徐羽, 卞紫萱, 刘丹,等. 基于 CNN 的单通道的语音增强[J]. 智能计算机与应用,2026,16(3):201-205. DOI:10.20169/j.issn.2095-2163.24043002

# 基于 CNN 的单通道的语音增强

徐 羽, 卞紫萱, 刘 丹, 张俊杰
(南京工程学院 信息与通信工程学院, 南京 211167)

**摘 要**: 语音增强是语音信号处理领域目前最具挑战性的任务之一。传统方法大多基于无监督学习,且噪声去除效果不明显,非平稳噪声消除效果较弱,性能在混响模型中较差。近些年来,由于深度神经网络计算能力的提升以及大数据的发展,在各个领域中得到了广泛应用,因此本文提出了一种基于升降编解码全卷积神经网络(IDEDFCNN)的语音增强算法。该模型首先利用编码器对输入语音信号的帧进行上下文信息采集,然后利用解码器将编码器采集信息与语音帧相互联系,从而达到更好地实现语音增强的目的。

**关键词**: 语音增强; 深度学习; 单通道; IDEDFCNN
**中图分类号**: TN912.35 **文献标志码**: A **文章编号**: 2095-2163(2026)03-0201-05

## Single channel speech enhancement based on CNN

XU Yu, BIAN Zixuan, LIU Dan, ZHANG Junjie
(School of Information and Communications Engineering, Nanjing Institute of Technology, Nanjing 211167, China)

**Abstract**: Speech enhancement is currently one of the most challenging tasks in the field of speech signal processing. And traditional methods are mostly based on unsupervised learning, with their poor performance in the reverberation model in which the noise removal effect is not obvious and the non-stationary noise elimination effect is weak. In recent years, deep neural networks have been widely used in various fields due to the improvement of their computational power and the development of big data. Therefore, this paper proposes a speech enhancement algorithm based on Increase Decrease Encoder Decode Full Convolutional Neural Network (IDEDFCNN). The model firstly uses an encoder to collect contextual information from the frames of the input speech signals, and then uses a decoder to interconnect the information with the speech frames, in order to achieve speech enhancement in a better way.

**Key words**: speech enhancement; deep learning; single channel; IDEDFCNN

## 0 引 言

早期语音增强方法的提出是为了提高语音的可懂度,这是由多种历史因素导致的[1-5]。一方面,起初的声学知识体系并不完善,一些语音的基本分类、概念才开始逐步形成、并涌现;另一方面,最开始的语音增强是为了服务其他科技领域的技术。随着数字信号与通信技术原理的逐步完善与发展、通信设备的更新迭代,对于语音的质量与可懂度需求逐步提高,语音增强技术得到了高速发展。

Lim 等学者经过研究与实验提出以维纳滤波作为语音增强算法[6-10]。Boll 等学者提出了基本谱减法。1984 年,Ephraim 等学者提出了基于最小均方误差(MMSE)的短时幅度谱语音增强方法,并在此基础上对其进行改进。

近些年,随着技术的革新与迭代,以及神经网络相关算法的完善,其计算能力逐步提升。同时,大数据技术也得到了快速发展,各类数据可供选择,并且获取途径多样,传统的语音增强算法对非平稳噪声的去除能力也有限,且适用环境也有限。深度学习网络(DNN)在图像处理与语音处理领域均取得不错的成效,为其在语音增强研究中的应用提供了条件。基于 DNN 的语音增强算法,对数据的标识和量具有较大的需求,并且对时间依赖性问题不是很敏

**作者简介**: 卞紫萱(2003—),女,本科生,主要研究方向:机器学习; 刘 丹(2003—),女,本科生,主要研究方向:系统推荐; 张俊杰(2003—),男,本科生,主要研究方向:计算机算法。
**通信作者**: 徐 羽(2003—),女,本科生,主要研究方向: 机器学习,数据分析。Email:2054703191@qq.com。

