Computer Science and Application 计算机科学与应用, 2024, 14(4), 230-241
Published Online April 2024 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2024.144093

Hans汉斯

基于CBAM-CRN的面向会议场景的多通道回声消除模型

孙慧冰[1]，丁碧云[1]，孙成立[2*]

[1]南昌航空大学信息工程学院，江西 南昌
[2]广州航海学院信息与通信工程学院，广东 广州

收稿日期：2024年3月20日；录用日期：2024年4月18日；发布日期：2024年4月25日

摘　要

本文研究了基于深度学习的多通道回声消除方法，提出了基于卷积块注意力模块(CBAM)融合卷积循环网络(CRN)的多通道回声消除方法。该方法利用U型网络的特征提取能力和LSTM网络处理时序信号的优势，结合了时频掩蔽算法和稀疏自适应归一化处理，同时融合了通道注意力和空间注意力联合机制，该混合域注意力能够有效地捕获关键特征并抑制无关特征。实验表明，CBAM-CRN方法在多种通话模式下均优于自适应滤波和其他深度学习方法，有效提高了远场免提通话的语音质量。

关键词

深度学习，多通道回声消除，U型网络，混合域注意力

Multi-Channel Echo Cancellation Model for Conference Scenarios Based on CBAM-CRN

Huibing Sun[1], Biyun Ding[1], Chengli Sun[2*]

[1]School of Information and Engineering, Nanchang Hangkong University, Nanchang Jiangxi
[2]School of Information and Communication Engineering, Guangzhou Maritime University, Guangzhou Guangdong

Received: Mar. 20th, 2024; accepted: Apr. 18th, 2024; published: Apr. 25th, 2024

Abstract

In this paper, we study the multi-channel echo cancellation method based on deep learning for

*通讯作者。

文章引用: 孙慧冰, 丁碧云, 孙成立. 基于 CBAM-CRN 的面向会议场景的多通道回声消除模型[J]. 计算机科学与应用, 2024, 14(4): 230-241. DOI: 10.12677/csa.2024.144093

acoustic echo problem, and propose a multi-channel echo cancellation method based on convolutional block attention module (CBAM) and convolutional recurrent network (CRN). This method takes advantage of the feature extraction ability of U-Net and the advantages of LSTM network in processing time series signals, combines the time-frequency masking algorithm and sparse adaptive normalization processing, and fuses the channel attention and spatial attention joint mechanism, the hybrid domain attention can effectively capture key features and suppress irrelevant features. Experimental results show that the CBAM-CRN method is superior to adaptive filtering and other deep learning methods in various call modes, and effectively improves the voice quality of far field hands-free calls.

Keywords

Deep Learning, Multi-Channel Echo Cancellation, U-Net, Mixed Domain Attention

1. 引言

回声消除是声学前端信号处理的重要环节，不管是远程音视频会议系统还是智能语音终端产品，在通话进行过程中由于现实环境复杂多变，语音信号可能会被回声掩盖。为了提高通信系统的音频质量，声学回声消除成为语音交互系统的必要条件，科研人员对声学前端信号处理算法越来越重视。回声消除的主要任务：一是消除扬声器和麦克风的耦合回声，二是消除扬声器播放的信号经过多路径反射后又被麦克风采集而形成的回声。现代智能语音设备多数采用麦克风阵列设计，当麦克风和扬声器的数量增加时，就会出现多通道回声消除(Multi-channel Echo Cancellation, MCAEC)的问题[1]，如果继续采用单通道回声消除的方法对每个麦克风逐一处理，这势必会对语音质量产生不良的影响，且结合麦克风阵列技术和回声消除技术的模型也难达到理想的效果。因此，如何在低信噪比的复杂环境下，有效提高目标语音的音质和可懂度是值得探究的，回声消除尤其是多通道回声消除技术成为智能语音信号处理的重点研究方向，对于推动社会发展和提升人们生活质量具有重要意义。

