2025 年 第 25 卷 第 20 期
2025, 25(20): 08331-16

科 学 技 术 与 工 程
Science Technology and Engineering

ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2404954
引用格式:王华朋, 冯嘉琪. 基于深度学习的语音增强方法综述[J]. 科学技术与工程, 2025, 25(20): 8331-8346.
Wang Huapeng, Feng Jiaqi. Review of speech enhancement methods based on deep learning[J]. Science Technology and Engineering, 2025, 25(20): 8331-8346.

自动化技术、计算机技术

# 基于深度学习的语音增强方法综述

王华朋, 冯嘉琪*
(中国刑事警察学院公安信息技术与情报学院, 沈阳 110854)

**摘 要** 随着深度学习技术的兴起,基于深度学习的语音增强方法日益广泛应用,性能普遍优于传统方法。概述语音增强中降噪信号处理的基本框架,逐步分析深度学习驱动的语音增强模型的最新进展。对基于深度学习的语音增强算法进行全面整理,详细介绍不同神经网络的语音增强方法的原理、特点、评价指标及代表性研究,综合评估这些方法的优势与不足。最后,结合当前发展状况,分析语音增强过程中面临的核心挑战,并对未来发展路径进行讨论与预测。

**关键词** 语音增强; 深度学习; 语音降噪; 神经网络
**中图法分类号** TP391; **文献标志码** A

# Review of Speech Enhancement Methods Based on Deep Learning

WANG Hua-peng, FENG Jia-qi*
(College of Public Security Information Technology and Intelligence, Criminal Investigation Police University of China, Shenyang 110854, China)

[**Abstract**] With the emergence of deep learning technologies, speech enhancement methods based on deep learning have seen widespread application and generally surpass traditional approaches in performance. The fundamental framework of noise reduction signal processing in speech enhancement was outlined and progressively delved into the latest advancements in deep learning-driven speech enhancement models. A comprehensive organization of deep learning-based speech enhancement algorithms was provided, detailing the principles, characteristics, evaluation metrics, and representative studies of various neural network-based methods. The advantages and limitations of these approaches were thoroughly assessed. Finally, in light of the current developmental landscape, the core challenges encountered in the speech enhancement process were analyzed, and future developmental trajectories were discussed and predicted.

[**Keywords**] speech enhancement; deep learning; speech denoising; neural network

在现代通信和音频处理领域,语音增强技术扮演着至关重要的角色,其目的是利用技术手段抑制背景噪声的干扰,从而提高带噪声语音信号的质量和清晰度[1]。语音增强不仅在日常生活的噪声环境中发挥作用,对于助听器、自动语音识别[2]、会议系统等专业领域也具有重要意义。随着深度学习技术的快速发展,其在语音增强中的应用已经超越了传统的信号处理方法,展现出更加强大的性能和潜力。

传统的语音增强方法主要集中在信号处理技术方面,旨在改善在噪声环境中录制的语音信号的质量。主要包括谱减法[3]、维纳滤波法[4]、最小均方误差估计[5],线性预测编码[6]等。然而在面对复杂、未知或非平稳噪声环境时,这类方法往往难以准确地识别和分离噪声与纯净语音,最终可能导致语音输出出现失真。随着语音处理应用需求的增长,这些传统方法在复杂场景中的不足日益凸显,促使研究者寻求更为先进的解决方案。

近年来,随着深度学习的迅速发展,基于深度学习的语音增强方法得到了广泛的研究和探索,并

**收稿日期**: 2024-07-02; **修订日期**: 2025-04-11
**基金项目**: 国家重点研发计划(2017YFC0821000);司法部司法鉴定重点实验室(KF202117);中国刑事警察学院研究生创新能力提升项目(2024YCZD05)
**第一作者**: 王华朋(1979—),男,汉族,山东菏泽人,博士,教授。研究方向:说话人识别、深度学习、人工智能。E-mail:huapeng.wang@hotmail.com。
* **通信作者**: 冯嘉琪(2001—),女,汉族,河南新乡人,硕士研究生。研究方向:深度学习、语音检验。E-mail:18240668287@163.com。

在各种复杂场景下展现出良好的性能。受到语音识别[2]、说话人识别[7]等语音处理领域的启发,目前基于深度学习的语音增强研究主要集中在利用先进的神经网络架构、设计更有效的模型训练策略以及针对特定噪声环境的模型优化等方面。从早期基于循环神经网络(recurrent neural network, RNN)[8]的方法[如长短期记忆网络(long short-term memory, LSTM[9])]到最近受到关注的基于生成对抗网络(generative adversarial network, GAN)的方法[10](如CycleGAN[11]),基于注意力机制的方法(如Transformer[12])等,各种深度学习技术已经被应用于处理语音增强任务。这些方法解决了传统语音增强技术难以克服的问题(如非平稳噪声抑制、回声消除等)。前人开展了多种基于深度学习的语音增强相关研究工作,涵盖有监督学习、无监督学习、端到端模型、多任务学习等多种方法。这些研究不仅提高了语音信号的可懂度和自然度,也为语音增强方法带来了新的视角和挑战。

基于此,本文研究旨在全面综述深度学习在语音增强中的最新研究进展,系统分析现有方法的优势与不足,探讨其未来发展方向。从语音增强的基础知识和其重要性入手,然后深入分析深度学习技术在语音增强中的应用,包括但不限于卷积神经网络(convolutional neural networks, CNN)、RNN和Transformer模型。此外,将讨论当前主流方法面临的挑战,以及如何通过创新的深度学习架构和算法来解决这些问题,并展望未来的研究方向。

## 1 语音增强模型概述

在现实世界的环境中,语音信号很容易被噪声干扰。噪声(包括混响在内),可以分为平稳噪声(随时间不变)和非平稳噪声(随时间变化)。平稳噪声包括空调的嗡嗡声、电子设备的背景噪声等。非平稳噪声有街道噪声、火车噪声、喋喋不休的背景声(其他说话者的声音)以及乐器声等。语音与噪声在时间域的关系可表示为

$$y_t = s_t h_t + n_t \tag{1}$$

式(1)中:$s_t$为纯净的语音信号;$h_t$为房间脉冲响应;$n_t$为加性噪声;$y_t$为噪声语音,将$x_t = s_t h_t$作为目标语音,可以将式(1)重写为

$$y_t = x_t + n_t \tag{2}$$

式(2)中:$t$为时间索引;噪声语音信号$y_t = [y_1, y_2, \cdots, y_T]$,其中$T$为语音的长度。

基于深度学习的语音增强模型旨在通过从噪声语音中估计出清晰的语音信号来提高语音的质量和可懂度,模型的性能主要取决于以下几个重要因素:输入数据、特征提取、训练目标、深度学习模型、评估指标。

### 1.1 数据集

语音增强模型的性能受其训练数据的质量和数量影响显著,高质量的纯净语音和噪声样本有助于模型更好地捕捉语音和噪声之间的特征差异,从而提高增强效果。此外,数据集的多样性也是影响模型性能的重要因素,包含不同环境和背景下的语音和噪声样本可以提高模型的泛化能力,使其在各种复杂场景下表现更稳定。最后,对训练数据进行适当的预处理也至关重要。选择合适的采样频率、帧大小、帧重叠百分比和窗函数[13]等预处理步骤可以优化模型的训练过程[14],提高其性能和稳定性。因此,在设计和选择训练数据集时,需要综合考虑数据质量、多样性、规模和预处理等因素,以确保模型能够取得理想的增强效果。表1对这些数据集进行了概述。

**表1 数据集概述**

**Table 1 Review of datasets**

| 数据集类型 | 数据库 | 语言 | 规模 | 采样率/kHz |
|---|---|---|---|---|
| 纯净语音 | TIMIT | 英语 | 630位说话者 | 16.00 |
| | VoiceBank | 英语 | 110位说话者、44 h语音 | 48.00 |
| | LibriSpeech | 英语 | 1 000 h语音 | 16.00 |
| | WSJ0 | 英语 | 800 h | 16.00 |
| 噪声 | UrbanSound8K | — | 10种噪声、8 732条记录 | 22.05 |
| | Demand | — | 16通道、6种大环境噪声 | 48.00 |
| | Noise-92 | — | 15种噪声类型 | 19.98 |
| | ESC 50 | — | 50种环境噪声、2 000条记录 | 44.10 |
| 含噪语音 | CHiME3 | 英语 | 342 h语音、50 h嘈杂环境音频 | 16.00 |
| | VoiceBank + Demand | 英语 | 两个子集:28位说话者和56位说话者 | 48.00 |
| | AMI | 英语 | 100 h会议录音 | 16.00 |
| | DAPS | 英语 | 20位说话者 | 44.10 |
| | Aurora-2 | 英语 | 8 440条记录 | 8.00 |
| | NOISEX-92 | — | 8种噪声、1.4 G | 16.00 |

### 1.2 特征提取

特征提取在语音增强模型中扮演着关键角色,其目标是从原始语音信号中提取出最具代表性和有助于语音增强的特征表示。首先需要选择合适的时域或频域表示方式。时域表示通常采用波形形式,而频域表示则通过傅里叶变换将时域信号转换为频谱表示。图1显示了频域中的特征提取,使用短时傅里叶变换(short-time fourier transform, STFT)导出幅度谱和功率谱,用于提取非负矩阵分解(non-negative matrix factorization, NMF)[15]、梅尔

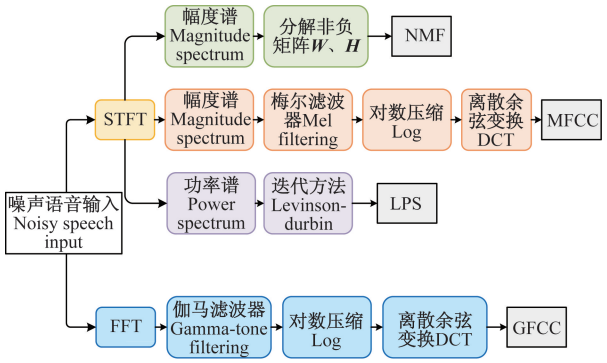



图 1　频域特征提取

Fig. 1　Features extraction in frequency domain

频率倒谱系数(Mel frequency cepstral coefficient, MFCC)[16]、对数功率谱(logarithmic power spectrum, LPS)[17]等。从快速傅里叶变换(fast Fourier transform,FFT)中提取伽马通频率倒谱系数(gammatone frequency cepstral coefficients, GFCC)[18]。此外,特征提取还包括对频谱图的后续处理,如归一化、对数变换等,这些处理有助于进一步改善特征的表示形式,提高模型对不同语音信号的鲁棒性和稳健性。