**收稿日期**: 2024-04-30

感,但 CNN 方法利用卷积的方法,较好地处理时频间的依赖关系。

综上,为了提升算法的语音增强的性能,本项目基于 CNN 网络进行相关研究。

## 1 系统设计

### 1.1 语音增强模型

本项目采用的是基于全卷积神经网络(FCN)的语音增强算法[11-16]。但是全卷积神经网络仍然存在降噪效果不明显等问题[17-20],因此本项目通过加入基于升降编解码卷积神经网络(IDEDFCNN)的模型。本项目的算法基本框架如图 1 所示。

本项目将纯净语音与噪声进行混合,再通过选取合适的窗函数进行分帧处理,将语音帧信号进行短时傅里叶变换以及维度转换,将信号输入相关模型得到处理后的频域语音信号,在此基础上通过逆变换和波形叠加得到最终的增强语音。

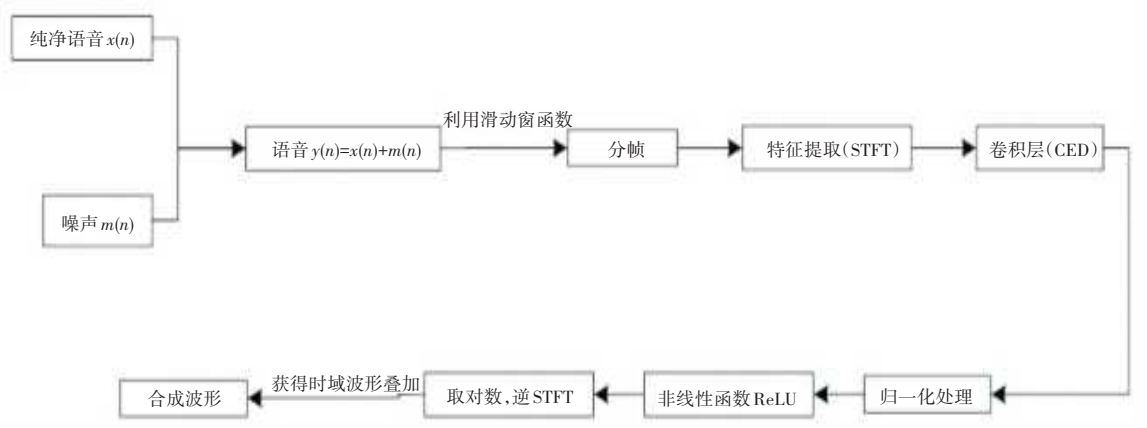


**图 1 算法的基本框架**

**Fig. 1 Basic framework of the algorithm**

### 1.2 分帧

本文对采集的语音数据进行了分帧加窗操作。在语音增强时,训练特征选取的是傅里叶对数幅度谱特征,因此研究中选取合适的频段,将语音信号安装顶下标准逐步提取一个短信号,作为一帧。由于吉布斯效应会对分帧后的信号产生影响,因此加入窗函数,来确保傅里叶变换后语音帧的连续性。但加窗会使帧信号两端出现信号的衰弱,因此信号帧一部分与前后帧是重叠的。汉明窗相关表达式如下:

$$\omega(n)=\begin{cases}0.54-0.46\cos[2\pi n/(N-1)], & 0\leqslant n\leqslant N-1\\ 0, & \text{其他}\end{cases} \tag{1}$$

### 1.3 语音信号特征提取

语音信号特征提取是将每个语音信号定义为一个多维特征向量。为了方便本项目对语音特征的理解、采集与操作,本项目选取傅里叶对数幅度谱特征作为训练模型的训练特征,实现语音增强效果。

传统的傅里叶变换作用于语音信号整体,对于局部的相关特征提取效果较弱,对短时信号的处理上存在不足,而短时傅里叶变换(the Short-time Fourier Transform,STFT)根据前文所提取的短时帧信号进行频域转换,将其与时域有所联系。语音信号经过 STFT 可以提取其幅度谱及对数谱得到相关关系。让网络根据所提取的特征生成的增强信号符合人耳听觉效应,本文利用幅度谱及其对数谱作为训练特征,并取其中系数部分。对此可以表示为:

$$I_A(t,f)=\log(X_A(t,f)) \tag{2}$$

本文通过 STFT 变换以及取其对数变换获取频域对数谱特征。

### 1.4 评价指标

对于语音增强效果,人耳的听取质量为其一,但对差别度不是很大的语音,就需要引入客观评价指标。以下介绍常见的客观评价指标:PESQ 和 STOI。

(1) PESQ 算法需要带噪的衰减信号和一个原始的参考信号。提取 2 个信号参数的时频特征,综合其时频特性,得到最后的评价参考,一般取值范围在-0.5~4.5 间,得分越高表示和原语音吻合度越高,即语音增强效果较好。

