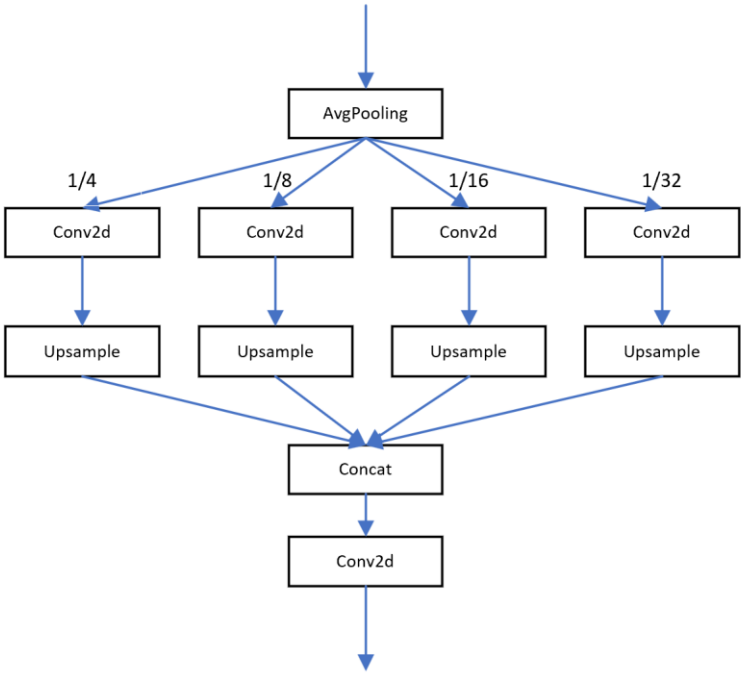


图 5-3 Pyramid layer 结构[33]

除以上网络的主体部分外，sDPCCN 使用一个 Speaker encoder 提取目标说话人声学信息，其使用注册语音的幅度谱作为输入特征，使用多层一维卷积层和 Relu 激活函数提取高维特征，最后使用平均池化层和点乘操作将生成的包含目标说话人声学信息的嵌入向量融合到主体网络中，在训练阶段使用负尺度不变信噪比（SISNR）损失函数[71]衡量预测的语音与实际干净语音之间的差异，并同时对主体网络部分和 Speaker encoder 部分进行参数更新。以上是 sDPCCN 的原理介绍，本文将其应用到单通道个性化语音增强任务中作为基线系统，并在 sDPCCN 基础上提出了动态声学补偿方法和自适应焦点训练机制，整体系统流程如图 5-4 所示。

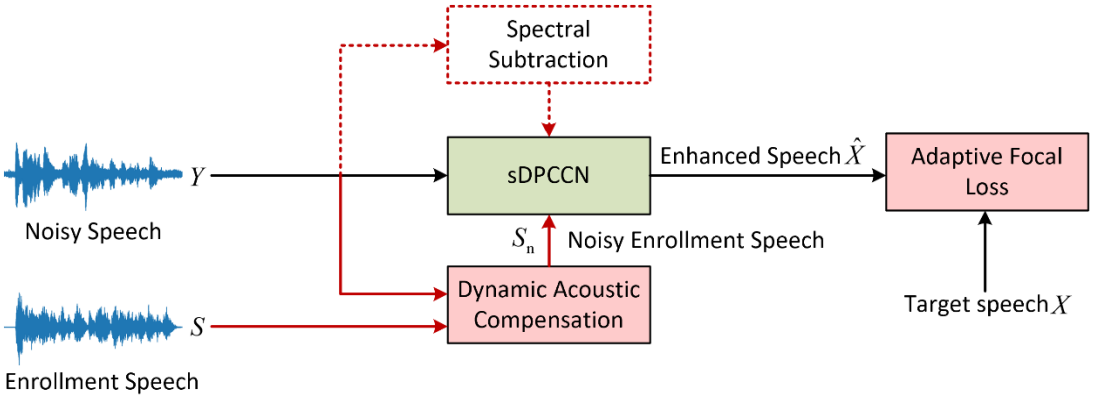


图 5-4 系统整体流程，图中红色部分为本文所做改进部分

5.3 动态声学补偿

个性化语音增强需要额外的目标说话人注册语音的辅助，而实际的测试语音可能与注册语音处于完全不同的声学环境[72]。由于同一个说话人的两条语音不可能同时录制，因此这样的声学环境不匹配是难以避免的。通常个性化语音增强系统使用的注册语音是数据集中提供的不含噪声的干净语音，而测试语音中可能包含严重的背景噪声，这样的声学环境不匹配可能导致系统既无法准确去除背景噪声也无法去除干扰说话人语音，很大程度上影响了个性化语音增强的性能。为改善测试语音与注册语音的声学环境不匹配问题，本文提出了两种动态声学补偿方法：（1）基于谱减法的动态声学补偿（2）基于波形叠加的动态声学补偿。下面对两种方法分别进行介绍。

5.3.1 基于谱减法的动态声学补偿

谱减法利用语音的前若干帧信号的均值估计噪声的能量谱，并将其从整个

语音信号的能量谱中减去以达到去除噪声的目的。由于注册语音是干净不含有噪声的，本文提出对测试语音应用谱减法消除其包含的背景噪声，从而达到消除或减弱测试语音与注册语音之间声学环境不匹配的问题。

5.3.2 基于波形叠加的动态声学补偿

第 5.3.1 节提出的基于谱减法的动态声学补偿虽然能够在一定程度上消除测试语音中的噪声带来的声学环境不匹配，但谱减法有其明显的缺点，即谱减法只利用前若干帧估计噪声能量谱，当噪声非平稳时估计得到的能量谱可能大于实际的噪声能量谱，使经谱减法增强后的语音产生失真。虽然背景噪声造成的声学环境不匹配的问题被缓解了，但对测试语音造成的失真是不可逆的，从而影响系统输出语音的质量。

因此本章节提出基于波形叠加的动态声学补偿，避免了基于谱减法的动态声学补偿的缺陷，实现了在不造成语音失真的条件下消除注册语音与测试语音的声学环境不匹配，其结构如图所示。

