基于神经网络与 MMSE 结合的语音增强算法

王 刚 [1]
WANG Gang

摘 要 语音信号在传输过程中，往往因噪声等因素的干扰而使其质量和可辨度受到影响。为解决这一问题，语音增强算法被提出。语音增强算法分为基于信号处理的传统方法和基于深度学习的方法。文章主要研究了噪声场景下的单通道语音增强算法，提出了一种神经网络与传统算法结合的语音增强算法，即在最小均方误差算法中利用神经网络来估计先验信噪比，以提升语音增强效果。在使用最小均方误差算法进行语音增强时，为估计先验信噪比，需要用语音活动检测算法或噪声估计算法计算后验信噪比。先验信噪比对语音增强的结果影响最大，其估计的准确性直接决定了算法的降噪效果，但传统方法往往难以准确计算出先验信噪比。因神经网络具有强大的学习能力，文章提出了基于神经网络与最小均方误差算法结合的语音增强算法，利用神经网络来估计先验信噪比，后结合最小均方误差算法实现语音增强。实验结果表明，该算法在各项指标上都优于传统算法，具有更好的语音增强性能表现。

关键词 语音增强；深度神经网络；最小均方误差

doi：10.3969/j.issn.1672-9528.2025.07.016

0 引言

随着科学技术的日益进步，人们的日常生活充满了各种智能语音产品，涵盖了智能手机、车载语音系统、智能家居设备以及人机交互系统等多个领域。语音指令因其自然、简洁且明确的特性，已成为人机交互中不可或缺的重要形式。然而，在语音信号的传输过程中，可能受到各种干扰，造成语音信号的质量下降。为提升语音质量和可懂度，出现了语音增强算法，也称噪声抑制算法 [1]。

语音增强算法主要目的是从语音信号中分离出有用的信息，使语音信号更加清晰和准确，从而提高语音识别、语音合成和其他语音处理应用的精度和效果 [2]。根据所使用的麦克风通道数的差异，语音增强算法可被划分为两类：一是基于单通道的语音增强算法；二是基于多通道的语音增强算法。这两类算法在处理语音增强问题时，各有其独特的应用场景和优势。一般而言，语音信号所蕴含的信息主要体现在 3 个方面：频率、时间以及空间。其中单通道语音增强算法利用的是时域信息和频域信息两种信息，而多通道语音增强算法同时利用了频域信息、时域信息和空间域信息 3 种信息，其硬件成本较单通道算法的成本更高。单通道语音增强算法仅利用了时域和频域两个方面的信息，缺少声音的空间信息，因此实现单通道的语音增强更具有挑战性。从实现技术路线上讲，语音增强算法分为基于信号处理的传统方法和基于深度学习（deep learning, DL）的方法两种类型。

传统的语音增强方法凭借其算法简洁性、计算高效性以及硬件需求低等特点，一直得到广泛研究与应用。传统语音增强方法包括有谱减法、统计模型法和子空间法等。虽然这些方法具有较好的语音增强效果，但其中一些方法依赖于语音活动检测（voice activity detection, VAD）算法和噪声估计算法，而且只适用于平稳噪声消除，对非平稳噪声鲁棒性不佳 [3]。采用统计模型的方法，能够有效地通过追踪含噪语音信号的平滑功率谱中的频谱最小值，实现对噪声的准确估计，从而在一定程度上处理非平稳噪声 [4]，但是其降噪效果依然有待提升。

深度学习作为当前研究的热点，也引起了语音增强领域的研究者们的关注。深度神经网络（deep neural network, DNN）在语音增强中发挥着重要作用 [5]，其监督学习的特性使得它在处理语音增强这类具有明确输入输出关系的任务时表现出色。通过假设噪声为加性噪声或设定房间混响系数，研究者能够轻松构建出适用于神经网络训练的数据集。与传统的语音增强方法相比，基于深度学习的算法在面对非平稳噪声时能取得更好的降噪效果。本文主要研究了噪声场景下的单通道语音增强算法，提出了一种神经网络与传统算法结合的语音增强算法，即在最小均方误差算法中利用神经网络来估计先验信噪比，以提升语音增强效果。