回声消除的研究开始于二十世纪五十年代，经历多年的积累和创新，在不同的应用场景中的回声消除相继被提出，根据回声消除的处理方式不同，分为传统自适应滤波器算法和基于深度学习方法。传统自适应滤波可分为三大类：时域类 LMS 算法[2]、频域类 LMS 算法[3]和最小二乘(Least Square, LS)类算法[4]及其。在复杂的声学环境下，传统方法的信号估计不准确，需要处理延迟估计、双端检测和非线性残余回声等问题。近些年，随着深度学习的高速发展，基于深度学习的方法成为声学回声消除(Acoustic Echo Cancellation, AEC)的重要手段，在语音处理领域广泛应用，深度神经网络(Deep Nueral Network, DNN)擅长建立高维向量之间的非线性映射关系，因此基于深度学习的 AEC 系统展现出巨大潜力。其中基于端到端的网络如因果 DTLN (Dual-Signal Transformation LSTM Network) [5]模型在实时声学回声消除方面也表现出了优越性。依据有监督语音分离的理念，文献[6]提出了一种回声消除方法，该方法以双向长短期记忆神经网络(Bidirectional Long Short-Term Memory, BLSTM)结合理想比率掩蔽算法为基础。而根据输入信号的形式不同，近年来也有将语音波形作为输入的模型。比如生成对抗网络[7]。对于回声消除任务，UNet 模型的演变也备受关注，Jung-Hee Kim 和 Joon-Hyuk Chang 提出了具有注意力机制的 Attention Wave-U-Net [8]模型的回声消除方法，直接将时域语音波形输入模型进行回声消除。

针对多通道回声消除问题，文献[9]提出了出了一种基于 CRN 网络和波束形成的多麦克风和多通道 AEC 方法。该方法通过结合波束形成技术实现多麦克风的信号采集和处理，同时还采用多通道 AEC 模块对回声信号进行估计和抑制。通过这些措施的结合，该方法能够有效地消除多通道环境中的回声干扰。在 2023 年，文献[10]提出了一种两阶段的基于 CCRN 网络的立体回声消除方法，该方法将回声信号作为先验知识，通过在复数域中实现回声消除。文献[11]提出了多通道降噪和回声消除的二级系统，该系统利用多通道维纳滤波器进行降噪，然后估计回声路径。与传统算法不同的是，深度学习的回声消除模型不再需要双端检测模块，这给模型的搭建带来了极大的简化。

但是基于深度学习的回声消除方法需要收集大量数据集进行训练。随着网络层数的增加，其参数存储量也越大。此外，深度神经网络训练的目标通常只包含幅度谱，而未充分挖掘相位谱所涵盖的信息。因此，如何提升网络训练速度并实现有效回声消除，成为目前研究的重难点。在智能语音设备的设计中，麦克风阵列被广泛采用，这将把复杂的 AEC 问题转化为多通道回声消除问题。相比单通道 AEC 算法，基于多通道的 AEC 算法可能会引入非唯一性问题。

本文旨在解决回声消除中存在的问题，提出了一种嵌入混合域注意力机制的 CBAM-CRN 模型。在回声消除任务中应用了时频掩蔽算法，设计了基于深度学习模型的多通道回声消除方法，并获得了较好的语音质量和回声消除效果。

2. 信号模型

基于深度学习的多通道回声消除原理图如图 1 所示。

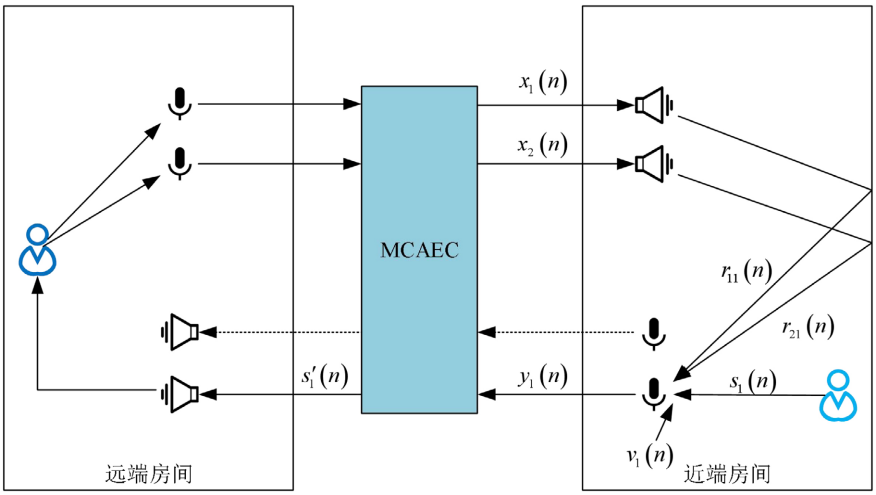

Figure 1. Multi-channel echo cancellation schematic based on deep learning
图 1. 基于深度学习的多通道回声消除原理图