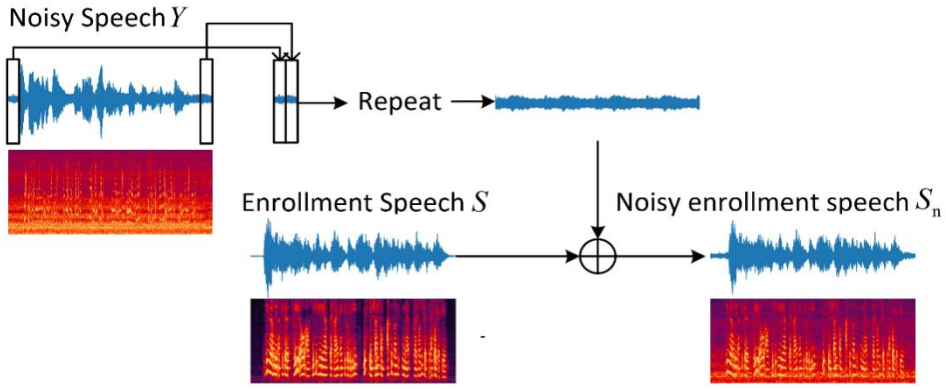


图 5-5 基于波形叠加的动态声学补偿

基于波形叠加的动态声学补偿方法同样受启发于谱减法，但使用与谱减法相反的操作，首先截取测试语音的前若干帧或最后若干帧信号，由于现实场景录制的语音的前若干帧和最后若干帧通常不包含人声部分，所以将截取得到的信号段视作背景噪声片段，并将其重复扩展至与注册语音相同时长，将其与干净的注册语音相加得到新的注册语音，计算方法如下式：

$$S_n = repeat(Y_1, \dots, Y_J, Y_{T-K+1}. \dots, Y_T) + S \qquad (5-1)$$

式中 S 为原始注册语音，J 和 K 分别表示截取测试语音的前J 帧和最后 K 帧，T 为测试语音的总帧数。基于波形叠加的动态声学补偿方法通过使注册语音带有与测试语音类似的背景噪声缓解两者声学环境不匹配的问题，该方法所有操

作均在时域完成。虽然该方法与谱减法类似，当测试数据中所含噪声是非平稳噪声时，截取的噪声段将不能包含完整的噪声信息，使得波形叠加后的注册语音与测试语音的声学环境依然匹配程度不足，但该方法不会造成测试语音失真，且在平稳噪声环境下效果优秀，从测试集整体或现实应用场景考虑，该方法依然可以明显提升单通道个性化语音增强系统的性能。

5.4 自适应焦点训练

困难样本问题是语音和图像领域的热点问题之一，近年来受到了广泛的研究[73][74]，但在单通道个性化语音增强领域中针对此问题的研究相对较少。目前衡量一个个性化语音增强系统的性能主要依据测试集的平均性能，而忽略了不同样本的性能差异。测试集上整体性能优秀的个性化语音增强系统，在个别样本上可能性能很差。当这样的个性化语音增强系统应用于现实场景中的任务时，这类性能较差的情况将严重影响系统用户的使用体验。在合成数据时，通常设置信噪比范围为一个适中的区间，这导致训练集中的困难样本一般情况下只占整体的非常小一部分，所以即使训练阶段的困难样本与非困难样本相比通常在计算损失函数时值更大，训练时为使得训练集总体的损失更小，模型仍将被训练为更“偏向”训练集中占多数的非困难样本，而在一定程度上忽视了困难样本的信息，这就是基于深度学习的模型会在个别情况下或个别样本上效果较差的原因。为解决单通道个性化语音增强任务中的困难样本问题，本文提出一种自适应焦点训练方法，下面分别对其中的自适应焦点损失函数和两阶段训练方法进行介绍。

5.4.1 自适应焦点损失函数

基线系统 sDPCCN 使用的训练损失函数是负尺度不变信噪比（Negative SISNR），其衡量预测干净语音与实际干净语音之间的信噪比，当 SISNR 值越大时预测的干净语音越接近实际干净语音，其计算方法如式$(5-2)$所示，其中 x 和 $\hat{x}$ 分别表示时域的增强后的语音信号和实际的干净语音信号，而后通过取负值计算训练损失函数优化网络参数。负尺度不变信噪比损失函数更常用于语音分离或目标说话人提取任务，本文将 sDPCCN 应用于单通道个性化语音增强任务，因此在该损失函数的基础之上添加常用于语音增强领域的最小均方误差损失函数，其计算方法如式$(5-3)$所示，其中 X 和 $\hat{X}$ 分别表示频域的增强后的语音信号和实际的干净语音信号。负尺度不变信噪比损失函数在时域进行计算，而最小均方误差损失函数在频域计算，使得模型训练时同时考虑语音时域与频域的误差，通过学习不同域的信息，进一步提升个性化语音增强系统的性能，

本文将该损失函数命名为 TF-Loss，其计算方法如式$(5-4)$所示。

$$\begin{cases} x_{target} = \dfrac{x\langle x, \hat{x}\rangle}{\|x\|_2} \\ e_{noise} = \hat{x} - x_{target} \\ SISNR = 10\log_{10}\dfrac{\|x_{target}\|_2}{\|e_{noise}\|_2} \end{cases} \tag{5-2}$$

$$L_{MSE} = \left|X - \hat{X}\right|^2 \tag{5-3}$$

$$L_{TF} = L_{-SISNR} + L_{MSE} \tag{5-4}$$

本文在 TF-Loss 的基础上提出一种自适应焦点损失函数，用于改善困难样本问题。其计算方式如下式所示：

$$L_{AFT} = \sum_{i=1}^{B} L_{TF}^{i} * sin\left[\frac{\pi}{2} * \left(\frac{L_{TF}^{i} - \mu}{\sigma}\right)\right] \tag{5-5}$$

其中 B 为 Batch size 大小，μ 和 σ 分别为 L_{TF} 在每批数据中的均值与标准差。自适应焦点损失函数首先对 TF-Loss 进行批归一化处理，将损失值调整至 0 到 1 的区间内，而后通过将 $\pi/2$ 与归一化的 TF-Loss 相乘，并应用正弦函数获得对应的权重，权重的取值范围为 0 至 1 之间，且生成的权重与 TF-Loss 的值成单调关系，即 TF-Loss 的值越大生成的权重值也相应越大，之后将权重与 TF-Loss 相乘完成加权。自适应焦点损失函数通过为有较大损失值的样本，即困难样本赋予较大的权值，而非困难样本赋予较小的权值，使得模型训练时困难样本的重要性被加强。为使总体的损失函数最小，模型将更多地学习困难样本的信息，从而使得困难样本的性能有所提升。