(2)STOI 表示的是在短时帧中原始干净信号,与增强后信号的相关性取值范围为[0,1],值越高越好。

本项目信噪比较低,所以采用 STOI 作为语音评价的主要指标。

## 2　语音增强算法

### 2.1　基于 CNN 的语音增强算法

卷积神经网络(CNN)在图像处理方面取得较为广泛的应用,其局部提取特征能力较为优异,因此被运用到语音信号的处理中。CNN 网络结构如图 2 所示。基于 CNN 的语音增强在语音信号被各种噪声干扰或淹没的情况下,该技术将能从噪声背景中提取出有用的语音信号,抑制和降低噪声的干扰。

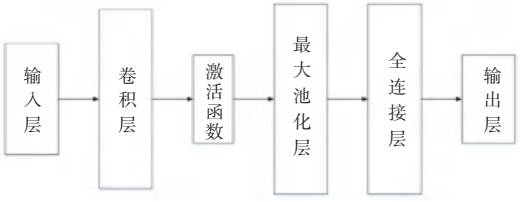


**图 2　CNN 网络结构**

**Fig. 2　Structure of CNN network**

基于 CNN 的语音增强算法的基本流程如下：

(1)将预处理语音与原语音信号进行短时快速傅里叶变换,得到其频域表示,而神经网络对实数的处理能力较强,要对信号功率谱做实数部分采集,作为神经网络输入层数据。

(2)经过输入层进入卷积层做特征提取,进行局部特征提取,学习干净语音的特征,有效地保留输入特征图的空间结构信息,并通过训练学习得到有用的特征表示。

(3)进入池化层(如最大池化)用于减小特征图的维度,降低计算复杂度,同时保留关键信息。重复卷积和池化操作增强模型的学习能力,得到最终的语音表示。

(4)再进入全连接层,将卷积层和池化层提取到的特征图展平为一维向量,并通过一系列全连接操作得到最终的输出。

传统的 CNN 的语音增强技术也存在一些问题。CNN 主要关注局部特征的提取,对于语音信号中的长期依赖关系建模不足。因此本文提出了基于 IDEDFCNN 的语音增强改进算法。

### 2.2　基于 IDEDFCNN 的语音增强改进算法

由于 Lee 等学者利用 R-CED(R-Convolution Encode Decode,R-CED)网络实现了 babble 噪声下助听器语音数据的增强,因此本文使用相应的卷积神经网络架构,即升降编解码全卷积神经网络(IDEDFCNN)来实现对噪声环境下的语音增强功能。

采用传统的 FCN 的语音增强,对于较长的语音序列,无法实现较大跨度的语音增强效果,使其无法有效取得预期效果。因此本项目采用改进的 CED 网络结构代替原卷积层。提出的编解码器结构在目前的深度学习环境下被多次引用,取得了较为不错的成效。

IDEDFCNN 语音增强算法、即改进的深度编码器-解码器全卷积神经网络,结合了编码器-解码器结构和全卷积神经网络(FCNN)的优势,旨在提取并恢复语音信号中的有用信息,同时抑制背景噪声和其他干扰。

由上文可知,解码器由对称分布的卷积层构成。为了根据编码器所提供的特征信息来重建增强信号,解码器还可能包含非线性激活函数(ReLU)和批归一化层,以增强模型的表达能力并加速训练过程。ReLU 函数的数学定义公式为：

$$f(x) = \max(0, x) \tag{3}$$

通过设置大于零域为线性函数有效解决了梯度下降问题。

模型的目标是最小化增强后语音与干净语音之间的某种损失函数,从而优化模型的参数。均方误差的定义公式如下：

$$\mathrm{MES} = \frac{\sum_{i=0}^{N} (T_{\mathrm{if}} - T)^2}{N} \tag{4}$$

其中,$T_{\mathrm{if}}$ 表示预测值,$T$ 表示真实值。MES 利用预测值与真实值的差值,计算相关误差,本次实验的训练指标就是根据该误差进行评判。