近端麦克风信号的数学模型如式 1 所示，其中，$x(n)$和 $s(n)$分别代表远端和近端说话人语音，j 代表麦克风数量，$r_{ij}(n)$代表近端扬声器语音经过第 i 条回声路径到达第 j 个麦克风的回声信号，$y(n)$代表近端麦克风混合语音，$s'(n)$为 MCAEC 系统估计的近端语音。

$$y_j(n)=s_{ij}(n)+\sum_{ij} r_{ij}(n)\ j=1,2,3,\cdots,M \tag{1}$$

3. 基于 CBAM-CRN 的时频掩蔽多通道回声消除算法

现已有很多研究表明，各种注意力机制嵌入到 CNN 中可以明显改善网络性能，主要有 SENet、BAM、

ECANet 等注意力机制。但这些只能在通道维度上获取局部信息，CBAM 是一种新型且轻量的结合通道和空间的注意力模块，该模块不仅拥有强有力的特征表示能力，而且能够覆盖更多的语音特征，在语音帧之间可以建立多通道的相关性信息，减少冗余的参数，增强 CRN 网络的泛化性能。

本文提出了一种基于 CBAM-CRN 模型的时频掩蔽多通道回声消除方法，该方法结合了时频掩蔽算法和 CBAM-CRN 模型，并且在训练网络之前，将语音数据集进行自适应归一化处理，这样做的目的是避免神经网络由于大量数据的繁杂性而不能精准建模，减少了人工选择合适的归一化操作的实验。图 2 详细描述了该方法的实现过程，同时在算法 1 中提供了具体的学习算法

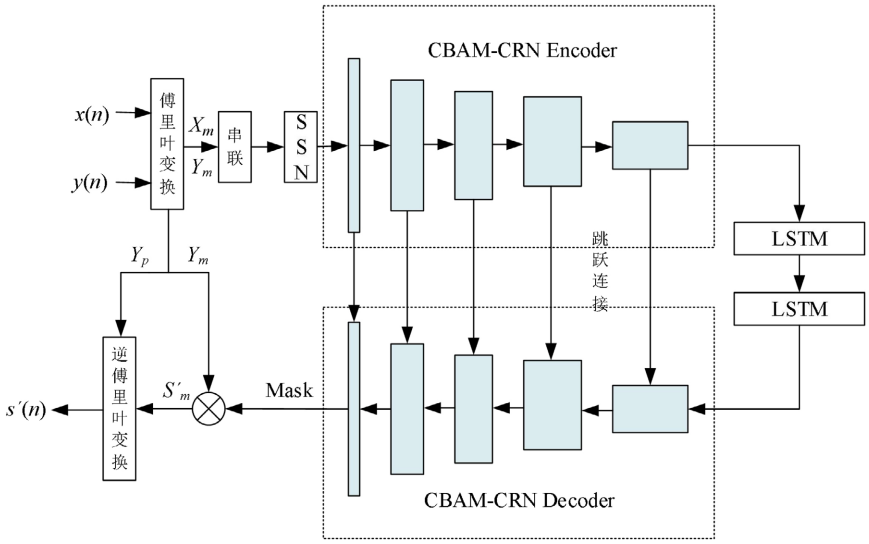

Figure 2. Time-frequency masked multi-channel echo cancellation method based on CBAM-CRN model

图 2. 基于 CBAM-CRN 模型的时频掩蔽多通道回声消除方法

算法 1：基于 CBAM-CRN 模型的时频掩蔽多通道回声消除算法

初始化：初始化 CBAM-CRN 模型，batch size = 4；

第一步：首先对语音进行切割，远端信号作为参考信号，然后对输入的近端麦克风信号 $y(n)$与远端信号 $x(n)$进行 STFT 处理，分别得到相应的幅度谱 Y_m 和 X_m 以及近端麦克风语音的相位谱 Y_p；