5.4.2 两阶段训练策略

第 5.4.1 章节提出的自适应焦点损失函数能够使模型在训练阶段更多关注困难样本，然而如果直接使用该损失函数进行模型训练，可能使得模型只关注困难样本，而因较小的权值忽视非困难样本，使得训练完成的单通道个性化语音增强性能整体性能降低，这显然是不可行的。为权衡困难样本和非困难样本在训练阶段的重要性，本文提出一种两阶段训练策略。第一阶段使用式$(5-4)$中不加权的损失函数进行训练，且使用早停训练方法，当验证集上的损失在连续 6 轮训练中没有降低时，则停止第一阶段的训练。经过第一阶段的训练，模型更关注训练集中占主体的非困难样本，充分学习到非困难样本的信息，使模型在整个训练集上有较好的性能。在第二阶段使用式$(5-5)$的损失函数对第一阶

段训练完成的模型继续进行训练，第二阶段的训练将在 20 轮训练内完成，通过第二阶段的训练，缓解了模型在第一阶段训练中对非困难样本的偏向，使得最终训练完成的单通道个性化语音增强系统鲁棒性更高，能更好地处理复杂声学场景下的各类样本。

5.5 实验设计与分析

本章节对上述提出的动态声学补偿方法和自适应焦点训练机制进行实验验证，并对实验结果进行分析。

5.5.1 数据集与实验配置

针对单通道个性化语音增强任务，由于需要说话人的注册语音数据，我们选择使用 The 4th Deep Noise Suppression (DNS) Challenge track2[75]的数据集。其中包含来自共 3230 个说话人的近 760 小时干净语音数据，每个说话人至少有两分三十秒的数据，干净语音数据包含英语、法语、德语、意大利语、西班牙语、俄罗斯语共六种语音，六种语言的比例如表 5-1 所示:

表 5-1 The 4th DNS Challenge track2 干净语音数据集不同语言的分布情况

Language	Speakers
English	2064
Italian	14
Spanish	224
German	874
French	47
Russian	7

如表 5-1 所示，其中英语数据主要来源自 Vocalset[76]，LibriVox[77]和 VCTK 数据集，非英语数据来自 OpenSLR[78]等。The 4th DNS Challenge 提供了 181 小时噪声数据，主要来自 Audioset[79]，Freesound[80]和 DEMAND[81]数据库，共包含 150 个种类的噪声。混响数据包含 3076 条真实数据和 115000 条合成数据，主要来自 OpenSLR26[78]和 OpenSLR28[78][78]数据库。为了实验的便利性，我们选用其中 500 个说话人的数据，合成了 160 小时的训练数据集，信噪比范围为-5 至 20dB，混响时间为 0.3 秒至 1.3 秒，每条数据时长为 10 秒。训练集共包含三种数据：（1）包含目标说话人语音和背景噪声，该类数据合计 100 小时（2）包含目标说话人语音和干扰说话人语音，该类数据合计 40 小时（3）同时包含目

标说话人语音，背景噪声和干扰说话人语音的数据，该类数据合计 20 小时。测试数据同样也选用 DNS 数据集进行合成，且与合成训练集部分不重合，参数配置与合成训练集时相同，共生成 800 条测试样本，并将该测试集命名为 DNS-test，其中含目标说话人语音和背景噪声的数据 500 条，该子测试集称为 t-noise；包含目标说话人语音和干扰说话人语音 200 条，该子测试集称为 t-mix；同时包含目标说话人语音，背景噪声和干扰说话人语音的数据 100 条，该子测试集称为 t-nmix。训练集和 DNS-test 测试集中三种类型样本的比例根据 DNS 比赛测试集划分。以上所有数据集中样本对应的注册语音时长为两分三十秒，所有数据均为 8kHz 采样率采样。

实验使用 Hanning 窗对语音进行加窗操作，帧长为 32 毫秒，帧移为 8 毫秒。Batch size 大小设置为 32，使用 Adam 优化器对模型进行优化，初始学习率为 0.001，其余实验配置均与 sDPCCN 原论文中一致。

5.5.2 评价指标

本章节实验主要使用五种评价指标，除以上章节使用的 PESQ 外，还使用 SISNR，DNSMOS，MOS，困难样本比例对本文所提出的动态声学补偿方法和自适应焦点训练的有效性进行验证，下面对这四种评价指标进行简单介绍。

SISNR：全称为 scale invariant signal to noise ratio，即尺度不变的信噪比，同样也是一种需要参考的客观评价指标，其主要衡量目标信号和干扰分量的相对比例关系，计算公式如(5 − 2)所示，SISNR 得分范围为负无穷至正无穷。

DNSMOS P.835[75]：该评价指标是一种客观评价指标，其通过深度神经网络预测主观 MOS (ITU-T Rec.P.835)得分，由 DNS 比赛官方发布已训练完成的模型，该模型通过语音数据和相对应的 MOS 得分进行训练，训练完成的模型可直接进行打分，不需要参考语音，DNSMOS 得分由三部分组成：1) SIG 衡量说话人语音的质量 2) BAK 衡量背景噪声的去除程度 3) OVRL 衡量语音总体质量。以上三个得分范围均为 1 至 5 之间，得分越高表示由模型预测的质量越高。

MOS (ITU-T Rec.P.808)[82]：该评价指标是一种主观评价指标，打分人根据语音的主观听觉感受进行打分，得分范围为 1 到 5 之间，得分越高代表语音的主观听觉感受越好。