### 1.3　深度学习在语音增强中的优势

深度学习技术在语音增强领域的应用中已经显示出显著的优势,主要体现在如下几个方面。

(1)自动特征学习。深度学习模型,尤其是CNN和循环神经网络RNN,能够自动从原始数据中学习到有用的特征表示,减少了对专家设计特征的依赖。这一点在语音增强中尤为重要,因为语音信号的复杂性和多变性要求模型能够捕捉到丰富的时频特征[19]。

(2)处理非平稳噪声。与传统的基于信号处理的方法相比,深度学习模型能够更好地处理非平稳噪声,如街道噪声、人声嘈杂等。这是因为深度学习模型通过多层次的抽象能够捕捉到更复杂的模式和时间依赖性[20],从而在动态变化的噪声环境中实现有效的语音增强。

(3)端到端学习。深度学习允许端到端的训练策略,这意味着模型可以直接从带噪语音到干净语音的映射进行学习,而无需复杂的预处理和后处理步骤[21]。这种端到端的方法简化了语音增强的流程,并有助于提高整体性能。

(4)泛化能力。深度学习模型通常具有较好的泛化能力,能够在不同的噪声环境和录音条件下保持稳定的性能[22]。这得益于深度学习模型的参数数量多,能够学习到更多的数据分布信息。

(5)多任务学习。深度学习模型可以同时处理多个相关任务,如语音增强、说话人识别和语音识别。通过共享表示和学习策略,多任务学习可以提高各个任务的性能,并在任务之间实现知识迁移[23]。

近年来,深度学习驱动的语音增强技术也在多个方向上取得了显著进展。Transformer模型因其自注意力机制能够捕捉长时间依赖特性,在复杂噪声环境下表现出优越的鲁棒性,其中Conformer模型结合了卷积神经网络和Transformer的优势,在语音识别中取得了优异的性能[24];多模态语音增强方法通过结合语音与视觉信息,在极端噪声场景中显著提升了语音信号的可懂度,代表性的有SEANet(sound enhancement network)模型[25]利用加速度计数据,在嘈杂环境下实现了语音增强;针对实时语音通信的需求,文献[26]提出了多种轻量化模型,显著降低了计算复杂度,使语音增强在嵌入式设备和低资源环境中得以应用,如SpeechFormer模型通过层次化的Transformer结构,提高了模型的效率和性能;通过对抗训练、数据增强和注意力机制的引入,深度学习模型在处理非平稳噪声和极端噪声场景中展现出更好的泛化能力和增强效果,张天骐等[27]提出了复谱映射下融合高效Transformer的语音增强方法,结合卷积模块与高效Transformer,提升了模型在复杂噪声环境下的表现。上述研究成果不仅推动了语音增强技术的性能提升,也为应对复杂环境中的语音处理挑战提供了新的解决思路。

## 2　训练目标

在语音增强的研究领域中,训练目标的设定通常基于两种不同的方法:基于映射(Mapping)的方法和基于掩蔽(Masking)的方法[28]。映射方法的核心在于将带噪语音直接转换为干净语音,这一过程可被视为一个回归问题,其目标是学习一个函数,该函数能够最小化输入和期望输出之间的差异。相对地,基于掩蔽的方法则关注于估计一个掩蔽矩阵,该矩阵用于对带噪语音的频谱进行加权,以区分并抑制噪声成分,从而恢复出清晰的语音信号。在这种方法中,掩蔽矩阵的估计问题被构建为一个分类任务,其输出被用作滤波器,以实现对噪声的抑制和语音信号的增强[29]。这种基于掩蔽的方法在实现上通常涉及到对噪声和语音的统计特性进行建模,并利用这些模型来指导掩蔽矩阵的估计过程。

### 2.1　基于映射的方法

在语音增强领域,监督学习通常采用预测方法,旨在建立受损语音与理想干净语音目标之间的最佳确定性映射。传统语音增强技术一般通过统计优化手段,以最大化特定的目标函数,从而寻找

最优增益函数。在这一过程中,普遍存在一种假设,即干净语音和噪声语音的频谱分布是复杂的高斯分布,且它们之间是统计不相关的。基于这一假设,语音增强的目标是通过最小化干净频谱和估计增强频谱之间的均方误差(mean squared error,MSE)来实现。这种方法在语音增强中被广泛应用,因为它提供了一种直接且有效的方式来改善语音信号的质量。基于映射的方法通过映射一个非线性函数 $F$,将噪声语音 $y_t$ 转换为增强语音 $x_t$,映射的方法可表示为

$$y_t \xrightarrow{F} x_t \tag{3}$$

原始语音信号存在快速波动的问题,因此语音信号的幅度谱图通常使用基于映射的方法。幅度谱图由 STFT 和时间窗滤波器来构建,通过 STFT 逆运算,基于频谱相位信息将频谱图重建回原始带噪语音信号。然而,这种映射方法是在时域中使用的[30-32]。基于映射的方法对信噪比(signal-to-noise ratio,SNR)波动的敏感性较低[33],适用于不同 SNR 条件下的应用。

在基于映射的方法中,神经网络根据观察到的输入来预测输出。观察到的输入来自有噪声的语音 $y_t$,而目标输出来自纯净的语音 $x_t$。为了学习 $F$ 函数,神经网络通过最小化均方误差(mean squared error,MSE)损失 $L_{\text{MSE}}$ 来调整参数,MSE 损失由噪声语音和增强的噪声语音计算得到。如。训练网络 $F$ 从带噪信号 $\boldsymbol{Y}$ 映射到干净信号 $\boldsymbol{X}$。

$$L_{\text{MSE}} = \| \boldsymbol{Y} - F(\boldsymbol{X}) \|^2 \tag{4}$$

平均绝对误差(mean absolute error,MAE)也用来表示损失,MAE 损失由原始语音信号和增强语音信号计算得到,计算公式为

$$L_{\text{MAE}} = \| \boldsymbol{Y} - F(\boldsymbol{X}) \| \tag{5}$$

式中:$\|\cdot\|^2$ 为平方损失;$\|\cdot\|$ 为绝对损失。

### 2.2 基于掩蔽的方法

语音增强的核心技术之一在于对语音信号的时频(TF 域)表示进行信噪比估计,并对噪声部分强度高于语音的部分进行抑制。该技术遵循听觉感知中的掩蔽效应原理,即在时间或频率上相邻的强信号会干扰弱信号的听觉感知[34]。内耳耳蜗中的听觉神经元沿基底膜分布不均导致信号重叠,进而干扰了向大脑传递的神经信息。如果能够有效地抑制那些干扰性噪声的 TF 单元,同时保留那些在语音信号中谐波分量占主导地位的 TF 单元,则经过处理的语音质量通常会得到显著提升。

基于掩蔽的方法的主要目标是从噪声语音信号中估计语音,该方法使用幅度谱图(TF 域)实现,其中估计的掩码作为滤波器应用于输入谱图。理想二进制掩码(ideal binary mask,IBM)和理想比率掩码(ideal ratio mask,IRM)是最常用的两种掩码。在 IBM 中,语音幅值较高(局部 SNR 大于阈值 $R$)的频谱图的频率范围设为 1,噪声强度较高的频谱图的频率范围设为 0,如式(6)所示。

$$\text{IBM}(k,l) = \begin{cases} 1, & \text{SNR}(k,l) > R \\ 0, & \text{其他} \end{cases} \tag{6}$$

式(6)中:$l$ 为时域;$k$ 为频域;$R$ 为根据实验试验确定的阈值。

研究表明,在噪声信号的信噪比低于 5 dB 时,可以获得最佳结果[35]。

IRM 是软屏蔽方法之一[36],用于提高分离语音的可理解性和质量。IRM 使用 0 ~ 1 的概率,可表示为

$$\text{IRM}(k,l) = \left[ \frac{X(k,l)^2}{X(k,l)^2 + N(k,l)^2} \right]^{\beta} \tag{7}$$

式(7)中:$X(k,l)^2$ 和 $N(k,l)^2$ 分别为语音和噪声的能量;$\beta$ 为用于掩码缩放的可调参数,通常设为 0.5。

目标二进制掩码通过对比每个 TF 单元的目标语音能量与固定干扰来进行分类,并据此赋予二进制标签。考虑到相位信息对语音质量的重要性,也存在相位掩码[37]。复理想比掩码(complex ideal ratio mask,cIRM)是一种包含相位信息的掩码方法[38],但其无界性导致优化过程面临无限搜索空间的挑战。频谱幅度掩码(spectrum amplitude mask,SAM)则基于干净语音和噪声语音的 STFT 幅度来定义。其他掩码技术还包括理想幅度掩码(ideal amplitude mask,IAM)和相位敏感掩码(phase sensitive mask,PSM)[39]等。

## 3 语音增强的深度学习模型

用于解决语音增强问题的深度学习模型包括深度神经网络(deep neural networks,DNN)、RNN、LSTM、CNN、GAN、Transformer。

### 3.1 深度神经网络(DNN)

DNN 是人工神经网络研究概念的延伸,也被称为前馈全连接层(feed-forward neural networks,FCL)或多层感知器(multilayer perceptrons,MLP),是语音增强常用的架构,如图 2 所示。

第一层的每个神经元都通过输入向量 $\boldsymbol{x}(n) = [x_1, x_2, \cdots, x_n]$ 与多层感知机相连。第一层的输出作为第二层的输入,以此类推,每一个中间层的输出都会成为下一层的输入,最后一层输出层得到输出向量 $\boldsymbol{y}$。

该网络被称为全连接网络,是因为每层的每个节点与前一层的每个节点共享一个连接,因此,DNN具有非常大的参数。Karjol 等[40]提出了一种基于多个 DNN 系统的增强方案,该系统涉及 $N$ 个 DNN,每个 DNN 都有助于提高语音输出,采用门控网络提供权重对多个 DNN 的输出进行加权平均。该模型使用 $N=4$,每一个都是三层深度。在 TIMIT 语料库上,平均信噪比为 -5 ~ 10 dB 的情况下,可见噪声的平均语音质量感知评估(perceptual evaluation of speech quality,PESQ)为 2.65,不可见噪声的平均 PESQ 为 2.19。Zhao 等[41]通过将语音可理解度度量纳入损失函数来扩展 DNN 的功能。使用 IEEE 语料库和 NOISEX 库中的噪声,结果显示,匹配信噪比条件下 PESQ 为 2.215,不匹配信噪比条件下 PESQ 为 2.230。Kawanaka 等[42]在基于函数逼近的方法中使用稳定技术,设计了一个新的基于 PESQ 的代价函数,采用自强化学习领域的稳定化技术,稳定地训练 DNN,在公共数据集上 PESQ 为 2.93,同时比传统方法获得了更好的主观质量。Fujimura 等[43]提出了一种不需要干净信号的训练策略,直接利用噪声信号作为目标训练 DNN。在训练数据集和测试数据集不匹配的情况下,取得了与使用干净信号作为目标相似的结果。Furnon 等[44]引入并扩展了一种基于 DNN 的分布式时频掩模语音增强方法,在空间不受约束的麦克风阵列中运行。通过利用节点间的合作和压缩信号的声学多样性,在多种实际声学环境中表现出了显著的噪声抑制和语音质量提升,在低信噪比条件下显示出了更高的鲁棒性。