第二步：将六个通道的 Y_m 和一个通道的 X_m 进行幅度谱串联，此时输入特征通道数为 7，之后通过稀疏自适应归一化(SSN)处理，其输出作为 CBAM-CRN 网络的输入进行迭代训练，而估计的近端语音幅度谱 S'_m 是由近端麦克风语音幅度谱 Y_m 与 CBAM-CRN 网络预测的掩蔽值 IRM 相乘得到的，其定义如式 2 所示，其中⊗代表哈达玛乘积：

$$S'_m = IRM \otimes Y_m \tag{2}$$

第三步：将网络估计的近端语音幅度谱 S'_m 与近端麦克风语音的相位谱 Y_p 进行重构，经过 ISTFT 得到回声消除后的近端信号 $s'(n)$；

第四步：将回声消除后的近端信号 $s'(n)$与纯净的近端语音 $s(n)$代入到损失函数 SI-SDR 公式中计算，保存 SI-SDR 的值；

第五步：选择 Adam 优化器优化 CBAM-CRN 模型的参数，接着重复上述所有的步骤，当 SI-SDR 损失函数的值收敛到最大值时，停止训练网络，保存具有最佳参数的 CBAM-CRN 模型；

第六步：对回声消除模型的性能进行评估。

4. 实验结果与分析

4.1. 实验环境配置

4.1.1. 实验设备设置

本文的实验在 Windows 10 系统上使用 PyCharm 软件的 Community Edition 2019.1.1 x64 版本进行调试和运行完成。实验使用的计算机处理器(CPU)为 AMD Ryzen5 2600X Six-Core 处理器，主频为 3.60 GHz，内存容量是 16 GB；图形处理器(GPU)是 NVIDIA RTX2060 型号，显存为 6 GB。实验采用 Pytorch 框架搭建本文提出的深度神经网络模型，语音的波形图和语谱图等图像使用软件 matplotlib 库绘制。

4.1.2. 语音数据集设置

实验中使用的语音数据集选取 LibriSpeech 语音库[12]，该数据库包含 1000 小时的多说话人的图书朗读音频，音频片段是以 16 kHZ 采样的，平均长度为 5 s，而添加的噪声有来自 WSJ0 Hipster Ambient Mixtures (WHAM!) [13]噪声库，WHAM!噪声库相比 wsj0-2mix 噪声库含有更加复杂的噪声，含有很多实际场景的背景噪声，比如旧金山湾的 cafe、restaurant、bar 等。LibriSpeech 语音库和 WHAM!噪声库的数据集组成如表 1、表 2 所示。

Table 1. Dataset of LibriSpeech speech library
表 1. LibriSpeech 语音库的数据集

数据集	总时长/h	说话人平均时长/min	女说话者数量/个	男说话者数量/个	总人数数量/个
验证集	5.4	8	20	20	40
测试集	5.4	8	20	20	40
训练集	100.6	25	125	126	251

Table 2. WHAM! noise library data set
表 2. WHAM!噪声库的数据集

数据集	总时长/h	语音片段数量
train	58	20,000
dev	14.7	5000
test	9	3000

针对远场信号的回声消除问题，实验需要模拟真实的远场免提通话的会议场景，从数据库平台的语料库中选择数据集，将其合成新的数据集。首先利用图形方法构造一个房间配置生成器，随机生成 20 个不同尺寸的房间。混响参数 RT60 用于模拟房间材料对声源吸收的系数，设置为 0.26 s~0.7 s。在房间内随机选取不同的位置以模拟说话人的位置，较大的房间采样 600 个点，较小的房间采样 250 个点。此外，配置 6 个麦克风作为 6 个输入通道，其中相邻麦克风之间的孔距为 50 mm，放置麦克风的高度设置为 1.2~1.8 m，左右两个扬声器按照 30~45 度对称放置，固定麦克风和扬声器的位置。

从 LibriSpeech 语音库中获取纯净语音作为近端和远端说话人的语音，用于构建训练、验证和测试集。每个集合都包括近端语音、近端麦克风语音、远端语音、回声信号和参考信号。将含有噪声和回声的混合近端语音作为近端麦克风信号，为了产生回声，需要借助 RIR 房间脉冲响应与远端信号卷积，加入噪声之后仿真出回声，而远端信号加入远场噪声生成参考信号。通过混合 100 小时的 LibriSpeech 语音库、噪声库 WHAM! (设置信噪比在−6 dB~+3 dB 之间)以及房间脉冲响应仿真的回声来创建混合语音数据集。