困难样本比例[83]（Hard Sample Rate，HSR），例如 HSR0%表示整个测试集中经过系统增强后信噪比仍低于 0dB 的样本占整个测试集的比例。

5.5.3 实验结果分析

首先在 DNS-test 测试集上对基线系统 sDPCCN 和本文提出的 TF-Loss 进行

验证，实验结果如表 5-1 所示。系统 1 是基线系统 sDPCCN 的性能，将其与原始测试集的各项得分对比可以发现 sDPCCN 系统将 SISNR 得分提升了 9.16，PESQ 得分提升了 1.05，DNSMOS 的三项指标分别变化了-0.01，0.08，0.02，这说明 sDPCCN 系统总体上能够较好地去除背景噪声和干扰说话人，且只对目标语音信号造成轻微的失真，是一个性能非常优秀的基线系统。对比系统 2 和系统 1 的得分，TF-Loss 的应用使得 PESQ 得分提升了 0.07，DNSMOS 的 OVRL 得分提升了 0.02，SISNR 得分基本没有变动。由于基线系统 sDPCCN 的训练损失函数是 negative SISNR，因此其 SISNR 得分会相对较高，添加频域均方误差损失函数后 SISNR 得分没有变化，而其他指标均有提升，足以证明在训练阶段计算损失时同时考虑时域和频域的信息的有效性。

表 5-1 在 DNS-test 测试集上，基线系统 sDPCCN 和 TF-Loss 的实验结果

ID	Model	SISNR	PESQ	DNSMOS(SIG/BAK/OVRL)
/	Noisy	5.95	2.16	3.80/3.27/3.20
1	sDPCCN	15.11	3.21	3.79/3.85/3.32
2	+TF-Loss	15.12	3.28	3.79/3.88/3.34

表 5-2 在 DNS-test 测试集上，动态声学补偿的实验结果

ID	Model	SISNR	PESQ	DNSMOS(SIG/BAK/OVRL)
/	Noisy	5.95	2.16	3.80/3.27/3.20
1	sDPCCN	15.11	3.21	3.79/3.85/3.32
2	+TF-Loss	15.12	3.28	3.79/3.88/3.34
3	+DAC(2/0)	15.75	3.33	3.79/3.89/3.35
4	+DAC(2/2)	15.88	3.36	3.79/3.91/3.36
5	+DAC(8/4)	15.81	3.35	3.77/3.91/3.35
6	+DAC(4/2)	15.96	3.38	3.80/3.92/3.37
7	+ DAC(UB)	16.15	3.41	3.81/3.94/3.39
8	+SS(4/0)	15.77	3.35	3.76/3.92/3.35
9	+MMSE-LSA	15.01	3.23	3.76/3.92/3.36

而后对提出的动态声学补偿方法进行实验验证，实验结果如表 5-2 所示，表中系统 3 至系统 9 均为在系统 2 的基础上应用不同的动态声学补偿方法。系统 3 至 7 是基于波形叠加的动态声学补偿，括号中表示截取测试语音的前 J 帧和最后 K 帧信号作为背景噪声，例如（2/2）表示截取前后各两帧信号。系统 7 中 UB 表示 Upper bound，由于训练数据和测试数据均是合成的，即所用的噪声数据是已知的，该系统直接将合成数据所用的完整噪声数据与注册语音叠加，用

以获得基于波形叠加的动态升学补偿方法的性能上限。系统 8 是基于谱减法的动态声学补偿方法，用于与基于波形叠加的方法进行对比。通过对比系统 3，4，5，6 的得分，系统 6 的性能最优，即截取前 4 帧信号和最后两帧信号作为背景噪声是最好的配置。对比系统 2 与系统 6，当基于波形叠加的声学补偿被应用时，各项得分均有明显提升，其中 SISNR 提升 0.84，PESQ 提升 0.10，DNSMOS 提升 0.01/0.04/0.03，这充分证明基于波形叠加的动态声学补偿能够很好地解决注册语音与测试语音声学环境不匹配地的问题。对比 6 与系统 7 可以发现截取前 4 帧和最后两帧作为背景噪声的基于波形叠加的动态声学补偿方法与该方法的上限性能差异不大，SISNR 相差 0.19，PESQ 相差 0.03，DNSMOS 相差 0.01/0.02/0.02，由于在实际测试时完整的噪声数据无法获得，因此可以认为将前 4 帧信号和最后 2 帧信号作为背景噪声已是非常优秀的处理方法。对比系统 2，6，8 的性能，基于谱减法的动态声学补偿方法也能在一定程度上改善声学环境不匹配问题，系统 8 相较系统 2 的 SISNR 提升 0.65，PESQ 提升 0.07，DNSMOS 提升-0.03/0.04/0.01，然而与系统 6 相比，两者性能仍有一定差距，通过 DNSMOS（SIG）得分可以发现，基于谱减法的动态声学补偿给目标语音造成了较严重的失真，这也是其总体性能上不如基于波形叠加的动态声学补偿方法的原因。

除谱减法外，本章节也尝试使用另一种传统语音增强方法 MMSE-LSA 消除声学环境不匹配问题，结果如表格 5-2 最后一行所示，可以看到其也会对语音造成严重失真，导致性能不佳。综上所示，两种动态声学补偿方法都能在一定程度消除声学环境不匹配，但是，通过去除测试语音中噪声实现消除声学环境不匹配的方法（如谱减法，MMSE-LSA）都会对语音造成不可逆失真，使得这类方法的性能受到限制，而通过波形叠加的方法则不会有语音失真的问题，能够很好地解决注册语音与测试语音的声学环境不匹配，之后实验中的动态声学补偿均为基于波形叠加的方法，且配置为（4/2）。

随后对 DNS-test 测试集中三种情况的样本进行详细分析，表 5-3 展示了基线系统 sDPCCN 和应用了 TF-Loss 和 DAC 之后的系统在三种不同样本组成的子测试集上的性能，表格中“/”前为基线系统 sDPCCN，“/”后为应用 TF-Loss 和 DAC 的系统，对比三个子测试集性能可以发现两个系统都在 t-noise 和 t-mix 测试集上性能优秀，而在 t-nmix 测试集上性能略差一些，这是由于 t-nmix 的样本相较另外两个子测试集更为复杂，同时包含背景噪声和干扰说话人使得目标语音很难被准确提取，且合成的训练集中该类样本占比相对较少。然而通过比较两个系统性能可以发现，TF-Loss 和 DAC 在 t-nmix 子测试集上带来的性能提升最为明显，这证明提出的这两项技术对于处理复杂环境下的单通道个性化语