1. 中国西南电子技术研究所 四川成都 610036

1 最小均方误差算法

最小均方误差（minimum mean-square error, MMSE）算法是一种基于统计模型的语音增强方法，其核心是通过最小化估计语音与真实语音的均方误差来优化降噪效果。本节将分步骤解析 MMSE 算法的推导过程及关键参数意义。

1.1 均方误差定义

假设含噪语音信号在时域表示为 Y_k，在频域表示为 $Y(\omega_k)$，纯净语音信号幅度为 X_k，增强后的估计值为 $\hat{X}_k$。MMSE 算法的目标是最小化估计值与真实值的均方误差：

$$e = E\{(\hat{X}_k - X_k)^2\} \tag{1}$$

式中：$E\{\bullet\}$表示期望运算通过贝叶斯准则，可得 X_k 的后验概率密度函数：

$$p(X_k \mid Y(\omega_k)) = \frac{p(Y(\omega_k) \mid X_k)p(X_k)}{p(Y(\omega_k))} \tag{2}$$

式中：$p(Y(\omega_k) \mid X_k)$为含噪语音在给定纯净语音下的条件概率（似然函数）；$p(X_k)$ 为纯净语音幅度的先验概率；$p(Y(\omega_k))$ 为含噪语音的边际概率（归一化因子）。

1.2 统计独立性假设与条件概率推导

假设傅里叶变换系数之间是统计独立的，估计值$\hat{X}_k$可表示为后验期望，即：

$$\hat{X}_k = E(X_k \mid Y(\omega_k)) = \frac{\int_0^\infty x_k p(Y(\omega_k) \mid x_k)p(x_k)\mathrm{d}x_k}{\int_0^\infty p(Y(\omega_k) \mid x_k)p(x_k)\mathrm{d}x_k} \tag{3}$$

进一步考虑相位 θ_x 的影响，假设噪声信号服从高斯分布，条件概率密度为：

$$p(Y(\omega_k) \mid x_k, \theta_x) = \frac{1}{\pi\lambda_d(k)}\exp(-\frac{|Y(\omega_k) - X(\omega_k)|^2}{\lambda_d(k)}) \tag{4}$$

式中：$\lambda_d(k)$ 为噪声频谱中第 k 个频谱分量的方差（反映噪声功率）；$X(\omega_k)$ 为纯净语音的频域表示。

1.3 增益函数推导

纯净语音信号的联合概率密度函数为：

$$p(x_k, \theta_x) = \frac{x_k}{\pi\lambda_x(k)}\exp(-\frac{x_k^2}{\lambda_x(k)}) \tag{5}$$

式中：$\lambda_x(k)$ 为干净语音信号频谱中第 k 个频率分量的方差（反映语音功率）。

通过积分运算最终得到 MMSE 幅度谱估计器：

$$\hat{X}_k = \frac{\sqrt{\pi}}{2}\frac{\sqrt{v_k}}{\gamma_k}e^{(-\frac{v_k}{2})}[(1+v_k)I_0(\frac{v_k}{2}) + v_k I_1(\frac{v_k}{2})]Y_k \tag{6}$$

式中：I_0 和 I_1 分别表示第零阶和第一阶修正贝塞尔函数；$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$表示中间变量，与信噪比有关，其中 ξ_k 和 γ_k 分别表示先验和后验信噪比，数学定义为：

$$\begin{cases} \xi_k = \dfrac{\lambda_x(k)}{\lambda_d(k)} \\ \gamma_k = \dfrac{Y_k^2}{\lambda_d(k)} \end{cases} \tag{7}$$