升降编解码全卷积神经网络结构如图 3 所示。由图 3 可知,输入语音为 8 帧信号,经过该网络输出 1 帧信号。前 5 层为编码器,后 4 层为解码器,最后 1 层做卷积操作。

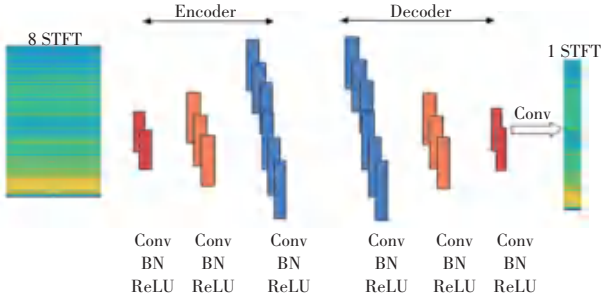


**图 3　升降编解码全卷积神经网络结构**

**Fig. 3　Increase Decrease Encoder Decode Fully Convolutional Neural Network architecture**

一旦模型训练完成,就可以用于实时语音增强。在推理阶段,给定一个带噪语音信号,IDEDFCNN 模型将其输入到编码器中,提取特征并传递给解码器。解码器根据这些特征重建增强后的语音信号,并输出。由此推得波形合成的公式如下：

$$S = S_{\mathrm{dB}} \cdot S(\lambda,\theta) \tag{5}$$

$$S_{\mathrm{out}} = \sum_{i=0}^{l} i\mathrm{STFT}[S,S'] \tag{6}$$

其中，$S_{\mathrm{dB}}$ 表示信号的分贝尺度，$S(\lambda,\theta)$ 表示信号的相位谱，得到频域合成语音。

此后取信号的共轭，组合成数组取逆变换得到时域频谱，最后重叠叠加，合成得到增强后语音信号。

综上所述，该算法将输入的待处理的语音经过分帧加窗归一化等操作，再经由傅里叶变换取功率谱和对数域幅度谱实数特征提取，规定连续 8 帧信号后，输入到卷积层中，经由 IDEDFCNN 算法对相邻帧的信息进行建模，由编码器交付于解码器，利用上下文信息加于待增强语音帧，去除噪声，还原原信号，实现语音增强。

## 3 实验结果

### 3.1 实验环境设置

本项目是在 Windows10 系统下，使用 TensorFlow 模型，语音数据集采用 TIMIT 语音集，训练集为 9 600 条语音，测试集为 3 200 条语音，噪声数据集采用 ESC-50 噪声集，训练集和测试集均为 4 000 条。在 0 dB 信噪比时，随机添加噪声来增强鲁棒性测试集。

模型的训练采用升降编解码全卷积神经网络进行训练，学习设置为 200 000 轮，批大小为 128，学习率设置为 $a = 0.01$，损失函数为均方误差损失函数。语音信号的相关输入是根据 256 点的 FFT 取有效点，共 8 帧所构成的矩阵，整个特征提取环节设置 10 个卷积层，前 5 个为编码器，后 4 个为解码器，最后为生成语音输出层。训练轮次设为 $2\times10^9$ 轮次，学习率最初设置为 $a = 0.001$，在训练过程中，若验证其损失率 4 次变化不大，则以 $a/2$，$a/4$ 逐步减小学习率形成较为优质模型。

### 3.2 实验结果分析

增强阶段输入不同噪声的含噪语音数据，进行预加重、分帧加窗和特征提取后输入到预训练好的模型中，进行增强和语音重构处理，获得增强后的数据。

在训练阶段，将处理后的信号输入到对应的网络模型后，在训练过程中每 100 轮计算一次均方误差估计，查看语音模型的训练效果，通过多次训练得到语音增强模型，可以选取不同轮次进行语音增强效果的对比，本次训练采用 160 000 轮次模型进行语音增强。经由网络训练处理的语音帧重新叠加合成出信号。