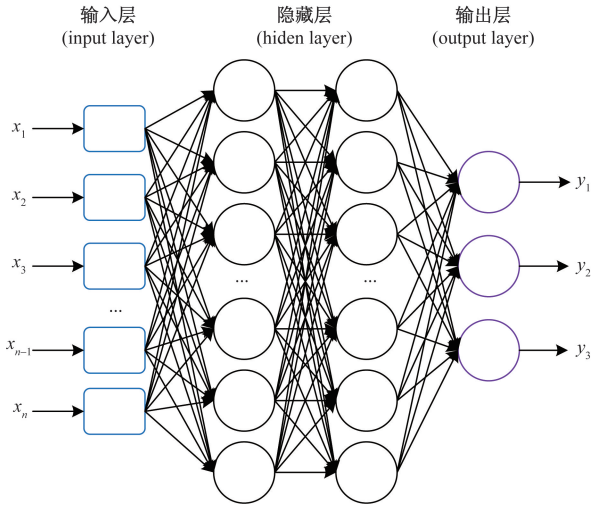



图 2　深度神经网络(DNN)模型

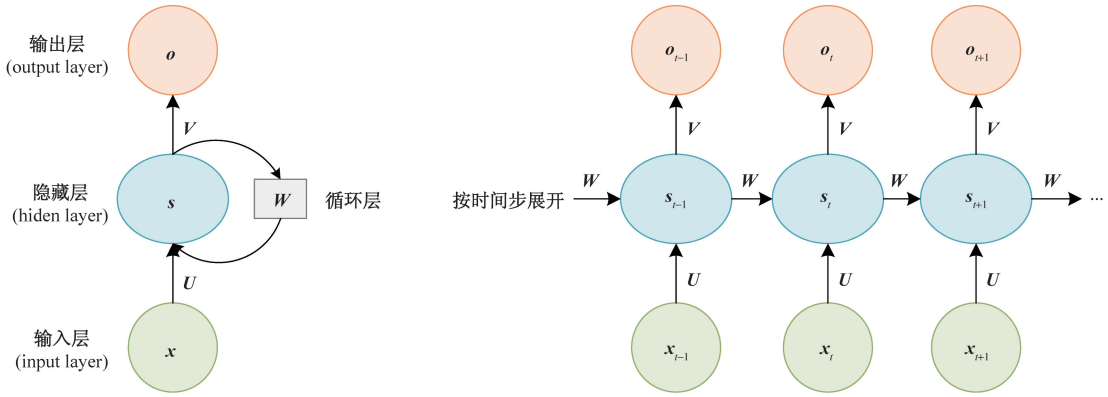
Fig. 2　Deep neural network (DNN) model

上述研究表明,基于 DNN 的语音增强方法能够通过不同的学习策略和网络架构设计,有效地改善在噪声环境下的语音质量。而且通过多目标学习[45]和感知激励的损失函数[41],增强了模型对语音特征的理解和处理能力。Noisy-target Training[43]的提出为缺乏干净语音数据的场景提供了新的解决方案,而稳定化技术的应用[42]则为提高 DNN 训练的稳定性和可靠性提供了有效途径。

### 3.2　循环神经网络(RNN)

RNN 具备处理变长语音信号并捕捉时间动态上下文的能力,从而优化语音增强效果,在接收噪声语音输入时,RNN 同时生成增强后的语音。典型的 RNN 采用 Elman 结构,如图 3 所示,其中包含从过去状态到当前状态的递归连接。

Arisoy 等[46]探索了将双向循环神经网络(bidirectional recurrent neural network,BRNN)和 LSTM 应用于语音识别中的语言模型构建,并针对英语广播新闻转录任务进行了评估,结果显示,双向 RNN 在词




$\boldsymbol{x}$ 为输入向量;$\boldsymbol{o}$ 为输出向量。输入向量 $\boldsymbol{x}_t=[x_1,x_2,\cdots,x_T]$ 为当前时间步的输入序列,其和输出向量 $\boldsymbol{o}_t=[o_1,o_2,\cdots,o_T]$ 的维度根据特征和输出需求而异;$\boldsymbol{s}$ 为在特定时间步的隐藏状态,是 RNN 能够处理序列数据的关键;$\boldsymbol{s}_t$ 为根据前一隐藏状态 $\boldsymbol{s}_{t-1}$ 确定的当前时间步的状态;$\boldsymbol{W}$、$\boldsymbol{U}$ 和 $\boldsymbol{V}$ 为权重矩阵

图 3　循环神经网络(RNN)模型

Fig. 3　Recurrent neural network (RNN) model

错误率(word error rate,WER)上相较于单向 RNN 取得了显著改进,但双向 LSTM 并未在单向 LSTM 的基础上带来额外的性能提升。Takeuchi 等[8]提出了一种基于均衡循环神经网络(equilibriated recurrent neural network,ERNN)的实时语音增强方法,该方法通过引入迭代训练机制,显著减少了模型参数数量,同时避免了梯度消失问题,从而在保持性能的同时降低了计算复杂度。相较于传统的 LSTM,所提出的 ERNN 结构在参数数量上减少了 1/5,在多项语音增强指标上展现出与 LSTM 相媲美的性能。Tan 等[47]提出了卷积循环神经网络(convolutional recurrent neural network,CRN),结合卷积编码器-解码器(convolutional encoder-decoder,CED)和 LSTM 层,用于实现噪声和说话者独立的实时单声道语音增强任务。在短时客观可懂度(short-time objective intelligibility,STOI)和 PESQ 评分方面均优于当时的基于 LSTM 的模型,并且具有更少的可训练参数,显示出更高的参数效率和更快的收敛速度。

结合卷积网络与循环网络的模型在语音增强任务中得到了广泛应用。Wahab 等[48]提出了基于复杂频率域的紧凑神经网络模型,通过卷积编码-解码器捕获语音的时间频率特征,同时利用循环网络(如 LSTM、GRU、SRU)捕获时序依赖性。实验结果表明,该模型在资源受限设备上的实时语音增强任务中表现优异,在 STOI 和 PESQ 等指标上显著优于许多现有方法。

总体而言,这些研究通过不同的网络结构和训练策略,实现了对语音信号的有效增强。研究者们强调了模型对噪声环境泛化的适应性,在减少模型参数的同时保持了性能[8],而 CRN 模型[47]则在参数效率和实时处理方面取得了突破。它们都通过利用 RNN 的时序数据处理能力,提高了语音增强的性能和实用性,以实现实时、低延迟的语音增强目标,为未来在更广泛的应用场景中的研究和应用奠定了基础。

RNN 的梯度消失问题阻碍了对长期时间依赖型的建模,研究者们提出了几种专门的 RNN 架构,包括 LSTM 和门控循环单元(gated recurrent unit,GRU)等以解决这一问题。

### 3.3 长短时记忆网络(LSTM)

LSTM 的开发是为了克服 RNN 中的梯度消失问题。图 4 显示了 LSTM 架构,它使用门控制各层之间的信息流,这样梯度就能以相对稳定的方式长期传递。读取、存储和写入决策都基于激活函数,其输出是介于(0,1)的值。遗忘门和输出门决定是否存储或丢弃新信息。模型决策基于 LSTM 模块的

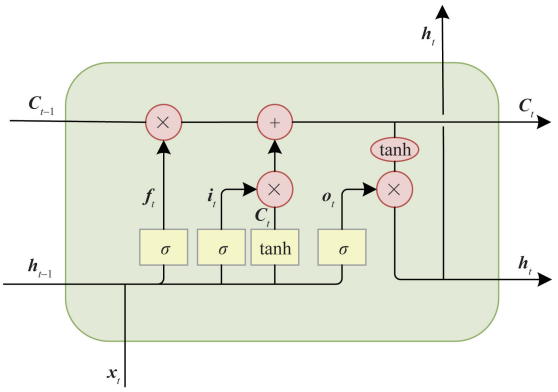



图 4 LSTM 模型

Fig. 4 LSTM model

内存和输出门的条件,然后将输出重新作为输入,创建循环序列。

LSTM 单元可以在时间步 $t$ 处使用式(8)计算。

$$\begin{cases} \boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t] + \boldsymbol{b}_f) \\ \boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t] + \boldsymbol{b}_i) \\ \tilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_C[\boldsymbol{h}_{t-1},\boldsymbol{x}_t] + \boldsymbol{b}_C) \\ \boldsymbol{C}_t = \boldsymbol{f}_t \odot \boldsymbol{C}_{t-1} + i_t \odot \tilde{\boldsymbol{C}}_t, \\ \boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t] + \boldsymbol{b}_o) \\ \boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh \boldsymbol{C}_t \end{cases} \tag{8}$$

式(8)中:$\boldsymbol{x}_t$ 为当前时间步 $t$ 的输入向量;$\boldsymbol{h}_{t-1}$ 为前一时间步 $t-1$ 的隐藏状态(短期状态)向量;$\boldsymbol{C}_{t-1}$ 为前一时间步 $t-1$ 的记忆单元(长期状态)向量;$\tilde{\boldsymbol{C}}_t$ 为候选向量,可理解为中间变量;$\boldsymbol{i}_t$、$\boldsymbol{o}_t$ $\boldsymbol{f}_t$ 分别为输入、输出和遗忘门的激活向量;$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_C$、$\boldsymbol{W}_o$ 分别对应遗忘门、输入门、候选状态、输出门的权重矩阵;$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_C$、$\boldsymbol{b}_o$ 分别为对应遗忘门、输入门、候选状态、输出门的偏置项;$\sigma$ 为 sigmoid 函数;$\odot$ 为向量之间的逐元素乘法。