式中：Y_k 表示含噪语音在频率 ω_k 处的幅度。先验信噪比 ξ_k 对语音增强的影响最大，是主要参数，只要 ξ_k 足够低，就可以得到较强的抑制效果；后验信噪比 γ_k 是校正参数，并且只在先验信噪比较低的时候影响衰减量。

式（6）可分解含噪语音幅度 Y_k 与频谱增益函数 G 的乘积形式：

$$\hat{X}_k = G \cdot Y_k \tag{8}$$

因此，计算 MMSE 估计器的前提条件是已知先验信噪比 ξ_k 和噪声方差 λ_d。在假定噪声为平稳信号的基础上，利用语音活动检测算法或噪声估计算法在非语音段来计算噪声方差。其中先验信噪比估计是至关重要的一环，其估计的准确性将很大程度上影响降噪系统的整体性能。

2 DNN 与 MMSE 结合的语音增强算法

2.1 基于 MMSE-DNN 的语音增强算法流程

基于 MMSE 的语音增强算法流程如图 1 所示，首先需要进行语音活动检测或噪声估计，然后估计先验信噪比和后验信噪比进而计算出增益函数，再将增益函数与含噪信号谱相乘，得到纯净语音谱，最后对估计的纯净语音谱进行傅立叶反变换，得到时域估计信号。

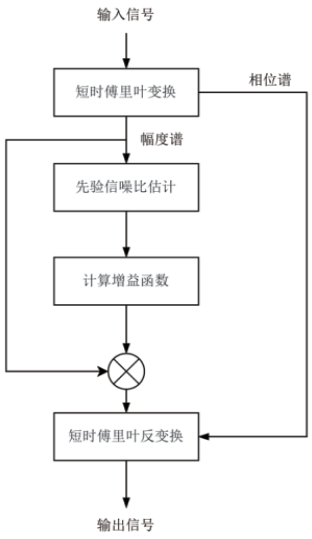


图 1 基于 MMSE 的语音增强算法流程

在 MMSE 算法中，估计先验信噪比 ξ_k 是十分重要的一环，其估计的准确性直接影响了算法的降噪效果。考虑到神经网络具有强大的学习能力，本文采用神经网络与 MMSE 算法相结合的方法来实现语音增强。

本文使用神经网络对先验信噪比 ξ_k 进行估计，然后利用

MMSE 的增益函数，对含噪语音进行降噪处理，其具体流程如图 2 所示。

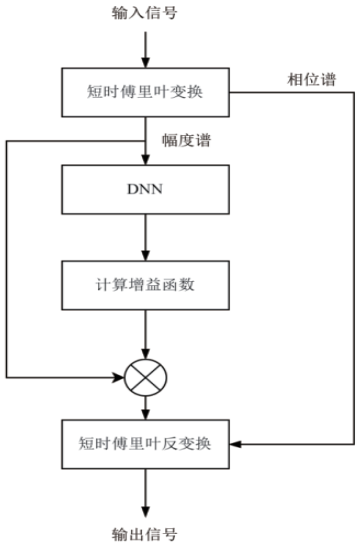


图 2 基于 MMSE-DNN 的语音增强算法流程

首先将含噪语音的时域信号进行短时傅里叶变换，得到其频域表示，再将含噪语音的幅度谱作为网络的输入，然后由神经网络估计出的先验信噪比和后验信噪比应用于增益函数的计算中：

$$G(t,f)=\mathrm{DNN}(|Y(t,f)|) \tag{9}$$

式中：t 表示帧序号；f 表示频率序号。降噪后的幅度谱由增益函数计算得到，经 ISTFT 将其转换为时域信号：

$$\hat{x}=\mathrm{ISTFT}(G|Y|e^{j\theta_Y}) \tag{10}$$

式中：θ_Y 指含噪信号的原始相位。

2.2 DNN 网络结构

本文所提出的基于 MMSE-DNN 的语音增强算法中，DNN 的网络结构如图 3 所示，采用卷积循环网络 CRN[6] 的结构，其中包含卷积编解码器模块（convolutional encoder-decoder, CED）和双路径循环（dual-path recurrent neural network, DPRNN）模块 [7]。