音增强任务有很好的效果。

表 5-3 在 DNS-test 三个子测试集上的实验结果

Metrics	t-noise	t-mix	t-nmix
SISNR	16.43/17.36	14.74/15.32	9.26/10.08
PESQ	3.31/3.49	3.32/3.41	2.51/2.69
DNSMOS(OVRL)	3.34/3.38	3.33/3.35	3.22/3.30

接着对本文提出的自适应焦点训练方法进行实验验证，表 5-4 给出了在 DNS-test 测试集上，本文提出的三项技术对 sDPCCN 系统性能的提升效果，TF-Loss 和 DAC 的效果已在上述段落中论证，而添加了自适应焦点训练（AFT）技术后，SISNR，PESQ 和 DNSMOS 三项指标均没有明显变化，其有效性需要针对测试集中的困难样本的实验予以验证。

表 5-4 在 DNS-test 测试集上，自适应焦点训练的实验结果

ID	Model	SISNR	PESQ	DNSMOS(SIG/BAK/OVRL)
/	Noisy	5.95	2.16	3.80/3.27/3.20
1	sDPCCN	15.11	3.21	3.79/3.85/3.32
2	+TF-Loss	15.12	3.28	3.79/3.88/3.34
3	+DAC	15.96	3.38	3.80/3.92/3.37
4	+AFT	15.89	3.36	3.79/3.91/3.37

表格 5-5 给出了在 DNS-test 测试集上，TF-Loss，动态声学补偿和自适应焦点训练对困难样本比例（HSR0,HSR5,HSR10）的改善情况，DAC 和 TF-Loss 由于使得系统整体性能有较大提升，困难样本比例相应有所下降。而自适应焦点训练在对整体性能不产生明显影响的条件下，使得测试集的困难样本比例明显下降，其中 HSR0 下降 0.50%，HSR5 下降 1.25%，HSR10 下降 4.62%。困难样本比例的下降能够充分说明自适应焦点训练对于改善困难样本问题是十分有效的。

表 5-5 在 DNS-test 测试集上，各系统困难样本比例的实验结果

ID	Model	HSR0(%)	HSR5(%)	HSR10(%)
1	sDPCCN	1.38	5.75	18.13
2	+TF-Loss	1.50	5.38	16.63
3	+DAC	1.13	4.00	13.25
4	+AFT	0.63	2.75	8.63

图 5-6 展示了 DNS-test 测试集在经过应用自适应焦点训练机制前后的两个个性化语音增强系统增强后的样本的信噪比分布，图中蓝色样本点代表未使用自适应焦点训练的系统，而橙色样本点代表使用自适应焦点训练后的系统，通过对比不难发现，在整体信噪比水平上，两者非常接近，然而橙色样本点在纵坐标上更为集中，在 10dB 以下的范围内，橙色样本点明显少于蓝色样本点，这也说明了自适应焦点训练提升了个性化语音增强系统的鲁棒性，使其能更好地处理困难样本。

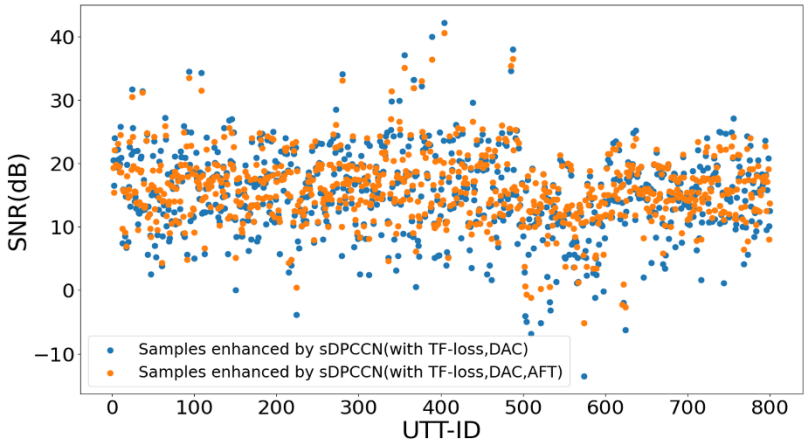


图 5-6 DNS-test 测试集经过个性化语音增强后的样本信噪比分布

表格 5-6 展示了不同系统在 DNS-test 测试集中的困难样本子集（经基线系统 sDPCCN 增强后信噪比仍小于 10dB 的样本集合）上的性能，在该实验中使用了主观 MOS 打分，对比表格中的各个系统，可以得出 TF-Loss，动态声学补偿，自适应焦点训练均可使得困难样本的性能有所提升，其中自适应焦点训练带来的性能提升尤为明显，各项指标中 SISNR 提升了 1.83dB，PESQ 得分提升了 0.16，DNSMOS（OVRL）得分提升 0.08，而主观得分 MOS 分提升了 0.18。综合以上的实验结果，可以充分说明自适应焦点训练对于改善困难样本性能，提升测试集中困难样本的性能下限有明显的效果，且对于整体的系统性能没有造成任何损失，是一项非常有效的技术。

表 5-6 在 DNS-test 困难样本子测试集上各系统的实验结果

ID	Model	SISNR	PESQ	DNSMOS(OVRL)	MOS
1	sDPCCN	5.93	2.12	3.19	3.28
2	+TF-Loss	6.01	2.18	3.23	3.35
3	+DAC	7.22	2.33	3.31	3.49
4	+AFT	9.05	2.49	3.39	3.67

5.6 本章小结

本章节介绍了针对单通道个性化语音增强任务，基于 sDPCCN 系统提出的 TF-Loss，动态声学补偿方法和自适应焦点训练方法。首先对单通道个性化语音增强的原理进行介绍，并对该任务的基线系统 sDPCCN 的技术进行详细阐述。随后介绍了结合时域与频域信息的 TF-Loss，并提出了基于波形叠加的动态声学补偿方法和基于谱减法的动态声学补偿方法，分别通过为注册语音添加噪声和为测试语音去除噪声的方法解决注册语音与测试语音声学环境不匹配的问题。为了解决个性化语音增强任务中的困难样本问题，提出自适应焦点训练机制，其通过为 TF-Loss 使用正弦函数加权使得模型训练时更关注困难样本，并将训练分为两个阶段权衡困难样本和非困难样本的重要性，在不降低系统总体性能的条件下提升了困难样本性能。最后在 DNS-test 测试集上的实验结果表明本章节提出的方法均有效，TF-Loss 和动态声学补偿方法使得个性化语音增强性能明显提升，且验证了基于波形叠加的动态声学补偿方法优于基于谱减法的方法。自适应焦点训练能够大幅提升困难样本性能，提升模型的鲁棒性。