Geiger 等[49]采用 LSTM 循环神经网络进行隐马尔可夫模型(hidden markov model,HMM)状态预测,提升了噪声环境下的语音识别鲁棒性。实验结果显示,相较于传统的音素预测方法,所提模型在 CHiME 挑战赛中达到更低的词错误率(WER)。Weninger 等[50]采用基于 LSTM 的 RNN 进行语音增强,并将其应用于提高噪声环境下自动语音识别(automatic speech recognition,ASR)的鲁棒性。通过判别性训练 LSTM 网络以优化语音重建目标,实现了在 CHiME-2 数据集上的 WER 显著降低,达到了当时最佳性能。Sun 等[51]通过对直接映射(LSTM-DM)、理想比例掩码(LSTM-IRM)和间接映射(LSTM-IM)比较分析,发现在低信噪比条件下 LSTM-DM 能更好地恢复语音可懂度,而 LSTM-IRM 在高信噪比条件下表现更佳。因此设计一种多目标联合学习框架,提高 PESQ 和 STOI。

Wang 等[52]提出了 LCLED(LSTM-convolutional-BLSTM encoder-decoder)网络,结合 LSTM 和 CNN 的优势,捕获语音信号的时间-频率特征,引入转置卷积和跳跃连接。与全 LSTM 结构相比,不仅降低了模型复杂度和训练时间,而且在多种噪声条件下均提高了增强语音的质量和可懂度。Oruh 等[53]提出了改进的 LSTM 模型,将 RNN 集成作为“遗忘门”到 LSTM 单元中,有效处理连续输入流并优化网络参数的使用。在公认的英语数字语音识别基准数据集上达到了 99.36% 的高准确率。

双向 LSTM(bi-directional long short-term, BiLSTM)的性能优于单向 LSTM。Graves 等[54]提出 BiLSTM 并进行了全梯度学习算法的改进。结果表明,BiLSTM 在 TIMIT 数据库上的表现优于标准 RNN 和 MLP,且训练速度更快、准确率更高,证实了 BLSTM 在处理需要丰富上下文信息的语音任务中的有效性。Zhang 等[55]提出了高速公路 LSTM(highway long short-term memory,HLSTM)和延迟驱动的 BiLSTM(latency-controlled bi-directional long short-term memory,LC-BiLSTM),在相邻层之间的记忆单元引入带门控的直接连接。在 AMI(augmented multi-party interaction corpus)远场语音识别(distant speech recognition,DSR)任务上取得了优于先前工作的性能,实现了更低的 WER。Xue 等[56]提出了改进的 LC-LSTM,通过采用不同类型的神经网络拓扑结构来初始化 BiLSTM 记忆单元状态,提高了解码速度并可用于实时语音增强。

最初 LSTM 被引入用于解决 RNN 中梯度消失的问题,通过引入记忆单元和控制信息流动的门控机制,LSTM 能够有效地捕捉长期依赖关系,这对语音增强和识别至关重要,然而在处理长序列数据时仍然面临梯度消失的挑战。双向 LSTM 被提出以克服 LSTM 的局限性,通过在两个方向上处理输入数据来同时利用过去和未来的上下文信息,在序列分类任务中表现出优越的性能,但其缺点是增加了计算复杂度。多目标联合学习和改进的网络拓扑结构结合不同学习目标的优势,可以提高模型的泛化能力,并提升效率和性能。

### 3.4　卷积神经网络(CNN)

CNN 采用网格状结构处理数据,在语音处理等领域受到了研究者的广泛关注。CNN 通过局部连接来捕获相邻帧中的模式,从而在特征提取方面表现出色。基于 CNN 的语音增强方法利用其在特征提取方面的卓越能力,通过学习噪声和干净语音的映射关系,有效地从噪声语音信号中分离出目标语音。CNN 的函数描述为

$$(y_k)_{ij} = (\boldsymbol{w}_k \otimes \boldsymbol{x}) + a_k \tag{9}$$

式(9)中:$y_k$ 为特征输出图;$\boldsymbol{x}$ 为特征输入图;$\boldsymbol{w}_k$ 和 $a_k$ 分别为滤波器的权重和偏置值;$\otimes$ 为二维卷积。

CNN 架构通常包含多个卷积层和池化层,用于提取语音信号的时间频率特征,并通过全连接层输出增强后的语音。

图 5 显示了典型的 CNN 架构,由 3 个基本层组成:卷积层、池化层和全连接层。卷积层由一组应用于输入的滤波器(即内核)组成,滤波器应用于输入时,可以生成特征图。通过堆叠卷积层,用户可以生成更复杂的模型,从图像中学习更详细的特征。池化层是深度学习中使用的卷积层类型之一,使用池化层时,输入会在空间上缩小,从而更容易处理,消耗的内存也更少,池化还可以减少参数数量,加快训练过程。池化可分为两种类型:最大池化和平均池化。在最大池化中,每个特征图都取最大值。平均池化则使用每个特征图的平均值。为了减少输入的大小,通常在卷积层之后添加池化层。全连接层将每个神经元与前几层中的其他神经元连接起来,作为 CNN 的最后一层,全连接层使用从前几层学到的特征进行预测。与 DNN 和 RNN 相比,CNN 更易于训练。

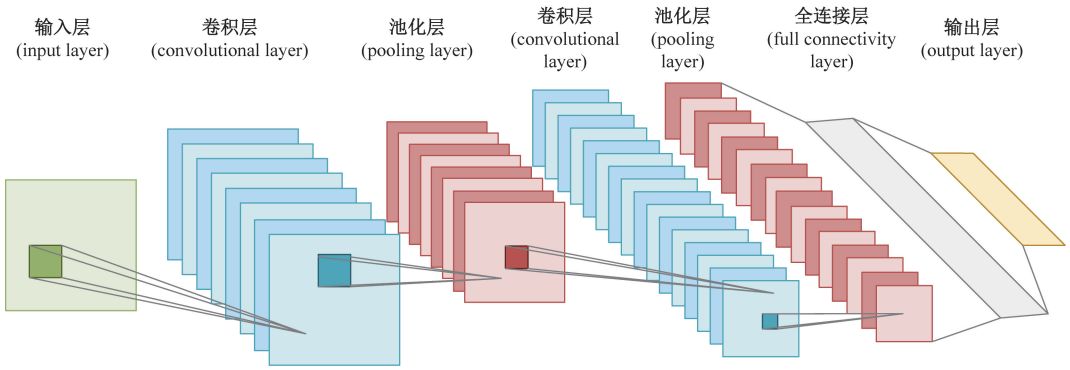



图 5　CNN 模型

Fig. 5　CNN model

Pandey 等[57]提出了用于时域内语音增强的全卷积神经网络,引入频域损失函数,利用 STFT 幅度的 MAE 损失来优化网络性能,在多个标准评价指标上显著优于当时的其他方法。Fu 等[58]提出了一种基于信噪比(signal-to-noise ratio,SNR)感知的 CNN 模型,通过采用多任务学习(multi task learning,MTL)框架和 SNR 自适应去噪策略,显著提升了在不同 SNR 水平下的去噪性能。相较于传统 DNN,模型在标准化客观评估中表现更优,尤其是在未见过的高 SNR 输入条件下,验证了其泛化能力和提高语音质量的潜力。Ouyang 等[59]通过结合一维和二维频率扩张卷积以及残差学习与跳跃连接结构,有效减少参数并提升性能。该 CNN 在相位估计方面表现出色,在处理女性语音时,能够显著提高重建语音的感知质量。Pandey 等[60]提出了一种时间卷积神经网络(temporal convolutional neural network,TCNN)用于时域内实时语音增强,在编码器和解码器之间插入时间卷积模块(temporal convolutional module,TCM),有效利用过去帧的信息。在说话人和噪声独立的情况下,相比于现有实时卷积循环模型展现出了更优的效果,在参数数量上有显著减少。

基于 U-Net 的语音增强网络因其对称的结构和较低的参数量,广泛用于语音增强任务。针对传统 U-Net 的深层特征提取能力不足和特征丢失问题,许春冬等[61]提出了结合残差和双注意力机制的 DA-Res-U-Net(dual attention residual U-Net)模型。该模型通过引入残差结构和双注意力机制增强深层特征提取能力,并在编码层和中间层引入 ASPP(atrous spatial pyramid pooling)模块,实现多尺度特征融合。相较于卷积循环神经网络(convolutional recurrent network,CRN)、深度复数 U-Net(deep complex U-Net,DCU-Net)和深度复卷积循环网络(deep complex convolutional recurrent network,DCCRN)等模型,该方法在 PESQ、STOI 和对数谱距离(log-spectral distance,LSD)指标上均有显著提升。

此外,针对传统卷积神经网络在复杂噪声场景中的局限性,徐浩森等[62]提出了一种结合通道注意力机制和循环网络的卷积循环神经网络 AR-CED,该方法通过为卷积核分配权重,有效提升了对多种噪声特性的处理能力,实验结果表明,其在 PESQ 和 STOI 等指标上均优于传统模型。

基于 CNN 的语音增强方法通过利用其在时频特征提取上的优势,实现了对噪声的有效抑制和语音质量的提升。早期研究集中在使用传统 CNN 架构,模型通常采用简单的卷积层和池化层来提取特征。全卷积网络(fully convolutional networks,FCN)的提出通过去除全连接层来降低模型的参数数量,同时保持对空间特征的捕捉能力。为了更好地处理序列数据,TCNN 通过使用因果和扩张卷积层来增加感受野,同时保持时间序列的因果关系。不同的研究通过引入如多任务学习、SNR 感知算法、复杂语谱图处理和时间卷积模块等技术,进一步提高了模型的性能和实时性,尽管参数数量较少,但 CNN 模型的计算复杂度仍然较高。

### 3.5 生成对抗网络(GAN)

GAN 由生成器网络(G)和判别器网络(D)组成,如图 6 所示,GAN 的训练通常基于卷积层或全连接层。基于 GAN 训练的语音增强(speech enhancement GAN,SEGAN)最早由 Pascual 等[10]提出。生成器网络学习将噪声语音的特征映射到干净语音中。随后,作为二元分类器的判别器网络会评估样本是来自干净语音(真实的)还是来自增强语音(虚假的)。根据鉴别器的结果,生成器会尝试调整分布以产生更好的输出,直到鉴别器无法区分输出是真还是假。该方法采用对抗性训练机制,通过迭代过程不断提升生成器产生高质量语音信号的能力,以及判别器识别生成语音与真实语音差异的精准度。

此外,GAN 框架可与其他语音处理技术相结合,实现语音的去噪、分离及增强等功能,显著提升增强语音的自然度和可懂度。然而,训练 GAN 既困难又不稳定,许多其他研究都试图提高 SEGAN 的性能。

Baby 等[63]在带梯度惩罚的鉴别器网络中实施了相对论损失函数,结果表明,改进后的鉴别器能产生更纯净的语音。此外,该研究还利用梯度惩罚作为训练的稳定器。Phan 等[32]建议使用多个发生