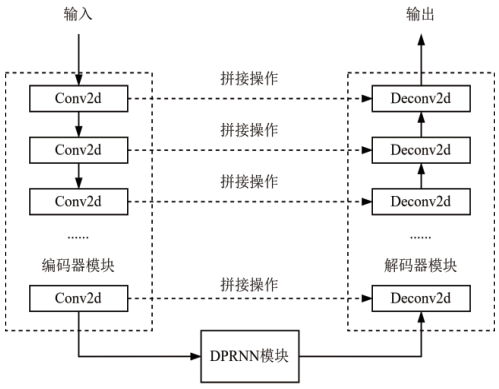


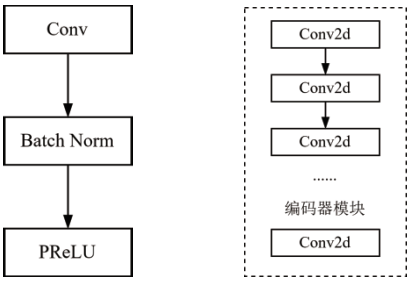
图 3 基于 MMSE-DNN 的语言增强算法 DNN 网络结构

DNN 网络结构实现了对先验信噪比的估计，同时兼顾了较低的网络参数和较好的降噪性能。下文将对所提出算法中的编码器模块，解码器模块和双路径循环网络模块做出介绍。

2.2.1 CED 模块

CED 的优点是可以有效地将输入信号压缩成较低维度的特征，以减少模型的参数同时还能保留重要的特征信息，并从这些特征中重新构建出目标信号。编解码器由两部分组成，分别是编码器模块（Encoder）和解码器模块（Decoder）。

编码器模块的结构如图 4 所示，由多个二维的卷积神经网络（Conv2d）块构成，Conv2d 包含了二维卷积层、批量归一化层（batch norm, BN）和带参数的线性整流函数（parametric rectified linear unit, PReLU）[8] 层。

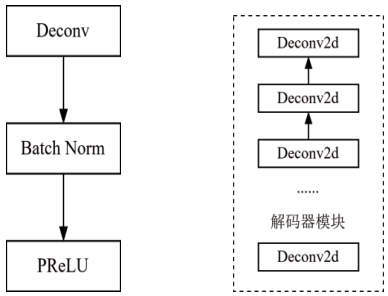


（a） 编码器中的 Conv2d （b） 编码器的网络结构

图 4 编码器网络结构

编码器模块中的二维卷积沿着时间和频率维度进行计算，提取语谱图中的局部特征。批量归一化层能够加速深度学习的训练过程，同时提高模型的稳定性。PReLU 是一种改进的激活函数，可以缓解梯度消失问题，增强模型的泛化能力。

解码器模块的结构如图 5 所示。解码器模块由多个二维转置卷积神经网络（Deconv2d）块构成，Deconv2d 包含了二维转置卷积层、批量归一化层和 PReLU 层。二维转置卷积层负责恢复特征尺寸，将 Encoder 压缩后的特征恢复到原始维度。值得注意的是，Decoder 在最后一层转置卷积采用了双曲正切函数作为激活函数，目的是将网络的输出控制在 -1 到 1 的范围。



（a） 解码器中的 Deconv2d；（b）解码器的网络结构

图 5 解码器网络结构

为更好地利用输入数据的特征，缓解梯度消失问题，本文在 Encoder 和 Decoder 之间还加入了跳跃连接机制。在解码器中，将当前层 Decoder 的输出与对应层 Encoder 的输出沿通道维度进行拼接，作为下一层 Decoder 的输入。