该数据集包括251对不同说话人的语音片段(包含125个女性说话人和126个男性说话人)与噪声片段的混合，共计139,00个训练混合语音集。

混合语音测试集由多个数据源混合而成，包括LibriSpeech语音库的40对不同说话人朗读片段(包含20个男性说话人和20个女性说话人)、WHAM!噪声库的3种不同类型噪声(cafe、restaurant、bar)以及房间脉冲响应仿真的回声。测试集共包含3000种语音，其中20个男性说话人和20个女性说话人组成性别平衡的远端语音，近端语音则随机选择40个说话人，构成40个近端与远端语音对。此外，本实验将语音片段的采样频率降到16 kHZ，时间间隔为4 s。会议通话状态设为三种：近端通话、远端通话和双端通话，每种状态下使用的数据集为所有数据集的三分之一，包括训练集、验证集以及测试集。在语音预处理中，采用512点离散傅里叶变换和汉明窗函数对所有语音片段进行分帧，帧长为512个采样点，帧移为256个采样点。

4.1.3. 模型参数设置

实验中模型的优化器选择Adam，该优化器初始化的学习率为10^{-4}，训练的迭代周期设为100 epoch，为了避免训练过程卡顿，批处理batchsize设置为4。如果在连续3个迭代epoch内验证集的损失值没有发生变化，学习率就逐次减半；而当连续10个迭代epoch内验证集的损失值没有上升趋势时，训练提前截止，训练损失函数选择SI-SDR。CBAM-CRN网络结构及参数如表3所示。

Table 3. Network structure and parameters of CBAM-CRN
表3. CBAM-CRN的网络结构及参数

网络层结构	输入输出通道大小	输入维度	卷积核大小	步长	输出维度
Conv2d_1	7 × 16	B × 7 × T × 257	3 × 3	(1, 2)	B × 16 × T × 128
Conv2d_2	16 × 16	B × 16 × T × 128	3 × 3	(1, 2)	B × 16 × T × 63
Conv2d_3	16 × 32	B × 16 × T × 63	3 × 3	(1, 2)	B × 32 × T × 31
Conv2d_4	32 × 32	B × 32 × T × 31	3 × 3	(1, 2)	B × 32 × T × 15
Conv2d_5	32 × 64	B × 32 × T × 15	3 × 3	(1, 2)	B × 64 × T × 7
Reshape	——	B × 64 × T × 7	——	——	B × T × 448
LSTM_1	——	B × T × 448	——	——	B × T × 448
LSTM_2	——	B × T × 448	——	——	B × T × 448
Reshape	——	B × T × 448	——	——	B × 64 × T × 7
TransConv2d_1	128 × 32	B × 128 × T × 7	3 × 3	(1, 2)	B × 32 × T × 15
TransConv2d_2	64 × 32	B × 64 × T × 15	3 × 3	(1, 2)	B × 32 × T × 31
TransConv2d_3	64 × 16	B × 64 × T × 31	3 × 3	(1, 2)	B × 16 × T × 63
TransConv2d_4	32 × 16	B × 32 × T × 63	3 × 3	(1, 2)	B × 16 × T × 128
TransConv2d_5	32 × 6	B × 32 × T × 257	3 × 3	(1, 2)	B × 6 × T × 257

表中Conv2d、LSTM、TransConv2d分别表示二维卷积层、长短期记忆网络层和二维转置卷积层，reshape表示对数据维度进行重新调整。由表3可知，卷积层的输入特征通道数是转置卷积层的一半，输出的特征维度逐层减半，而解码器的转置卷积层输出特征维度逐层加倍。此外，每个单层卷积都添加了CBAM注意力模块，并将CBAM模块的缩放参数ratio设置为16。输入数据包含6个通道的混合语音信号和1个通道的远端参考语音，因此输入特征图的通道为7。网络输入输出维度的参数表示为B × C × T × F (批量数 ×

特征通道数 × 帧数 × 特征维度)，5 个卷积模块的通道数为[16,16,32,32,64]，5 个反卷积的通道数为[32,32,16,16,6]。根据 STFT 切割的帧长度和 batch size 设置，每批次的输入特征维度为 4 × 7 × 251 × 257。

4.2. 回声消除算法的性能评价方法

目前业界通用的回声消除算法性能评价指标主要分为主观评价和客观评价。前者依据人耳对增强语音的听觉感受，侧重感性认识，而后者利用数学统计方式计算出增强语音质量的得分，更客观科学地衡量回声消除方法的性能。