第 6 章 总结与展望

6.1 本文总结

本文对单通道的实时语音增强和个性化语音增强任务展开研究。针对单通道实时语音增强任务，基于 PercepNet 系统从相位感知结构，多任务学习策略，后处理方法等方面展开详细研究。针对单通道个性化语音增强任务，基于 sDPCCN 系统从声学环境不匹配问题和困难样本问题等方面展开深入研究，并提出相对应的解决方法。本文的主要内容如下：

（1）首先介绍了单通道实时语音增强任务的目的以及目前基本的建模方式，接着对该任务的基线系统 RNNoise 和 PercepNet 的技术从声学特征，网络结构，基音滤波器，后处理技术等方面进行详细阐述。然后介绍了复现 PercepNet 实验所用数据集和包括 PESQ，STOI 在内的评价指标，最后给出了在 VCTK 测试集上的复现性能，实验结果表明，复现的 PercepNet 和原论文的 PercepNet 系统在训练数据量存在明显差距的情况下，性能差异不大，因此可以认为 PercepNet 复现正确，之后章节针对单通道实时语音增强任务的研究将以复现的 PercepNet 为基线系统。

（2）基于 PercepNet 系统提出相位感知结构，包括相位感知的声学特征，通过在 PercepNet 原有特征的基础上，加入语音频谱的实部与虚部特征，使得神经网络能够间接地学习到语音的相位信息。为进一步提升经系统增强后的语音质量，本文提出将 PercepNet 中只增强幅度的掩蔽替换为实虚部掩蔽使得系统能够同时恢复出干净语音的幅度与相位。进一步地，提出 TF-GRU 网络模块，通过矩阵转置的操作使得 GRU 层分别对语音的时间相关性和频率相关性建模。更好地利用提取的声学特征信息。最后，给出在 VCTK 测试集上的相关实验，通过与基线系统 PercepNet 比较，证明了提出的相位感知结构的有效性。

（3）基于 PercepNet 系统提出多任务学习策略与后处理方法。针对语音增强任务中存在的过度抑制问题，添加信噪比估计任务，使用多任务学习策略，使得模型学习到语音信噪比的信息，并为实虚部掩蔽添加额外的惩罚项，用于训练时惩罚过度抑制情况的产生。除此之外，针对高信噪比语音数据的严重过度抑制，本文创新性地提出基于信噪比估计的后处理方法，由于高信噪比的语音通常不需要后处理模块去除残留噪声，反而会由于后处理造成严重的过度抑制，因此使用神经网络预测的信噪比与设定的阈值进行比较，判断后处理模块是否被执行，从而缓解高信噪比语音的过度抑制。最后，给出在 VCTK 测试集

上的相关实验，实验结果表明，信噪比估计任务，过度抑制损失函数和基于信噪比的后处理方法均能有效解决过度抑制问题，其中基于信噪比的后处理方法对于高信噪比语音的过度抑制现象的改善尤为明显。

（4）针对单通道个性化语音增强任务，首先介绍了个性化语音增强的建模方法以及该任务的基线系统 sDPCCN 的技术原理。本文提出使用时域和频域相结合的损失函数使模型在训练阶段考虑多个域的信息，提升模型性能。然后，为解决该任务中存在的注册语音与测试语音声学环境不匹配的问题，提出了基于波形叠加的动态声学补偿方法和基于谱减法的动态声学补偿方法，分别通过为注册语音添加噪声和为测试语音去除噪声缓解声学环境不匹配。接着为了解决个性化语音增强任务中的困难样本问题，提出自适应焦点训练机制，通过两阶段的训练机制和自适应的加权损失函数，使得模型在训练阶段对非困难样本和困难样本的关注达到平衡。最后，给出在 DNS-test 测试集上的相关实验，实验结果说明了提出的 TF-Loss，动态声学补偿以及自适应焦点训练均能有效提升系统性能或解决相应的问题，且基于波形叠加的动态声学补偿由于不会对语音造成失真，效果会优于基于谱减法的动态声学补偿。

6.2 研究展望

本文主要研究了单通道的实时语音增强与个性化语音增强的建模方法，从相位感知的声学特征，实虚部掩蔽，TF-GRU 网络模块，多任务学习策略和基于信噪比估计的后处理技术等方面提升单通道实时语音增强性能。此外，为解决个性化语音增强任务中的声学环境不匹配问题和困难样本问题，提出了动态声学补偿方法和自适应焦点训练策略。但随着应用需求的多变化与技术的进步，单通道实时语音增强和单通道个性化语音增强仍然有许多问题需要解决。

（1）可用数据不足与数据集复杂性问题。目前的许多应用场景中为提高语音的听感质量，都使用 48kHz 采样率采样的语音，然而目前 48kHz 采样率的开源数据集相对较少，大多数据集都是 16kHz 采样率的。数据的匮乏导致建立一个适用于不同环境的鲁棒性全频带单通道实时语音增强系统是困难的。使用数据增强技术增加训练数据的多样性是一个比较实用的解决方法，但目前的数据增强方法仍比较单一，后续工作将从如何设计更符合现实场景的多样性数据增强方法展开研究。

（2）个性化语音增强任务中的注册语音问题。本文的实验中所有说话人均有两分三十秒的注册语音，然而现实应用场景中，注册语音的获取情况是相对随机的，一个说话人的注册语音可能只有十几秒甚至几秒，也可能一个说话人有多条的注册语音。所以如何建立一个对于不同时长的注册语音更鲁棒的个性

化语音增强系统，以及如何更好地利用可能存在的多条注册语音是未来需要研究的一个方向。

参考文献

[1] Puder H . Speech enhancement for hands-free car phone by adaptive compensation of harmonic eninge noise components[C]// European Conference on Speech Communication & Technology. DBLP, 2003.