在进行对比实验之前，利用选取好的测试语音对本项目模型进行测试，增强效果如图 4 所示。由图 4 可以观察到增强后的语音相比原语音，效果差别不是很大，对于噪声消除的效果较为明显，因此本项目的语音增强算法对噪声的削弱具有较好的效果。后续将设置 FCN 与 IDEDFCNN 的对比实验，比较两者差异以及相关评价指标的对比。

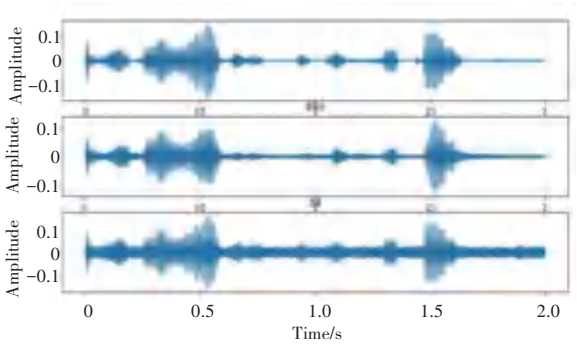


**图 4 较高噪声滤除效果图**

**Fig. 4 Diagram of higher noise reduction effect**

本次对比实验添加对比的噪声为键盘声和电流声。本实验对不同种类噪声进行选取，观察在不同两算法之间的增强效果有何影响。通过实验验证，得到对比结果见表 1。从表 1 观察可得，IDEDFCNN 算法与 FCN 相比，对于键盘噪声，IDEDFCNN 的 STOI 比 FCN 增加了 0.13，对于电流声增加了 0.08，平均值增加了大约 0.1，本文提出的 IDEDFCNN 能够实现较好的语音增强效果，提高语音的质量和可懂度。

**表 1 FCN 与 IDEFCNN 增强对比**

**Table 1 Comparison between FCN and IDEFCNN enhancements**

| 算法 | STOI(键盘声) | STOI(电流声) | 平均值 |
|---|---|---|---|
| FCN | 0.56 | 0.58 | 0.57 |
| IDEFCNN | 0.69 | 0.66 | 0.67 |

为增强前语音波形和经由 2 种算法得出的增强后的语音波形如图 5~图 7 所示。对比 3 个波形图，可以看出，FCN 对于噪声的去噪效果较弱，而 IDEDFCNN 的语音增强算法对噪声的消除明显提升，所以可知基于 IDEDFCNN 的语音增强算法是有效的，对语音增强有实际意义。

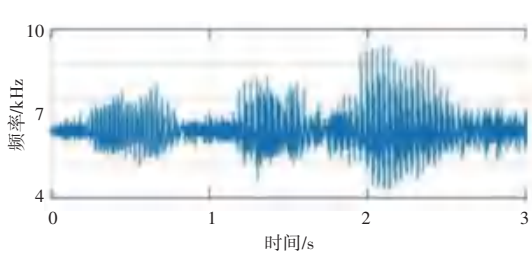


**图 5 增强前语音信号波形**

**Fig. 5 Speech signal waveform before enhancement**

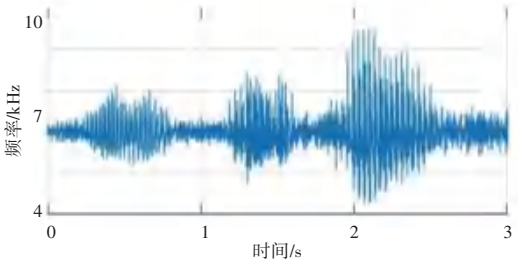


**图 6　FCN 增强后的语音信号波形**

**Fig. 6　Waveform of the speech signal enhanced by FCN**

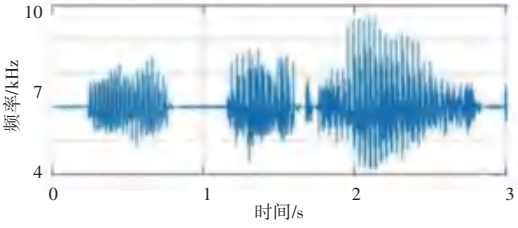


**图 7　IDEDFCNN 增强后的语音信号波形**

**Fig. 7　Waveform of the speech signal enhanced by IDEDFCNN**

## 4　结束语

本项目在 FCN 网络的基本架构下，利用 CED 模型进行语音降噪。但是由于 FCN 网络对于较长的语音序列，无法实现较大跨度的语音增强，使得其无法有效取得预期效果。