2.2.2 DPRNN 模块

Luo 等人 [7] 提出了双路径循环神经网络，该网络用于解决时域的语音分离问题，其网络结构如图 6 所示。

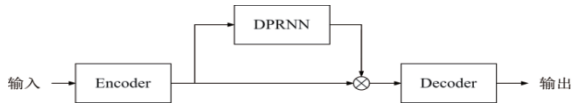


图 6 双路径循环神经网络

双路径循环神经网络包含了编码器（Encoder）、双路径 RNN 模块（dual-path RNN, DPRNN）和解码器（Decoder）3 个部分。在文献 [7] 中，Encoder 提取语音中的特征后，将语音按顺序分割为具有重叠部分的块，并将所有的块拼接起来，然后将这些特征块输入到 DPRNN 模块中。DPRNN 经过对块内和块间特征建模，将预测出的掩码与编码器编码后的特征相乘，作为 Decoder 的输入。Decoder 负责恢复特征，并使用重叠相加法（overlap-add method）将网络的输出转换回时域信号。

DPRNN 模块的结构如图 7 所示，包括两部分：块内处理和块间处理。这种处理方式背后的原理是：将一维声音信号通过重叠分块转换成二维矩阵形式，这些矩阵中的信息不只是块内有相关性，在不同的块之间也有相关性。传统的网络架构主要关注于块内信息的处理，而忽略了块与块之间的信息交流。DPRNN 通过其特定的设计有效地结合了这两方面的信息。

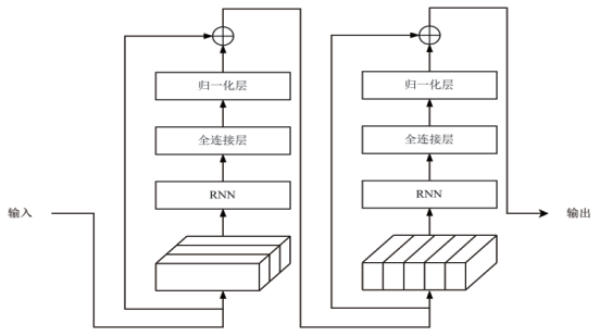


图 7 DPRNN 模块示意图

在 DPRNN 的处理流程中，首先对每个块内的信息进行处理，利用 RNN 来捕获块内的特征动态，随后，通过归一化层并结合残差连接的方式对提取的特征进行进一步处理，最后形成块内处理的输出。处理完块内信息后，接下来便是块间信息的处理，这需要将前一步骤的输出进行转置，以便在块间应用 RNN 进行信息的特征提取。块间 RNN 处理完成后，特征会被送入一个全连接层和归一化层，最终通过加入残差连接来生成最后的输出特征。这样的设计使得 DPRNN 能够更全面地利用音频信号中的时间和频率信息。

3 实验设置

3.1 实验目的

本文旨在验证所提出的 MMSE-DNN 的有效性和性能表现。通过与基于判决引导法的 MMSE（MMSE-DD）和基于最大似然法的 MMSE（MMSE-ML）的对比，评估 MMSE-DNN 算法在不同噪声环境下的语音增强效果。

3.2 数据集

为了对本文提出的算法进行评估，本文选用了 VCTK[10] 数据集。该数据集常被用于训练语音增强网络，包含了一系列含噪语音及其对应的纯净语音样本。这些纯净语音样本源自 VoiceBank[11] 语料库，涉及训练集中的 28 位讲话者共 11 572 个语音片段，以及测试集中的 2 位讲话者共 824 个语音片段。本文从 DEMAND[12] 噪声库中选取了若干噪声样本，并按照不同的信噪比（0、5、10、15 dB）将噪声加到纯净语音中，以此来生成含噪语音。在训练集的制备过程中，本文随机选择了 10 种噪声，包括 8 种来自现实场景录制的噪声和 2 种人工合成的噪声。而在测试集的构建中，本文引入了另外 5 种噪声，并同样按照 0、5、10、15 dB 的信噪比进行了含噪语音的生成。