[2] Meyer J , Simmer K U , Kammeyer K D . Multi-channel speech enhancement in a car environment using Wiener filtering and spectral subtraction[C]// IEEE International Conference on Acoustics. IEEE, 1997.

[3] Han S , Hong J , Jeong S , et al. Robust GSC-based speech enhancement for human machine interface[J]. IEEE Transactions on Consumer Electronics, 2010, 56(2):965-970.

[4] 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, 23.

[5] Wang Y , Wang D L . A structure-preserving training target for supervised speech separation[C]// IEEE International Conference on Acoustics. IEEE, 2014.

[6] Schrter H , Escalante-B. A N , Rosenkranz T , et al. DeepFilterNet: A Low Complexity Speech Enhancement Framework for Full-Band Audio based on Deep Filtering[J]. arXiv e-prints, 2021.

[7] Schrter H , Escalante-B. A N , Rosenkranz T , et al. DeepFilterNet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio[J]. 2022.

[8] Ding H , Soon I Y , Chai K Y . Over-Attenuated Components Regeneration for Speech Enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2010, 18(8):2004-2014.

[9] Boll S F . Suppression of acoustic noise in speech using spectral subtraction[J]. Acoustics Speech & Signal Processing IEEE Transactions on, 1979, 27(2):113-120.

[10] Lim, J. S , Oppenheim, et al. Enhancement and bandwidth compression of noisy speech[J]. Proceedings of the IEEE, 1979.

[11] Ephraim Y , Malah D . Speech Enhancement Using Spectral Amplitude Estimation[J]. IEEE Transactions on Acoustics Speech & Signal Processing, 2003, 33(2):443-445.

[12] Park S R , Lee J . A Fully Convolutional Neural Network for Speech Enhancement:, 10.21437/Interspeech.2017-1465[P]. 2016.

[13] Kolboek M , Tan Z H , Jensen J . Speech enhancement using Long Short-Term Memory based recurrent Neural Networks for noise robust Speaker Verification[C]// 2016 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2016.

[14] Zhang Q , Nicolson A , Wang M . Monaural Speech Enhancement Using Deep Multi-Branch Residual Network with 1-D Causal Dilated Convolutions[J]. 2019.

[15] Strake M , Defraene B , Fluyt K , et al. INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising[C]// Interspeech 2020. 2020.

[16] Wang Y , Wang D . Towards Scaling Up Classification-Based Speech Separation[J]. IEEE Transactions on Audio Speech & Language Processing, 2013, 21(7):1381-1390.

[17] Narayanan A , Wang D L . Ideal ratio mask estimation using deep neural networks for robust speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2013.

[18] Zhao Y , Wang D L , Merks I , et al. DNN-based enhancement of noisy and reverberant speech[C]// 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2016.

[19] Williamson D S , Wang Y , Wang D L . Complex ratio masking for joint enhancement of magnitude and phase[C]// IEEE International Conference on Acoustics. IEEE, 2016.

[20] Lu X G , Tsao Y , Matsuda S , et al. Speech Enhancement Based on Deep Denoising Autoencoder[C]// Conference of the International Speech Communication Association. ISCA, 2013.

[21] Tan K , Wang D L . Complex Spectral Mapping with a Convolutional Recurrent Network for Monaural Speech Enhancement[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[22] Hu Y , Liu Y , Lv S , et al. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement:, 10.21437/Interspeech.2020-2537[P]. 2020.

[23] Lv S , Fu Y , Xing M , et al. S-DCCRN: Super Wide Band DCCRN with learnable complex feature for speech enhancement[J]. 2021.

[24] Le X , Chen H , Chen K , et al. DPCRN: Dual-Path Convolution Recurrent Network for Single Channel Speech Enhancement[J]. 2021.

[25] Ronneberger O， Fischer P， Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer, Cham, 2015.

[26] 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). IEEE, 2018.

[27] Valin J M， Isik U， Phansalkar N , et al. A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech:, 10.21437/Interspeech.2020-2730[P]. 2020.

[28] Ji X， Yu M， Zhang C , et al. Speaker-Aware Target Speaker Enhancement by Jointly Learning with Speaker Embedding Extraction[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.

[29] Giri R， Venkataramani S， Valin J M , et al. Personalized PercepNet: Real-Time, Low-Complexity Target Voice Separation and Enhancement:, 10.21437/Interspeech.2021-694[P]. 2021.

[30] Eskimez S E， Yoshioka T， Wang H , et al. Personalized Speech Enhancement: New Models and Comprehensive Evaluation[J]. arXiv e-prints, 2021.

[31] Wang Q， Muckenhirn H， Wilson K , et al. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking:, 10.48550/arXiv.1810.04826[P]. 2018.

[32] Wang Q， Moreno I L， Saglam M , et al. VoiceFilter-Lite: Streaming Targeted Voice Separation for On-Device Speech Recognition:, 10.21437/Interspeech.2020-1193[P]. 2020.

[33] Han J， Long Y， Burget L , et al. DPCCN: Densely-Connected Pyramid Complex Convolutional Network for Robust Speech Separation And Extraction[C]// 2021.

[34] Valentini-Botinhao C， Xin W， Takaki S , et al. Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech[C]// 9th ISCA Speech Synthesis Workshop. 2016.

[35] 徐勇.基于深度神经网络的语音增强方法研究[D].中国科学技术大学,2015.

[36] 屠彦辉.复杂场景下基于深度学习的鲁棒性语音识别的研究[D].电子科技大学,2019.

[37] Foundation X O . Vorbis I specification.

[38] Owens E. Introduction to the Psychology of Hearing[J]. Archives of Otolaryngology, 1977(103-12).

[39]Cho K， Merrienboer B V， Bahdanau D , et al. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches[J]. Computer Science, 2014.

[40]Talkin D . Speech coding and synthesis. 1995.

[41]Moore B . An Introduction to the Psychology of Hearing: Fifth Edition. 2012.

[42]McGill TSP speech database http://www-mmsp.ece.mcgill.ca/Documents/Data/

[43]RNNoise demo https://jmvalin.ca/demo/rnnoise/

[44]Ko T， Peddinti V， Povey D , et al. A study on data augmentation of reverberant speech for robust speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2017.

[45]Kingma D P， Ba J . Adam: A Method for Stochastic Optimization[J]. arXiv e-prints, 2014.