语音的客观评价方法主要是通过算法对测试语音的质量进行评价.不同场景下通话系统的输出信号代表的意义分为三种情况：一是只有近端单讲时输出信号要求和麦克风采集的近端语音信号保持一致，二是只有远端单讲时要将输出信号尽可能抑制，三是双端对讲时既要抑制回声又要保留近端说话人的声音。因此，单讲状态和双端对讲状态下的客观评价标准不同。单讲状态下用回声返回损失增益值 ERLE (Echo Return Loss Enhancement) [14]测试回声抑制量。双端对讲状态下的客观评价标准主要有语音质量感知评价 PESQ (Perceptual Evaluation of Speech Quality) [15]、短时客观可懂度 STOI (Short-Time Objective Intelligibility) [16]、信号失真比 SDR (Signal to Distortion Ratio)、尺度不变信号失真比 SI-SDR (Scale-invariant Signal to Distortion Ratio) [17]等。

回声返回损失增益值 ERLE：ERLE 是回声消除单通话工作状态下(只有回声信号，无近端说话人语音)特有的评价准则，*ERLE* 值代表回声信号 $d(t)$和残留回声 $e(t)$的能量比值，ERLE 值越大表示回声消除算法性能越好。其定义如式(3)所示：

$$ERLE\left(dB\right)=10\log_{10}\left\{\frac{\varepsilon\left[d^{2}\left(t\right)\right]}{\varepsilon\left[e^{2}\left(t\right)\right]}\right\} \tag{3}$$

其中 $\varepsilon[\ ]$表示统计参数的期望值。

语音质量感知评价 PESQ：PESQ 主要是对回声消除后的语音进行衡量其失真程度，是最常用的语音质量评价指标之一，计算流程包括语音信号电平调准、IRS (Intermediate Reference System)滤波、时间对齐、听觉变换处理、掩蔽预测结果等。其计算公式如式(4)所示：

$$PESQ=4.5-0.1dsym-0.0309dAsym \tag{4}$$

其中，*dsym* 和 *dAsym* 分别是听觉变换处理后产生的谱失真测度，称为对称与非对称干扰，用途为平衡测试时的表征能力和测试精度，PESQ 值的波动范围为[−0.5,4.5]。

短时客观可懂度 STOI：在一定程度上可以反映极短时间内(一般为 300~400 毫秒)语音被理解的程度。测试语音能完全被理解的 STOI 定义为 1，不能被理解的 STOI 定义为 0，在[0,1]范围内，得分越高，说明回声消除后的语音能被理解，模型的性能也越好。

信号失真比 SDR：SDR 的物理意义描述的是参考语音 y 和模型输出的语音 y'接近的程度，其定义如式(5)所示：

$$SDR=20\log_{10}\frac{\|y\|}{\|(y-y')\|} \tag{5}$$

尺度不变信号失真比 SI-SDR：SI-SDR 信号失真比 SDR 改进后的性能衡量指标。SI-SDR 的表达式如下式(6)所示，将模型生成的语音向量 y'进行分解，其中 y'_t 和 y'_e 代表平行分量和垂直分量。

$$SI-SDR=10\log\left(\frac{\|y'_t\|^2}{\|y'_e\|^2}\right) \tag{6}$$

$$y_t' = \frac{\langle y', y \rangle y}{\|y'\|^2} \tag{7}$$

$$y_e' = y' - y_t' \tag{8}$$

4.3. 实验结果

本文选取远场实际免提通话情况下输出的语音进行分析，分别是双端对讲、远端单讲、近端单讲三种不同阶段，并将本文提出的基于 CBAM-CRN 的回声消除方法分别与传统算法和深度学习方法进行比较，传统算法是指传统自适应滤波器算法，例如 LMS、NLMS、PBFDAF 等，深度学习方法包括基于 LSTM 的回声消除方法、基于时间卷积网络[18] (Temporal Convolutional Networks, TCN)的回声消除方法、基于 DC-UNet 的回声消除方法、以及基于 CRN 和 DC-CRN 的回声消除方法。最后根据具体各性能评价方法对结果进行数值上的比较。

为证明 CBAM-CRN 模型消除回声的性能，实验将 CRN 和 CBAM-CRN 方法得出的评价指标与 LMS、NLMS、RLS、PBFDAF 等传统算法在统计数值上比较。