3.3 实验参数设置

在本文提出的 MMSE-DNN 算法中，短时傅里叶变换所采用的窗函数是汉宁窗，窗长和帧移分别为 512 点和 256 点。数据集中所有的语音采样频率为 16 kHz，因此窗长和帧移就是 32 ms 和 16 ms，傅里叶变换点数采用的 512 点。在编解码器模块中，编码器模块的卷积层数量设置为 3 层，卷积层的通道数分别为 64、64、64，卷积层的卷积核大小分别为 (5, 2) (3, 2) (3, 2)，卷积层的步长分别为 (2, 1) (2, 1) (1, 1)。解码器模块的转置卷积层数量设置为 3 层，转置卷积层的通道数分别为 64、64、1，转置卷积层的卷积核大小分别为 (5, 2) (3, 2) (3, 2)，转置卷积层的步长分别为 (2, 1) (2, 1) (1, 1)。在 DPRNN 模块中，采用了 LSTM 对块内和块间特征进行处理，每一个 LSTM 有两层，每层有 64 个隐藏单元，其中块内 LSTM 采用了双向的结构，由于要保证系统的因果性，块间 LSTM 采用了单向结构。

本实验中的训练语音样本采样率被设置为 16 kHz，训练时所有语音被截取成长度为 4 s 的语音片段。使用 Adam 优化器对网络进行优化，初始学习率设置为 0.001，模型训练的周期上限被设置为 100 轮，当损失函数连续两轮迭代没有下降时，将学习率衰减为原来的 0.8 倍。如果模型的损失超

过 10 轮迭代都没有下降时，则提前终止训练。本文使用均方误差损失函数来训练网络，压缩瞬时信噪比采用的 k=1，c=0.1，并且本文以 PESQ、STOI 和综合客观语音质量作为算法降噪性能的评价标准。

3.4 实验结果分析

图 8 是经过各种算法降噪处理后的语谱图，可以清晰地看到，原始语音中存在明显的背景噪声。

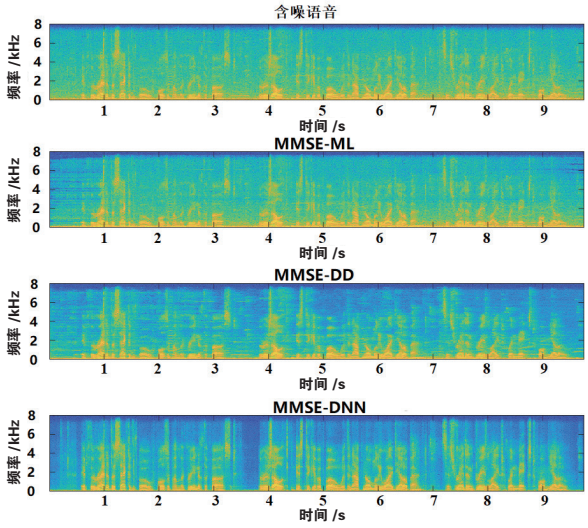


图 8 算法处理后的语谱图对比

从图 8 中可以看到，经过不同算法处理后，噪声得到了不同程度的抑制，然而，每种算法在降噪效果上存在差异。分析原始语音的频谱，可以发现原始语音的各个频段都受到了噪声干扰，特别是低频部分受到的影响最为明显。观察含噪语音的频谱，可以看到 0~1 s 和 3~4 s 为静音段，MMSE-DNN 算法展现出了比其他两种算法更为优越的语音增强能力。在有语音活动期间，MMSE-DD 算法在处理过程中会导致部分语音谐波丢失，而 MMSE-ML 算法在抑制噪声方面表现不佳。相比之下，MMSE-DNN 算法不仅在静音段表现出色，而且在语音活动期间同样能够有效地抑制噪声，同时还能保留住语音的谐波成分，说明本文提出的结合 DNN 和 MMSE 的语音增强算法在性能和稳定性方面均表现出优越性。