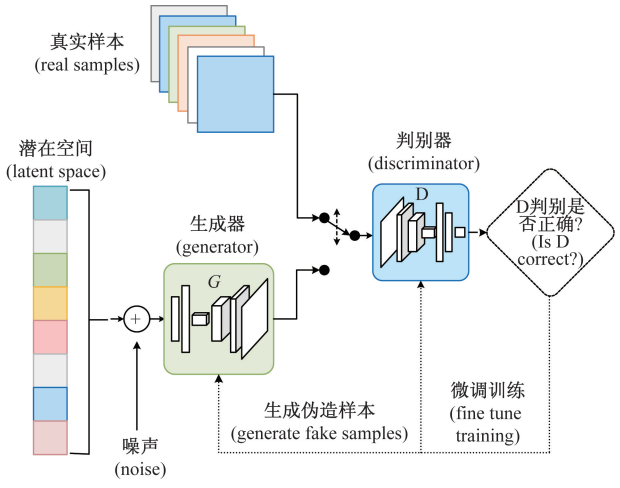



图 6 GAN 模型

Fig. 6 GAN model

器代替单个发生器,逐步精细化地处理带噪语音信号,引入迭代 SEGAN(iterated SEGAN,ISEGAN)和深度 SEGAN(deep SEGAN,DSEGAN)两种新架构。在 PESQ、综合信号失真度(composite signal distortion,CSIG)、综合背景噪声度(composite background noise,CBAK)、综合整体质量(composite overall quality,COVL)和分段信噪比(segmental signal-to-noise ratio,SSNR)方面都优于 SEGAN。

Pandey 等[64]引入条件 GAN 进行时频掩模估计,将网络训练方式从单一的 L1 损失函数转变为结合对抗性训练的框架,同时对 L1 与 L2 损失函数在语音增强中的适用性进行系统评估,结果表明,L1 损失在提高语音感知质量方面更为有效。Soni 等[65]用 GAN 来预测掩码,应用最小均方误差(minimum mean square error,MMSE)的正则化目标函数来改进 GAN。结果表明,与当时基于 GAN 的语音增强方法相比,PESQ 和 STOI 得到了改善。

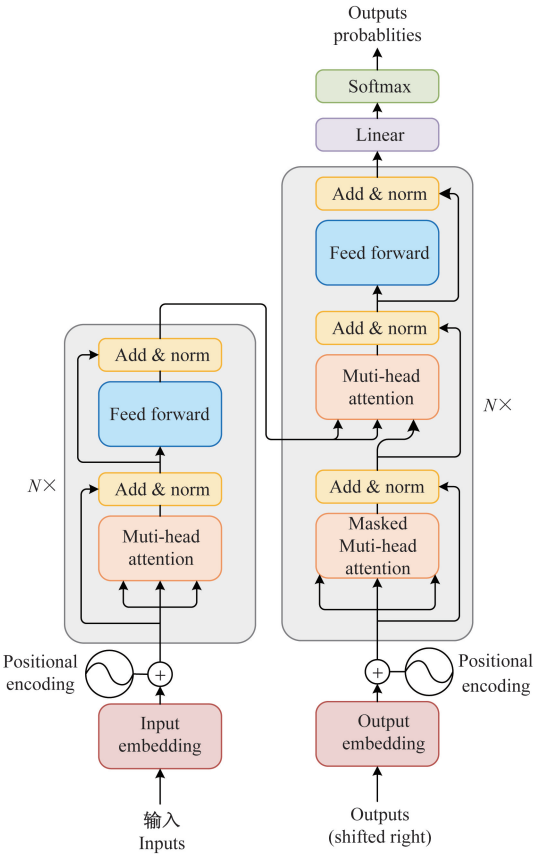
高级 GAN(high-level GAN,HLGAN)[66]消除了 SEGAN 在低 SNR 环境中去除噪声时丢失语音成分的问题。G 网络同时将噪声和干净语音作为输入,每个子网络中间层的输出用于计算损失函数。L1 准则损失特征被纳入为 SEGAN 模型提出的损失函数中。与 SEGAN 会损失一些语音成分相比,HLGAN 有显著改进。

Metric-GAN[67]通过优化感知指标取得了显著效果,而 CMGAN(conformer-based metric-GAN)[68]进一步将 Conformer 架构引入其中,利用其捕捉长时依赖和局部特征的能力,不仅在语音降噪任务上表现优异,还扩展到去混响和超分辨率任务,展示了其在多任务语音增强中的潜力。结果表明,CMGAN 在 PESQ 和 SSNR 等指标上均显著超越许多现有方法。

初期基于 GAN 的语音增强技术在模型泛化和稳定性方面存在局限。随后研究者提出各种优化方法来稳定训练过程并提高语音增强的性能,如渐进式生成器、多尺度鉴别器等,或引入滤波器和梯度惩罚机制。MSAE(multiscale autoencoder)框架[69]就是通过多尺度自动编码器对输入波形进行分解,形成多尺度嵌入,结合基于 U-Net 的掩蔽估计网络,实现端到端语音增强。结果表明,该框架在语音质量(如 PESQ、STOI)和自动语音识别性能上显著优于现有方法,尤其在噪声和混响环境下效果突出。这些研究不仅逐步提高了语音增强的性能,也为后续研究提供了新的方向,尤其是在低信噪比和资源受限环境下的应用。尽管如此,这些方法在实际应用中仍面临挑战,包括训练效率、模型复杂度和泛化能力,这些问题需要在未来的研究中得到进一步的解决和优化。

### 3.6　Transformer

Transformer 模型用于处理序列到序列的任务,它的核心思想是利用自注意力(Self-Attention)机制来捕捉序列内部各元素之间的全局依赖关系。图 7 显示了基础的 Transformer 模型,由编码器(Encoder)和解码器(Decoder)两部分组成,都由多个相同的层堆叠而成,每层都集成了自注意力和前馈神经网络。自注意力机制使模型能够同时考虑序列中所有位置的信息,而位置编码确保模型理解元素的顺序。Transformer 的并行化能力显著提升了训练速度,与依赖于顺序处理的 RNN 形成对比,使其在自然语言处理(NLP)领域得到了广泛应用,并催生了 BERT、GPT 和 T5 等,应用于语音识别、图像处理等其他领域。




N 为堆叠的层数;Output(shifted right)为输出右移;Input embedding 和 Output embedding 分别为输入嵌入和输出嵌入;Positional encoding 为位置编码;Muti-head attention 为多头注意力层;Masked Muti-head attention 为掩蔽的多头注意力层;Add & norm 为层归一化;Feed forward 为前置反馈层;Linear 为线性层;Softmax 函数;Outputs probablities 为输出概率

图 7　Transformer 编码器和解码器

Fig. 7　Transformer encoder and decoder

基于 Transformer 的语音增强方法利用其架构的自注意力机制,有效捕获时间序列中的长距离依赖关系,从而在单通道语音处理任务中实现卓越的性能。编码器用于提取语音特征,而解码器则负责重建干净的语音信号,通过引入位置编码和多头注意力,能够在不牺牲计算效率的情况下处理变长序列并捕捉复杂的时间模式。

Kim 等[70]提出 T-GSA(transformer with gaussian-weighted self-attention)用于语音增强,引入高斯加权自注意力机制,对注意力权重进行调整,以反映目标和上下文符号之间的距离。实验表明,T-GSA在语音增强性能上显著优于现有的 RNN 和原始 Transformer 模型。

Wang 等[12]提出新型两阶段 Transformer(two-stage transformer based neural network,TSTNN),并行处理局部信息和全局依赖信息,通过掩蔽模块与解码器重构出增强语音。在多个评估指标上超越了现有的时域和频域方法,同时显著降低了模型复杂度。Yu 等[71]提出了的双分支 Transformer(dual-branch attention-in-attention transformer,DB-AIAT)模型,引入注意力-注意力变换器(AIAT)模块,并行处理幅度掩蔽和复杂细化两个分支,有效恢复语音的幅度和相位信息。在 Voice Bank + DEMAND 数据集上达到了最先进的性能,同时保持了相对较小的模型规模。Yu 等[72]融合 LocalLSTM 和多头注意力机制提出 SETransformer,有效学习长期依赖关系而无需位置编码。结果表明,PESQ 和 STOI 相较于标准 Transformer 和 LSTM 模型展现出更优的去噪性能,并具有更高的计算效率。Saleem 等[73]提出基于卷积注意力 Transformer 网络(NSE-CATNet)的深度神经语音增强系统,引入基于 Transformer 的瓶颈层,利用一维卷积层和多头注意力(multi-head attention,MHA),通过 T-F 注意力模块的二维注意力能够更加精确地量化时频语音分布,从而在语音增强任务中实现更优的性能。DBT-Net 模型是一种双分支解耦式频谱估计框架,结合基于 Attention-in-Attention的 Transformer 网络,能够同时处理幅度估计和相位修复任务[74]。实验结果表明,该方法在 WSJ0-SI84 和 VoiceBank-DEMAND 数据集上的性能优于现有方法,达到最先进水平。

DPHT-ANet[75]是一种新型双路径高阶 Transformer 风格全注意力网络,通过高效的递归门控卷积模块替代传统多头注意力模块,有效捕捉时间和频率维度的深层特征信息,同时显著降低了模型的参数量和计算复杂度。实验结果表明,该模型在语音质量[PESQ、信号失真比(sigral to distortion ratio,SDR)]和可懂度[(STOI)、ESTOI(extended STOI)]上均优于现有方法,展现了卓越的性能。

从尝试将 Transformer 架构应用于语音信号处理任务开始,研究者们通过引入自注意力机制的不同变体,来增强模型对语音信号时频特性的捕捉能力。采用双分支结构允许模型同时关注语音的粗略和精细特征,以及通过时频注意力模块来显式地利用位置信息,从而在时间-频率分布上生成更为精确的注意力图。这些方法的共同优点在于它们能够显著提高语音增强的性能,尤其是在非平稳噪声环境下的鲁棒性。然而,这些复杂的网络结构也带来了更高的计算复杂度和参数数量,这可能对实时应用和资源受限的设备构成挑战。

## 4 不同模型架构对比

在深度学习驱动的语音增强领域,不同的模型架构根据其设计原理和应用目标展现出各自的特点和局限性。表 2 为对几种主流模型架构的主要优缺点对比分析。

GAN 通过对抗过程生成高质量的语音信号,能够学习到复杂的数据分布,适用于高保真语音增强任务,如语音重建。但其训练过程复杂,存在模式崩溃问题,可能导致生成器生成的样本多样性不足。

RNN 及其变体 LSTM 和 GRU,擅长捕捉语音信号的长期依赖关系,在处理具有明显时间序列特性的语音信号时有更显著的优势。如语音增强后需要进行的语音识别和语音合成。然而在训练时可能出现梯度消失或爆炸问题,且计算效率相对较低。