因此本文在原有网络的基础上，引入了基于升降编解码卷积神经网络，编码器在逐步将信号的高维特征进行提取保存处理、并输入后续编码，解码器卷积核压缩特征，获取语音信号的更为详细的频谱特征。通过观察损失函数的损失值的变化以及 STOI 参数，可以看出，在引入 IDEDFCNN 模型之后降噪效果更加明显，系统性能得到了一定的提升。并且 IDEDFCNN 语音增强方法具有以下优势：通过利用卷积神经网络的强大特征提取能力，IDEDFCNN 能够快速有效地处理语音信号，实现实时语音增强；IDEDFCNN 对不同类型的噪声和干扰具有较强的鲁棒性，能够在各种复杂环境下实现有效的语音增强。

目前，基于 IDEDFCNN 的语音增强技术在许多领域具有广泛的应用前景，如语音通信、语音识别、助听器设计以及智能家居等。通过这些应用，人们可以在嘈杂环境中获得更清晰、更易于理解的语音信号，从而提高通信质量和用户体验。

## 参考文献

[1] 吴卫鹏. 基于改进谱减的语音增强算法研究[D]. 南京：南京邮电大学，2019.
[2] 曹丽静. 步兵战车环境下战士口令识别研究[D]. 石家庄：河北经贸大学，2021.
[3] 孙思雨，张海剑，陈佳佳. 基于傅里叶卷积的多通道语音增强[J]. 无线电工程，2024，54(3)：580-588.
[4] 朱智慧. 基于深度学习的噪声鲁棒性语音识别技术研究[D]. 成都：四川大学，2023.
[5] 祝晓晨. 基于深度学习的单通道语音增强模型研究与实现[D]. 天津：天津理工大学，2023.
[6] 胡少东. 面向语音增强的深度神经网络结构优化研究[D]. 淄博：山东理工大学，2022.
[7] 夏俊杰. 基于深度学习的语音增强算法研究[D]. 重庆：重庆邮电大学，2022.
[8] 王潇. 基于深度学习的环境噪声下语音增强技术研究[D]. 成都：西华大学，2022.
[9] 李鑫元. 基于深度神经网络的单通道语音增强方法研究[D]. 西宁：青海师范大学，2022.
[10] ZHANG Wangyou，CHANG Xuankai，BOEDDEKER C. End-to-end dereverberation，beamforming，and speech recognition in a cocktail party [J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2022，30：3173-3188.
[11] 景浩. 基于深度学习的语音增强方法研究[D]. 焦作：河南理工大学，2021.
[12] 徐琅. 基于深度神经网络的单通道语音增强算法研究[D]. 赣州：江西理工大学，2021.
[13] 李如玮，李秋艳，赵丰年，等. 基于注意力和深度学习的双耳语音增强算法[J]. 华中科技大学学报(自然科学版)，2023，51(9)：125-131.
[14] 孔凡留. 基于深度学习的语音增强算法研究[D]. 南京：东南大学，2021.
[15] 李劲东. 基于深度学习的单通道语音增强研究[D]. 呼和浩特：内蒙古大学，2020.
[16] 李斌. 基于深度神经网络的单通道语音增强方法研究[D]. 杭州：浙江大学，2020.
[17] 彭川. 基于深度学习的语音增强算法研究与实现[D]. 成都：电子科技大学，2020.
[18] 鲍长春，项扬. 基于深度神经网络的单通道语音增强方法回顾[J]. 信号处理，2019，35(12)：1931-1941.
[19] 许雯婷，龚晓峰. 基于深度全卷积神经弹性网络 WCGAN-GP 模型的语音增强研究[J]. 计算机应用与软件，2024，41(2)：130-137.
[20] 邵晓光. 深度学习在语音增强技术中的应用研究[D]. 大庆：东北石油大学，2018.