表 1 展示了在各种噪声环境下，不同算法的 PESQ 得分。STRAFFIC 是指在交通路口环境下采集的背景噪声。PSTATION 是指在地铁站采集到的环境噪声。NRIVER 是在室外自然环境下采集到的流水声。

在 STRAFFIC 噪声中，按照信噪比从低到高的顺序，本文算法的 PESQ 得分比 MMSE-ML 高 10.9%、31.6%、49.6%、37.3%、29.6%，比 MMSE-DD 高 0.0%、13.6%、33.5%、40.6%、8.9%。在 PSTATION 噪声中，本文算法的 PESQ 得分比 MMSE-ML 高 8.5%、24.1%、37.1%、41.2%、30.5%，比 MMSE-DD 高 1.7%、5.1%、12.2%、5.9%、10.9%。在 NRIVER 噪声中，本文算法的 PESQ 得分比 MMSE-ML 高 29.6%、15.0%、40.98%、35.41%、37.39%，比 MMSE-DD 高 23.8%、10.9%、16.2%、9.7%、13.1%。从以上数据可以看出，当信噪比为 5 dB 和 10 dB 时，本文算法相对传统算法具有更加明显的优势。

表 1 不同算法的 PESQ 得分

噪声类型	算法	输入信噪比 /dB				
		-5	0	5	10	15
STRAFFIC	含噪语音	1.06	1.11	1.45	1.96	2.49
	MMSE-ML	1.10	1.20	1.49	2.09	2.53
	MMSE-DD	1.22	1.39	1.67	2.04	3.01
	本文模型	**1.22**	**1.58**	**2.23**	**2.87**	**3.28**
PSTATION	含噪语音	1.04	1.10	1.28	1.68	2.16
	MMSE-ML	1.05	1.16	1.40	1.77	2.26
	MMSE-DD	1.12	1.37	1.71	2.36	2.66
	本文模型	**1.14**	**1.44**	**1.92**	**2.50**	**2.95**
NRIVER	含噪语音	1.07	1.05	1.14	2.17	2.30
	MMSE-ML	1.08	1.06	1.22	2.40	2.38
	MMSE-DD	1.13	1.10	1.48	2.96	2.89
	本文模型	**1.40**	**1.22**	**1.72**	**3.25**	**3.27**

在对语音进行降噪处理时，不能仅仅关注降噪效果，还需要关注语音的失真情况，表 2 展示了在各种噪声环境下，不同算法的 STOI 得分对比。

表 2 不同算法的 STOI（%）得分

噪声类型	算法	输入信噪比 /dB				
		-5	0	5	10	15
STRAFFIC	含噪语音	69.27	84.23	94.83	97.56	99.19
	MMSE-ML	67.00	80.83	93.72	97.44	99.16
	MMSE-DD	64.93	77.94	87.91	92.51	97.95
	本文模型	**71.38**	**87.49**	**95.92**	**98.24**	**99.20**
PSTATION	含噪语音	63.34	77.87	87.80	93.83	97.76
	MMSE-ML	62.58	75.95	87.46	93.72	97.69
	MMSE-DD	60.99	73.52	84.16	90.99	96.10
	本文模型	**68.72**	**84.79**	**91.78**	**96.15**	**98.48**
NRIVER	含噪语音	87.18	68.32	80.14	99.10	98.86
	MMSE-ML	87.25	68.18	79.76	99.07	98.72
	MMSE-DD	86.47	67.39	77.31	98.05	97.92
	本文模型	**91.02**	**71.89**	**87.30**	**99.25**	**98.99**