CNN 因其能够有效捕捉局部特征,广泛应用于语音信号的初步处理和噪声估计,在语音增强中能够对带噪语音信号进行更好的噪声估计和特征增强。但是 CNN 对长距离依赖的捕捉能力有限,通常需要与其他模型(如 RNN 或 Transformer)结合使用,以提升语音增强性能。

Transformer 模型依赖自注意力机制,可并行处理序列数据并捕捉长距离依赖,计算效率高。其对语音增强任务(如实时语音通信)尤为适用,然而对序列的顺序信息处理能力有限,可能需要额外的位置编码来增强模型对序列顺序的感知。

混合模型也在语音增强中展现出独特优势。解元等[76]提出了一种基于混合混响模型的算法,结合了多通道线性预测模型和空间相干模型的优点,利用卡尔曼滤波器和矩阵特征值分解优化模型性能,在低混响和高混响环境下均取得了较好的增强效果。这类混合模型虽然在处理复杂场景时表现出色,但其结构复杂,训练和调参难度较大。而注意

表 2　基于深度学习的语音增强方法的优缺点

Table 2　Advantages and disadvantages of deep learning based speech enhancement methods

| 方法类型 | 优点 | 缺点 |
|---|---|---|
| DNN | 能够通过学习大量的训练数据来捕捉和模拟复杂的非线性关系;根据所需的结果进行优化调整,应用于多种数据类型 | 需要大量数据训练数据和高质量标签;模型复杂度高,计算和存储成本较大;可能存在延迟、梯度消失或爆炸问题,影响训练稳定性 |
| RNN | 能够捕捉语音信号的时间序列特性;与卷积层结合时可以扩展感受野 | 计算成本较高,训练困难;容易梯度消失或爆炸;对长时序依赖关系的建模效率较低 |
| LSTM | 解决了传统 RNN 可能梯度消失问题;可长期存储时间序列信息;适用于语音信号中较长时序依赖的任务 | 模型训练时间较长,内存占用高;延迟较大,尤其在实时应用中表现不佳;对参数初始化和超参数设置较为敏感 |
| GRU | 改进了 LSTM 的门控机制,减少了计算复杂度;更高效,适用于资源受限环境;能够较好解决梯度消失问题 | 在捕捉复杂长时间依赖关系上可能不如 LSTM;收敛速度慢,学习效率低 |
| CNN | 在局部特征提取和模式识别中表现优异,适合处理语音的时间频率特征;自动学习特征,无须人工监督 | 不适合处理长时间序列依赖性特征;对语音输入数据的变化缺乏强鲁棒性;需要与其他模型结合才能解决时序建模问题 |
| GAN | 能够生成高质量的增强语音信号,重建信号更接近原始语音;支持多种编码器-解码器结构,灵活性高 | 对抗训练比较困难和不稳定;对超参数的选择较为敏感,调优难度高;训练成本较高,尤其在高维复杂数据中表现出瓶颈 |
| AE | 能够学习到信号的压缩表示;可以实现降维,有助于特征提取 | 传统 AE 可能在处理非线性和复杂数据分布方面存在局限性;复杂语音增强性能不如更高级的深度学习模型 |
| DAE | 能够学习从噪声语音中恢复干净语音的映射;在训练阶段引入噪声,提升模型鲁棒性 | 对未知类型噪声或非平稳噪声的处理能力有限 |
| VAE | 通过引入概率生成模型,实现自然和多样化的语音信号生成;适用于复杂信号建模,能够生成连续性更好的特征 | 涉及复杂的变分推断步骤,训练难度较大;对超参数敏感性高,优化和调试成本较高 |
| Transformer | 自注意力机制能够有效捕捉长距离依赖关系;较好处理非局部特征;并行化能力强,计算效率高 | 结构复杂,参数数量多,计算资源需求高;需要大量高质量训练数据;实时应用中可能存在延迟问题 |

力机制增强的 Transformer 在非平稳噪声环境下表现尤为突出。特别要提到的是时间-频率注意力网络,该网络能够显式地利用位置信息,为时间-频率分布的语音信号生成二维注意力图,提高语音增强的精度。在需要对语音信号进行精细分析和处理的场景中,如语音质量提升和语音识别预处理,该网络发挥着重要作用。

在语音增强的研究和应用中,每种模型架构都有其独特的优势和局限,选择合适的模型架构需要考虑任务的具体需求、计算资源的限制以及期望的性能指标。尽管深度学习方法在复杂噪声环境下表现优越,但在特定场景(如激光相干探测)的语音增强中,基于高斯混合模型(GMM)和矢量泰勒级数(vector taylor series,VTS)的方法依然展现出良好的性能,尤其是在噪声和信道特性的自适应估计方面[77]。

## 5　评估指标

语音质量评估最初始于主观语音质量评估指标,主观评价指标主要有:平均意见评分(mean opinion score,MOS)、失真平均意见评分(degradation mean opinion score,DMOS)和判断韵字测试法(diagnostic rhyme test,DRT)等。主观评估是评估语音质量和清晰度的最准确的方法,然而主观评估方法成本较高且耗时,因此,很多客观指标已经开发出来用于准确地预测语音质量和可懂度。已经提出的用于估计语音质量的客观度量包括 PESQ、STOI、SDR、巴克失真度(Bark distortion measure, BSD)、对数似然比距离(log-likelihood ratio distance, LLR)、分段信噪比(segmental signal to noise ratio, SegSNR)等。表 3[17,78-82]列举了部分常见用于评估语音增强算法中语音质量和可懂度的指标及其计算公式。其中常用的是 PESQ 和 STOI 两个指标。

### 5.1　PESQ

PESQ 对原始信号和通过被测系统的信号首先电平调整到标准听觉电平,再利用 IRS(intermediate reference system)滤波器模拟标准电话听筒进行滤波,将两个信号在时间上对准,并进行听觉变换,这个变换包括对系统中线性滤波和增益变化的补偿和均衡;将两个听觉变换后的信号之间的谱失真测度作为扰动(即差值),分析扰动曲面提取出的两个退化参数,在频率和时间上累积起来,映射到 MOS 的预测值[78]。PESQ 得分范围在 -0.5 ~4.5,得分越高表示语音质量越好。

### 5.2　STOI

STOI 是用于衡量语音可懂度的重要指标之一,它通过对语音信号的短时片段进行分析,评估在噪声存在的情况下语音的可理解性。将语音信号分割成短时间帧,对每帧进行预处理和频谱分析以提取包络信息,然后计算瞬时频谱并估计每个时间窗口内的信噪比[79]。基于这些估计计算出一个得分,从而预测该时

间窗口内语音的可懂程度。STOI 的得分范围为 0 ~ 1,得分越高,表示语音的可懂度越高。

**表 3 语音增强评估指标[17,78-82]**

**Table 3 Speech enhancement evaluations measures[17,78-82]**

| 评估方法 | 数学表达式 |
|---|---|
| PESQ[78] | $\mathrm{PESQ}=4.5-0.1d_{\mathrm{SYM}}-0.0309d_{\mathrm{ASYM}}$ |
| STOI[79] | $\mathrm{STOI}=\frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}d_j(m)$ |
| SDR[80] | $\mathrm{SDR}=10\lg\frac{\|\boldsymbol{s}\|^2}{\|\boldsymbol{s}-\hat{\boldsymbol{s}}\|^2}$ |
| BSD[81] | $\mathrm{BSD}(k)=\sum_{b=1}^{N_b}[S_k(b)-\overline{S}_k(b)]^2$ |
| LLR[82] | $d_{\mathrm{LLR}}(\boldsymbol{a}_x,\bar{a}_{\hat{x}})=\ln\left(\frac{\bar{\boldsymbol{a}}_{\hat{x}}^{\mathrm{T}}\boldsymbol{R}_x\bar{\boldsymbol{a}}_{\hat{x}}}{\boldsymbol{a}_x^{\mathrm{T}}\boldsymbol{R}_x\boldsymbol{a}_x}\right)$ |
| SegSNR[81] | $\mathrm{SNR}_{\mathrm{seg}}=\frac{10}{M_{\mathrm{s}}}\sum_{m=0}^{M_{\mathrm{s}}-1}\lg\left\{\frac{\sum_{n=Nm}^{Nm+N-1}x^2(n)}{\sum_{n=Nm}^{Nm+N-1}[x(n)-\hat{x}(n)]^2}\right\}$ |
| 综合评价指标[17] | $C_{\mathrm{sig}}=3.093-1.092S_{\mathrm{LLR}}+0.603S_{\mathrm{PESQ}}-0.009S_{\mathrm{WSS}}$<br>$C_{\mathrm{bak}}=1.634+0.478S_{\mathrm{PESQ}}-0.007S_{\mathrm{WSS}}+0.063S_{\mathrm{SNR}_{\mathrm{seg}}}$ |

注:$d_{\mathrm{SYM}}$ 为平均干扰值;$d_{\mathrm{ASYM}}$ 为平均不对称干扰值;$J$ 为频带数目;$M$ 为帧的总数;$d_j(m)$ 为第 $j$ 个频带、第 $m$ 帧的归一化相关系数;$\boldsymbol{s}$ 为目标纯净语音;$\hat{\boldsymbol{s}}$ 为增强后的语音信号;$S_k(b)$、$\overline{S}_k(b)$ 分别为干净信和增强信号在第 $k$ 帧、第 $b$ 个巴克频带的响度谱;$b$ 为当前频带索引;$d_{\mathrm{LLR}}$ 为对数似然比测度;$\mathrm{SNR}_{\mathrm{seg}}$ 为分段信噪比;$N_b$ 为频带总数;$\boldsymbol{a}_x^{\mathrm{T}}$ 为干净信号的 LPC 系数;$\bar{\boldsymbol{a}}_{\hat{x}}^{\mathrm{T}}$ 为增强信号的 LPC 系数;$\boldsymbol{R}_x$ 为干净信号的自相关矩阵;$x(n)$ 为原始(干净)信号;$\hat{x}(n)$ 为增强信号;$N$ 为帧长度;$m$ 为帧的索引;$Nm$ 为第 $m$ 帧的起始采样点索引;$M_{\mathrm{s}}$ 为信号中的帧数;$C_{\mathrm{sig}}$ 为预测语音信号失真的复合度量;$C_{\mathrm{bak}}$ 为预测背景噪声影响的复合度量;$S_{\mathrm{LLR}}$ 为 LLR 得分;$S_{\mathrm{PESQ}}$ 为 PESQ 得分;$S_{\mathrm{WSS}}$ 为加权谱斜率得分;$S_{\mathrm{SNR}_{\mathrm{seg}}}$ 为 SegSNR 得分。