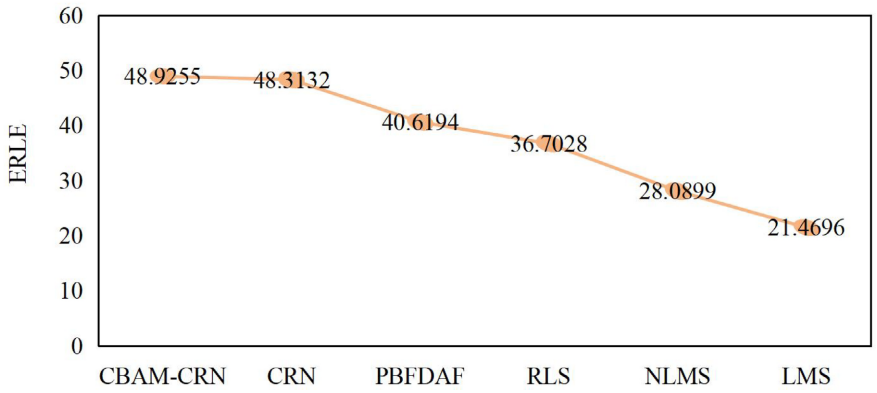

Figure 3. Comparison of single talk evaluation scores of traditional echo cancellation methods
图 3. 传统回声消除方法的单讲评价得分对比

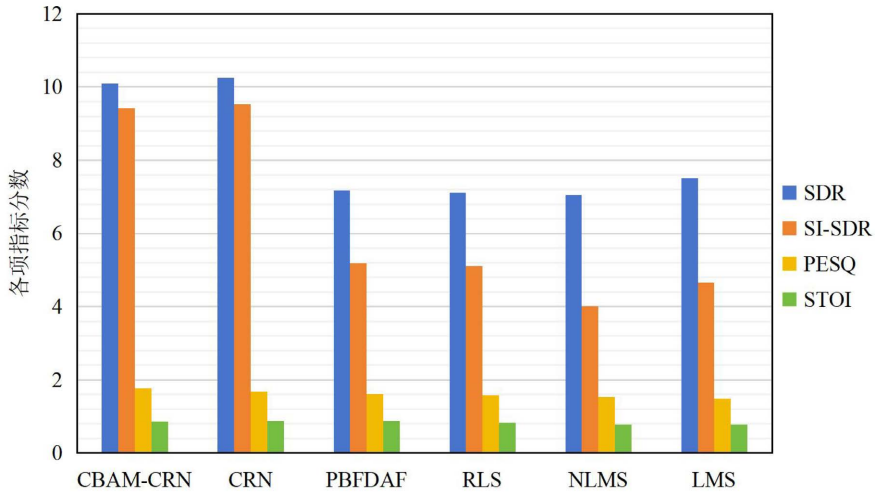

Figure 4. Comparison of evaluation scores of traditional echo cancellation methods in two-end intercom
图 4. 双端对讲情况下传统回声消除方法的评价得分对比

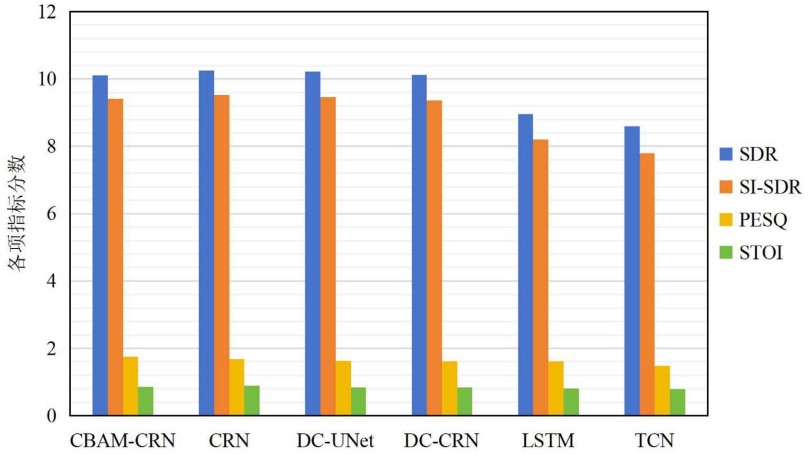

Figure 5. Comparison of index scores of each deep learning echo cancellation algorithm
图 5. 各深度学习的回声消除算法的指标分数对比

回声消除的性能指标主要包括 SDR、SI-SDR、PESQ、STOI 和 ERLE，其中 SDR 和 SI-SDR 反映消除干扰信号的程度，值越高，说明叠在近端纯净语音的干扰信号越小，输出的声音信号清晰，音质越高。PESQ 客观评价语音的质量，STOI 反映语音的可理解度，而 ERLE 反映通话系统中的回声抑制量。