从表 2 中可以看到，在含噪语音的低信噪较低时，传统算法会导致 STOI 分数下降。这是因为在传统算法处理的过程中，由于噪声估计不准确导致了算法对语音成分会过度抑制，损失了语音的一部分细节信息，因此造成了语音的可懂度下降 [13]。与其他的算法相比，本文提出的算法在各种信噪比情况和各种类型的噪声下的 STOI 分数都获得了提高，证明本文的算法在抑制噪声的同时，能够很好地保留语音的细节信息，并提升语音的可懂度。

4 结论

本文提出了基于神经网络与 MMSE 算法相结合的语音增强算法。首先介绍了基于 MMSE 的语音增强算法，MMSE 需要使用噪声估计算法或语音活动检测算法对后验信噪比进行估计，然后利用最大似然法或判决引导法计算出先验信噪比。先验信噪比对语音增强的效果影响最大，是最主要的参数，因此本章设计了用于估计先验信噪比的神经网络，该网络采用 CRN 结构，网络中包含了 CED 模块和 DPRNN 模块。在训练阶段，网络的输入为幅度谱，输出为压缩后的先验信噪比估计值，使用均方误差损失函数计算真实值和预测值之间的误差。在推理阶段，网络输入为含噪语音的幅度谱，然后对网络的估计值进行解压缩，得到先验信噪比的估计值。计算出增益函数后，对含噪语音幅度谱进行增强，结合含噪语音的相位谱重构时域信号。经实验结果表明，该方案在各项指标上都超过了其他算法，具备更好的语音增强性能表现，同时能够带来更小的语音失真。

参考文献：

[1] LOIZOU P C. Speech enhancement: theory and practice[M]. USA: CRC press, 2013: 64-93.

[2] KINOSHITA K, OCHIAI T, DELCROIX M, et al. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Piscataway:IEEE,2020:7009-7013.

[3] 董胡 , 徐雨明 , 马振中 , 等 . 基于小波包与自适应维纳滤波的语音增强算法 [J]. 计算机技术与发展 ,2020,30(1):50-53.

[4] LEE G W, KIM H K. Multi-task learning U-Net for single-channel speech enhancement and mask-based voice activity detection[J]. Applied sciences, 2020, 10(9):3230.

[5] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE signal processing letters, 2014, 21(1):65-68.

[6] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[EB/OL]. (2024-09-06)[2025-01-23].https://www.isca-archive.org/interspeech_2018/tan18_interspeech.html.

[7] LUO Y, CHEN Z, YOSHIOKA T. Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Piscataway: IEEE, 2020:46-50.

[8] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//2015 International Conference On Computer Vision (ICCV), NewYork:ACM,2015: 1026-1034.

[9] LE X H, CHEN H S, CHEN K, et al. DPCRN: dual-path convolution recurrent network for single channel speech enhancement[EB/OL].(2021-07-21)[2024-11-25].https://doi.org/10.48550/arXiv.2107.05429.

[10] VALENTINI-BOTINHAO C, WANG X, TAKAKI S, et al, Investigating rnn-based speech enhancement methods for noise-robust text-to-speech[EB/OL].(2016-09-20)[2025-02-12].https://www.isca-archive.org/ssw_2016/valentinibotinhao16_ssw.html.

[11] VEAUX C, YAMAGISHI J, KING S. The voice bank corpus: design, collection and data analysis of a large regional accent speech database[C]//2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation(O-COCOSDA/CASLRE), Piscataway:IEEE, 2013:1-4.

[12] THIEMANN J, ITO N, VINCENT E. The diverse environments multi-channel acoustic noise database (DEMAND): a database of multichannel environmental noise recordings[C]// Proceedings of Meetings on Acoustics, NewYork:ASA,2013: 35-81.

[13] LIM J S, OPPENHEIM A V. Enhancement and bandwidth compression of noisy speech[J]. Proceeding of the IEEE, 1979, 67(12): 1586-1604.

【作者简介】

王刚（1986—），男，重庆人，硕士，工程师，研究方向：通信与信息系统。

（收稿日期：2025-03-06 修回日期：2025-07-03）