### 5.3 SDR

SDR 通过比较原始干净语音信号与处理后信号之间的能量比来衡量从噪声中恢复出的语音信号的质量。将原始信号与评估信号在时间上对齐,计算两者之间的误差,之后计算能量比[80],按照 SDR 的定义计算 SDR 进行结果分析,SDR 越高,表示语音增强算法的性能越好,恢复出的语音信号的失真越小。SDR 的计算公式如表 3 所示。

### 5.4 其他

WER 是一个评估语音识别系统性能的指标,它衡量的是语音识别结果与参考文本之间的差异,可以用来衡量语音增强后音频信号的识别准确性。尽管 WER 本身是语音识别领域的一个指标,它同样可以间接反映语音增强技术的效果。因为语音增强的主要目标之一是提高语音的可懂度和质量,从而使得语音识别系统能够更准确地识别和转录语音内容。

表 4 列出了不同基于深度学习的语音增强方法(不包含全部)所用的数据集、提取特征、训练类型、评估指标。

## 6 展望

尽管基于深度学习的语音增强技术目前已成为主流,在语音质量和可懂度方面明显优于传统模型,但当前技术仍存在以下不足,需进一步研究与优化。

(1)噪声泛化能力。由于噪声分布的多样性,现有模型在未知类型或非平稳噪声环境下的表现仍

**表 4 不同基于深度学习的语音增强方法的比较**

**Table 4 Comparison between different deep learning based speech enhancement models**

| 方法类型 | 数据集 | 提取特征 | 训练类型 | 评估指标 |
|---|---|---|---|---|
| DNN | NOISEX、CHiME-2、Voice Bank + DEMAND、TIMIT + AURORA | MFCC、LPS、(log/power) Mag | 监督学习 | PESQ、SDR、STOI、SegSNR |
| RNN | CHiME-3、Voice Bank + DEMAND、TIMIT | LPS、MFCC、Raw signal、(log/power) Mag | 监督学习 | PESQ、STOI |
| LSTM | CHiME-3、Voice Bank + DEMAND、WSJCAMO | LPS、MFCC、Raw signal、(log/power) Mag | 监督学习 | PESQ、STOI |
| GRU | CHiME-3、Voice Bank + DEMAND、TIMIT | LPS、MFCC、Raw signal | 监督学习 | PESQ、STOI |
| CNN | CHiME-4、Aurora-4 | Raw signal、(log/power) Mag | 监督学习 | PESQ、STOI、WER、SDR、CSIG、CBAK、COVL、SegSNR |
| GAN | TIMIT + NOISEX +SSN、Voice Bank +、DEMAND | MFCC、Raw signal | 无监督学习 | PESQ、STOI、CSIG, CBAK、COVL、SSNR、WER、SDR |
| AE | CHiME-2、WSJCAMO | MFCC、Log Me | 无监督学习 | PESQ、STOI |
| DAE | MFCC、(Log) Mel | CHiME-2 | 无监督学习 | PESQ、WER |
| VAE | Reverberant、ATR | MFCC、Log Mel | 无监督学习 | PESQ、STOI、SNR、MMSE |
| Transformer | QUT-NOISE-TIMIT、VoiceBank + DEMAND、TIMIT + Musan + Noise-92 | (log/power) Mag、Raw signal | 监督学习 | SDR、PESQ、CSIG CBAK、COVL、SegSNR |

有待提高,设计更加复杂的网络结构可能会增加训练难度。未来的研究可尝试引入层次化或多尺度分析方法,结合无监督或半监督学习技术,从未标记噪声数据中学习特征,从而提升模型对新噪声类型的适应性,使其在多场景、多干扰环境中仍能保持高质量的语音增强性能。

(2)实时性能。深度学习模型的计算复杂度可能导致实时处理的延迟,导致运行时能耗过大,在实时语音通信和远程会议等系统上表现有所欠缺。针对此需要探索更轻量级的模型或优化算法,如何将模型复杂度进一步压缩,并在低资源下仍能保证获得高质量的干净语音信号满足实时语音增强的需求。

(3)端到端学习。端到端学习模型因其直接映射输入与输出的能力在语音增强领域表现出巨大潜力。未来的研究可以通过引入注意力机制或多任务学习框架进一步优化端到端模型的性能,使其不仅在语音增强任务上表现优异,同时在其他相关任务(如语音识别、说话人识别)中也能实现协同提升。

(4)多任务学习。当前的语音增强模型在多任务学习框架上的研究不够充分。在语音增强的上下文中,多任务学习可以同时针对语音增强、语音识别和说话人识别等任务进行优化,利用任务间的相关性来提升模型对语音特征的理解和处理能力。这不仅能够提升语音增强效果,还为语音处理领域的集成化提供了一种高效的解决方案。

(5)鲁棒性。为适应不同录音设备产生的信号失真、环境噪声的变化以及信道效应的影响,更稳定、更通用的语音增强模型的研究有待开展。应重点关注训练数据的广泛性、模型的正则化,并对模型的优化方法进一步探究。还可以探索使用数据增强技术(如添加合成噪声、随机变速和变调处理等),以增强模型对各种噪声条件的鲁棒性。

深度神经网络强大的非线性拟合能力可以同时支撑多个任务在一个网络中同步完成。基于深度学习的语音增强方法不仅在语音增强领域表现出色,还在其他语音处理(语音识别、说话人识别等)领域有广泛的应用研究,并表现出优异的效果。通过持续的研究和创新,基于深度学习的语音增强技术有望在未来解决上述这些挑战,并在各种语音通信和处理系统中发挥更大的作用。

## 7　结论

(1)探讨深度神经网络模型在语音增强算法中的应用,借助神经网络的线性与非线性建模能力,完成对带噪语音信号从特征提取、去除噪声到还原干净语音信号的过程。包括 DNN、RNN、CNN、GAN 和 Transformer 在内的深度学习模型在自动特征提取、处理非平稳噪声、端到端学习、泛化能力以及多任务学习方面具有显著优势,已成为当前语音增强的主流方法。着重介绍了影响基于深度学习的语音增强方法的重要因素,梳理了主流深度学习模型在语音增强中的发展,对近年来的架构演变进行总结和对比。

(2)概括介绍了语音增强模型的组成和影响因素,对常用的语音增强数据集(包括纯净语音、噪声、带噪语音)进行整理。介绍不同深度神经网络在语音增强中的应用,对中外众多团队在相关领域的研究成果进行阐述和分析,并对各个模型优缺点、使用数据集、适用场景等进行了总结和对比,为该领域后续研究工作的进一步开展提供了参考。此外,还对常见的客观评价指标的计算方法进行总结,为探索更准确的评估指标、设计更优化的语音增强模型提供参考。

## 参　考　文　献


[1] Benesty J, Makino S, Chen J. Speech enhancement[M]. Berlin: Springer, 2006.

[2] 潘卫军, 王梓璇, 蒋培元, 等. 面向管制语音识别系统的性能评价方法[J]. 科学技术与工程, 2024, 24(33): 14278-14286.
Pan Weijun, Wang Zixuan, Jiang Peiyuan, et al. Performance evaluation methods for ATC speech recognition systems[J]. Science Technology and Engineering, 2024, 24(33): 14278-14286.

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

[4] Zalevsky Z, Mendlovic D. Fractional wiener filter[J]. Applied Optics, 1996, 35(20): 3930-3936.

[5] 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.

[6] Chennoukh S, Gerrits A, Miet G, et al. Speech enhancement via frequency bandwidth extension using line spectral frequencies[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. Salt Lake City: IEEE, 2001: 665-668.

[7] 万玫汐, 王华朋, 闫道申, 等. 基于改进 ecapa-tdnn 的法庭自动说话人识别[J]. 科学技术与工程, 2024, 24(27): 11763-11773.
Wan Meixi, Wang Huapeng, Yan Daoshen, et al. Forensic automatic speaker recognition based on enhanced ECAPA-TDNN[J]. Science Technology and Engineering, 2024, 24(27): 11763-11773.