[46]Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs[J]. 2007.

[47]Taal C H， Hendriks R C， Heusdens R , et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]// 2010 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2010.

[48]Loizou P C . Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum[J]. IEEE Transactions on Speech and Audio Processing, 2005, 13(5):857-869.

[49]Soon I Y， Koh S N， Chai K Y . Selective magnitude subtraction for speech enhancement[C]// International Conference on High-performance Computing in the Asia-pacific Region. IEEE Computer Society, 2000.

[50]Mowlaee P， Saeidi R . Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement[J]. IEEE Signal Processing Letters, 2013, 20(12):1235-1239.

[51]Nsa B， Mik B . Unsupervised single-channel speech enhancement based on phase aware time-frequency mask estimation[J]. Applied Speech Processing, 2021:75-99.

[52]Ge X， Han J， Long Y , et al. PercepNet+: A Phase and SNR Aware PercepNet for Real-Time Speech Enhancement[J]. 2022.

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

[54]Mowlaee P， Saeidi R . Iterative Closed-Loop Phase-Aware Single-Channel Speech Enhancement[J]. IEEE Signal Processing Letters, 2013, 20(12):1235-1239.

[55]Thoidis I， Vrysis L， Markou D K , et al. Temporal Auditory Coding Features for Causal Speech Enhancement[J]. Electronics, 2020, 2020(9(10)):1698.

[56] Saleem N , Gao J , Khattak M I , et al. DeepResGRU: Residual gated recurrent neural network-augmented Kalman filtering for speech enhancement and recognition[J]. Knowledge-based systems, 2022(Feb.28):238.

[57] Paul D B , Baker J M . The design for the wall street journal-based CSR corpus[C]// The Second International Conference on Spoken Language Processing, ICSLP 1992, Banff, Alberta, Canada, October 13-16, 1992. Association for Computational Linguistics, 1992.

[58] Ding H , Soon I Y , Chai K Y . Over-Attenuated Components Regeneration for Speech Enhancement[J]. IEEE Transactions on Audio Speech & Language Processing, 2010, 18(8):2004-2014.

[59] Ding H , Soon I Y , Koh S N , et al. A post-processing technique for regeneration of over-attenuated speech[C]// IEEE International Conference on Acoustics. IEEE, 2009.

[60] Li A , Liu W , Luo X , et al. ICASSP 2021 Deep Noise Suppression Challenge: Decoupling Magnitude and Phase Optimization with a Two-Stage Deep Network:, 10.1109/ICASSP39728.2021.9414062[P]. 2021.

[61] Tu Y H , Du J , Sun L , et al. LSTM-based iterative mask estimation and post-processing for multi-channel speech enhancement[C]// 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). 2017.

[62] Lv S , Hu Y , Zhang S , et al. DCCRN+: Channel-wise Subband DCCRN with SNR Estimation for Speech Enhancement:, 10.48550/arXiv.2106.08672[P]. 2021.

[63] Pei C Y , Nordholm S , Hai H D . Optimization and evaluation of sigmoid function with a priori SNR estimate for real-time speech enhancement[J]. Speech Communication, 2013, 55(2):358-376.

[64] Ephraim, Y, Malah, et al. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator[J]. Acoustics Speech & Signal Processing IEEE Transactions on, 1985.

[65] L. Chen et al. Multi-Stage and Multi-Loss Training for Fullband Non-Personalized and Personalized Speech Enhancement[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.

[66] Y. Ju et al. TEA-PSE: Tencent-Ethereal-Audio-Lab Personalized Speech Enhancement System for ICASSP 2022 DNS Challenge[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

IEEE, 2022.

[67] Thakker M , Eskimez S E , Yoshioka T , et al. Fast Real-time Personalized Speech Enhancement: End-to-End Enhancement Network (E3Net) and Knowledge Distillation[J]. 2022.

[68] Huang G , Liu Z , Laurens V , et al. Densely Connected Convolutional Networks: IEEE Computer Society, 10.1109/CVPR.2017.243[P]. 2016.

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

[70] Zhao H , Shi J , Qi X , et al. Pyramid Scene Parsing Network[C]// IEEE Computer Society. IEEE Computer Society, 2016.

[71] Roux J L , Wisdom S , Erdogan H , et al. SDR – Half-baked or Well Done?[C]// ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019.

[72] Deng C , Ma S , Zhang Y , et al. Robust Speaker Extraction Network Based on Iterative Refined Adaptation:, 10.48550/arXiv.2011.02102[P]. 2020.

[73] Lin T Y , Goyal P , Girshick R , et al. Focal Loss for Dense Object Detection[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, PP(99):2999-3007.

[74] Dong Q , Gong S , Zhu X . Class Rectification Hard Mining for Imbalanced Deep Learning[J]. IEEE, 2017.

[75] Reddy C , Dubey H , Gopal V , et al. ICASSP 2021 Deep Noise Suppression Challenge:, 10.48550/arXiv.2009.06122[P]. 2020.

[76] Wilkins J , Seetharaman P , Wahl A , et al. VocalSet: A Singing Voice Dataset[C]// International Symposium/Conference on Music Information Retrieval. International Society for Music Information Retrieval, 2018.

[77] Librivox https://librivox.org/

[78] Ko T , Peddinti V , Povey D , et al. A study on data augmentation of reverberant speech for robust speech recognition[C]// IEEE International Conference on Acoustics. IEEE, 2017.

[79] Gemmeke J F , Ellis D , Freedman D , et al. Audio Set: An ontology and human-labeled dataset for audio events[C]// IEEE International Conference on Acoustics. IEEE, 2017.

[80] Freesound https://freesound.org/

[81] Joachim, Thiemann, Nobutaka, et al. The diverse environments multi-channel

acoustic noise database: A database of multichannel environmental noise recordings.[J]. The Journal of the Acoustical Society of America, 2013.

[82] Naderi B , Cutler R . Subjective Evaluation of Noise Suppression Algorithms in Crowdsourcing:, 10.48550/arXiv.2010.13200[P]. 2020.

[83] K. Wang, Y. Peng, H. Huang, Y. Hu, and S. Li. Mining hard samples locally and globally for improved speech separation[C]// 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022.