[8] Takeuchi D, Yatabe K, Koizumi Y, et al. Real-Time speech enhancement using equilibriated RNN[C]//ICASSP 2020-2020 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 851-855.
[9] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[M]. Cham: Springer International Publishing, 2015.
[10] Pascual S, Bonafonte A, Serrà J. SEGAN: speech enhancement generative adversarial network [J]. arXiv Preprint, 2017: arXiv: 1703.09452.
[11] Meng Z, Jinyu L, Gong Y, et al. Cycle-Consistent speech enhancement[J]. arXiv Preprint, arXiv: 1809.02253.
[12] Wang K, He B, Zhu W P. TSTNN: two-stage transformer based neural network for speech enhancement in the time domain[C]// ICASSP 2021—2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto: IEEE, 2021: 7098-7102.
[13] Rakshit H, Ullah M A. A comparative study on window functions for designing efficient FIR filter[C]//9th International Forum on Strategic Technology (IFOST). Cox's Bazar: IEEE, 2014: 91-96.
[14] Ren M, Liao R, Urtasun R, et al. Normalizing the normalizers: comparing and extending network normalization schemes[J]. arXiv Preprint, 2017: arXiv: 1611.04520.
[15] Kang T G, Kwon K, Shin J W, et al. NMF-based speech enhancement incorporating deep neural network [C]//Interspeech. New York: ISCA, 2014: 2843-2846.
[16] Chauhan P M, Desai N P. Mel frequency cepstral coefficients (MFCC) based speaker identification in noisy environment using wiener filter[C]//International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE). Coimbatore: IEEE, 2014: 1-5.
[17] Philipos C L. Speech enhancement: theory and practice[M]. 1st ed. Boca Raton: CRC Press, 2007.
[18] Yang S, Wang D L. Robust speaker identification using auditory features and computational auditory scene analysis [C]//2008 IEEE International Conference on Acoustics, Speech and Signal Processing. Las Vegas, NV: IEEE, 2008: 1589-1592.
[19] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.
[20] Melve O K. Speech enhancement with deep neural networks[D]. Oslo: Norwegian University of Science and Technology, 2016.
[21] Purwins H, Li B, Virtanen T, et al. Deep learning for audio signal processing[J]. IEEE Journal of Selected Topics in Signal Processing, 2019, 13(2): 206-219.
[22] Neekhara P, Hussain S, Pandey P, et al. Universal adversarial perturbations for speech recognition systems[J]. arXiv Preprint, 2019: arXiv: 1905.03828.
[23] Pandey A, Wang D. Attentive training: a new training framework for speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1360-1370.
[24] Gulati A, Qin J, Chiu C C, et al. Conformer: convolution-augmented transformer for speech recognition [J]. arXiv Preprint, 2020: DOI: 10.48550/arXiv.2005.08100.
[25] Tagliasacchi M, Li Y, Misiunas K, et al. SEANet: a multi-modal speech enhancement network [J]. arXiv Preprint, 2020: DOI: 10.48550/arXiv.2009.02095.
[26] Chen W, Xing X, Xu X, et al. SpeechFormer: a hierarchical efficient framework incorporating the characteristics of speech [J]. arXiv Preprint, 2022: DOI: 10.48550/arXiv.2203.03812.
[27] 张天骐, 罗庆予, 张慧芝, 等. 复谱映射下融合高效 Transformer 的语音增强方法[J]. 信号处理, 2024(2): 406-416.
Zhang Tianqi, Luo Qingyu, Zhang Huizhi, et al. Speech enhancement method based on complex spectrum mapping with efficient Transformer[J]. Journal of Signal Processing, 2024(2): 406-416.
[28] Odelowo B O, Anderson D V. A study of training targets for deep neural network-based speech enhancement using noise prediction [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, 2018: 5409-5413.
[29] Nossier S A, Wall J, Moniri M, et al. Mapping and masking targets comparison using different deep learning based speech enhancement architectures [C]//International Joint Conference on Neural Networks (IJCNN). Glasgow: IEEE, 2020: 1-8.
[30] Germain F G, Chen Q, Koltun V. Speech denoising with deep feature losses[J]. arXiv Preprint, 2018: arXiv: 1806.10522.
[31] Rethage D, Pons J, Serra X. A wavenet for speech denoising[J]. arXiv Preprint, 2018: arXiv: 1706.07162.
[32] Phan H, McLoughlin I V, Pham L, et al. Improving GANs for speech enhancement[J]. IEEE Signal Processing Letters, 2020, 27: 1700-1704.
[33] Zhang X L, Wang D. A deep ensemble learning method for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(5): 967-977.
[34] Miller G A. The masking of speech[J]. Psychological Bulletin, 1947, 44(2): 105-129.
[35] Wang Y X, Narayanan A, Wang D L. On training targets for supervised speech separation [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.
[36] Wang D. On ideal binary mask as the computational goal of auditory scene analysis [M]. Boston: Kluwer Academic Publishers, 2005: 181-197.
[37] Erdogan H, Hershey J R, Watanabe S, et al. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015: 708-712.
[38] 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): 483.
[39] Wang D, Chen J. Supervised speech separation based on deep learning: an overview[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2018, 26(10): 1702-1726.
[40] Karjol P, Ajay Kumar M, Ghosh P K. Speech enhancement using multiple deep neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, 2018: 5049-5052.
[41] Zhao Y, Xu B, Giri R, et al. Perceptually guided speech enhancement using deep neural networks[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (IC-

ASSP). Calgary, AB: IEEE, 2018: 5074-5078.
[42] Kawanaka M, Koizumi Y, Miyazaki R, et al. Stable training of DNN for speech enhancement based on perceptually-motivated black-box cost function [J]. arXiv Preprint, 2020: arXiv: 2002.05879.
[43] Fujimura T, Koizumi Y, Yatabe K, et al. Noisy-target training: a training strategy for DNN-based speech enhancement without clean speech[J]. arXiv Preprint, 2021: arXiv: 2101.08625.
[44] Furnon N, Serizel R, Illina I, et al. DNN-based mask estimation for distributed speech enhancement in spatially unconstrained microphone arrays[J]. arXiv Preprint, 2020: arXiv: 2011.01714.
[45] Xu Y, Du J, Huang Z, et al. Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement[J]. arXiv Preprint, 2017: DOI: 10.48550/arXiv.1703.07172.
[46] Arisoy E, Sethy A, Ramabhadran B, et al. Bidirectional recurrent neural network language models for automatic speech recognition [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). South Brisbane: IEEE, 2015: 5421-5425.
[47] Tan K, Wang D. A convolutional recurrent neural network for real-time speech enhancement [C]//Interspeech. Hyderabad: ISCA, 2018: 3229-3233.
[48] Wahab F E, Ye Z, Saleem N, et al. Compact deep neural networks for real-time speech enhancement on resource-limited devices[J]. Speech Communication, 2024, 156: 103008.
[49] Geiger J T, Zhang Z, Weninger F, et al. Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling [C]//Interspeech. New York: ISCA, 2014: 631-635.
[50] Weninger F, Erdogan H, Watanabe S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[M]. Cham: Springer International Publishing, 2015: 91-99.
[51] Sun L, Du J, Dai L R, et al. Multiple-target deep learning for LSTM-RNN based speech enhancement[C]//Hands-free Speech Communications and Microphone Arrays (HSCMA). San Francisco, CA: IEEE, 2017: 136-140.
[52] Wang Z, Zhang T, Shao Y, et al. LSTM-convolutional-BLSTM encoder-decoder network for minimum mean-square error approach to speech enhancement [J]. Applied Acoustics, 2021, 172: 107647.
[53] Oruh J, Viriri S, Adegun A. Long short-term memory recurrent neural network for automatic speech recognition[J]. IEEE Access, 2022, 10: 30069-30079.
[54] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and otherneural network architectures [J]. Neural Networks, 2005, 18(5/6): 602-610.
[55] Zhang Y, Chen G, Yu D, et al. Highway long short-term memory RNNs for distant speech recognition[J]. arXiv Preprint, 2016: arXiv: 1510.08983.
[56] Xue S, Yan Z. Improving latency-controlled BLSTM acoustic models for online speech recognition[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans, LA: IEEE, 2017: 5340-5344.
[57] Pandey A, Wang D. A New framework for CNN-based speech enhancementin the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179.
[58] Fu S W, Tsao Y, Lu X. SNR-aware convolutional neural network modeling for speech enhancement[C]//Interspeech 2016. New York: ISCA, 2016: 3768-3772.
[59] Ouyang Z, Yu H, Zhu W P, et al. A fully convolutional neural network for complex spectrogram processing in speech enhancement [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 5756-5760.
[60] Pandey A, Wang D. Temporal convolutional neural network for real-time speech enhancement in the time domain[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton: IEEE, 2019: 6875-6879.
[61] 许春冬, 王磊, 胡菁兰, 等. 结合残差与双注意力机制的U-net语音增强方法[J]. 计算机工程与设计, 2024(11): 3383-3389.
Xu Chundong, Wang Lei, Hu Jinglan, et al. U-Net speech enhancement method combining residual and dual attention mechanism [J]. Computer Engineering and Design, 2024 (11): 3383-3389.
[62] 徐浩森, 姜囡, 齐志坤. 基于注意力机制的卷积循环网络语音降噪[J]. 科学技术与工程, 2022, 22(5): 1950-1957.
Xu Haosen, Jiang Nan, Qi Zhikun. Speech denoising based on attention mechanism using convolution loop network[J]. Science Technology and Engineering, 2022, 22(5): 1950-1957.
[63] Baby D, Verhulst S. Sergan: speech enhancement using relativistic generative adversarial networks with gradient penalty [C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019: 106-110.
[64] Pandey A, Wang D. On adversarial training and loss functions for speech enhancement [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, 2018: 5414-5418.
[65] Soni M H, Shah N, Patil H A. Time-frequency masking-based speech enhancement using generative adversarial network[C]// IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB: IEEE, 2018: 5039-5043.
[66] Yang F, Wang Z, Li J, et al. Improving generative adversarial networks for speech enhancement through regularization of latent representations[J]. Speech Communication, 2020, 118: 1-9.
[67] Fu S W, Liao C F, Tsao Y, et al. MetricGAN: generative adversarial networks based black-box metric scores optimization for speech enhancement[J]. arXiv Preprint, 2019: DOI: 10.48550/arXiv.1905.04874.
[68] Abdulatif S, Cao R, Yang B. CMGAN: Conformer-based metric-GAN for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2477-2493.
[69] Borgström B J, Brandstein M S. A multiscale autoencoder (MSAE) framework for end-to-end neural network speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2418-2431.
[70] Kim J, El-Khamy M, Lee J. T-GSA: Transformer with gaussian-weighted self-attention for speech enhancement[C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020: 6649-6653.
[71] Yu G, Li A, Zheng C, et al. Dual-branch attention-in-attention

Transformer for single-channel speech enhancement[J]. arXiv Preprint, 2022: arXiv: 2110.06467.

[72] Yu W, Zhou J, Wang H, et al. SETransformer: speech enhancement transformer[J]. Cognitive Computation, 2022, 14(3): 1152-1158.

[73] Saleem N, Gunawan T S, Kartiwi M, et al. NSE-CATNet: deep neural speech enhancement using convolutional attention Transformer network[J]. IEEE Access, 2023, 11: 66979-66994.

[74] Yu G, Li A, Wang H, et al. DBT-Net: dual-branch federative magnitude and phase estimation with attention-in-attention Transformer for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2629-2644.

[75] Saleem N, Bourouis S, Elmannai H, et al. DPHT-ANet: dual-path high-order transformer-style fully attentional network for monaural speech enhancement[J]. Applied Acoustics, 2024, 224: 110131.

[76] 解元, 邹涛, 孙为军, 等. 基于混合混响模型的多通道语音增强算法[J]. 通信学报, 2024(11): 15-26.
Xie Yuan, Zou Tao, Sun Weijun, et al. Multichannel speech enhancement algorithm based onhybrid reverberation model[J]. Journal on Communications, 2024(11): 15-26.

[77] 芮小博, 孔欣玥, 伍洲, 等. 基于谱特征自适应估计的激光相干语音探测信号增强方法[J]. 仪器仪表学报, 2024(8): 326-335.
Rui Xiaobo, Kong Xinyue, Wu Zhou, et al. Enhancement of speech detected by laser coherent detection method based on spectral feature adaptation[J]. Chinese Journal of Scientific Instrument, 2024(8): 326-335.

[78] Rix A W, Beerends J G, Hollier M P, et al. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs[C]//IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City, UT: IEEE, 2001: 749-752.

[79] Taal C, Hendriks R, Heusdens R, et al. A short-time objective intelligibility measure for time-frequency weighted noisy speech[C]//IEEE International Conference on Acoustics, Speech and Signal Processing. New York: IEEE, 2010: 4214-4217.

[80] Vincent E, Gribonval R, Fevotte C. Performance measurement in blind audio source separation[J]. IEEE Transactions on Audio, Speech and Language Processing, 2006, 14(4): 1462-1469.

[81] Loizou P C. Speech quality assessment[M]. Berlin: Springer, 2011.

[82] Quackenbush S R, Barnwell T P, Clements M A. Objective measures of speech quality[M]. Englewood Cliffs, N. J.: Prentice Hall, 1988.