通过折线图和柱状图可以进一步直观比较各种算法性能的指标增减幅度变化，具体如图 3、图 4、图 5 所示。由图 3 可以看出，当通话处于单讲时，CBAM-CRN 的方法相比自适应滤波算法的 ERLE 有显著增幅，其得分甚至超出 LMS 算法的 56%。此外，在折线图中 CBAM-CRN 方法的得分最高。由图 4、图 5 可知，CBAM-CRN 的方法的在 SDR 和 SI-SDR 得分方面比传统回声消除方法有很大的提升，但对比其他深度学习算法，PESQ 值增长的趋势缓慢。

表 4 和表 5 分别为 CBAM-CRN 方法与传统自适应滤波算法和深度学习算法的评分对比表。

Table 4. Index score table of various algorithms in double end intercom and single talk
表 4. 双端对讲和单讲情况下各种算法的指标得分表

算法	双端对讲				单讲
	SDR	SI-SDR	PESQ	STOI	ERLE
CBAM-CRN	**10.1005**	**9.4134**	**1.7651**	**0.8671**	**48.9255**
CRN	10.2448	9.5287	1.6748	0.8850	48.3132
PBFDAF	7.1847	5.1808	1.6141	0.8754	40.6194
RLS	7.1201	5.1023	1.5810	0.8315	36.7028
NLMS	7.0471	4.0119	1.5301	0.7792	28.0899
LMS	7.5150	4.6605	1.4873	0.7802	21.4696

Table 5. Index score table for various deep learning algorithms in double-ended intercom and single-talk cases
表 5. 双端对讲和单讲情况下各种深度学习算法的指标得分表

算法	双端对讲				单讲
	SDR	SI-SDR	PESQ	STOI	ERLE
CBAM-CRN	**10.1005**	**9.4134**	**1.7651**	**0.8671**	**48.9255**
CRN	10.2448	9.5287	1.6748	0.8850	48.3132

续表

DC-UNet	**10.2153**	**9.4662**	**1.6374**	**0.8506**	**48.3934**
DC-CRN	10.1135	9.3703	1.6212	0.8471	47.9150
LSTM	8.9536	8.2038	1.6110	0.8133	34.6746
TCN	8.5972	7.7994	1.4956	0.8022	48.3072

通过对比表 4 中的指标得分结果可知，LMS、NLMS、RLS、PBFDAF 各个自适应滤波算法在双端对讲阶段指标的变化并不是很明显。尽管单讲情况下 PBFDAF 的 ERLE 得分有所改善，但与本文的 CRN 和 CBAM-CRN 方法相比仍显劣势。具体来说，CRN 方法和 CBAM-CRN 方法的 SDR 和 SI-SDR 得分都高于传统自适应算法。值得思考的是，由于存在非线性声学回声的情况，传统算法在鲁棒性方面较差，导致自适应滤波在双端对讲过程中难以收敛，这也是导致自适应算法的 PESQ 数值较小的原因。在单讲条件下，CBAM-CRN 方法的 ERLE 得分都要高于其他传统算法，差距最大可达 27.46 dB。综合表中的数据分析，CBAM-CRN 算法的 SDR、SI-SDR、ERLE 和 PESQ 得分都要高于其他的传统回声消除算法。虽然本文提出的 CBAM-CRN 方法的 STOI 得分低于 PBFDAF 传统算法，但从波形图和语谱图的角度考虑，PBFDAF 自适应算法会造成原始语音信号的损失。

对比表 5 中各种深度学习算法，CBAM-CRN 方法的各项得分都比 DC-UNet、DC-CRN、LSTM、TCN 方法略高。CBAM-CRN 方法的 SDR 和 SI-SDR 分数远超过 LSTM 方法和 TCN 方法，其 PESQ 和 STOI 分数与 DC-UNet 和 DC-CRN 方法相比得到了提高。这说明复数域的单层 CRN 网络消除回声的程度没有含 CBAM 注意力的 CRN 网络提升大，特别是在单讲情况下，ERLE 分值增大。而相较于 CRN 和 CBAM-CRN 这两者方法而言，CBAM-CRN 方法的 ERLE 和 PESQ 值均略高于 CRN 方法，PESQ 得分提高了 0.0903，ERLE 提高了 0.6123 dB，从而体现出注意力机制的优势，说明添加了 CBAM 注意力的 CRN 模型的回声消除性能略胜一筹。

综上所述，对于多通道的回声消除场景，基于深度学习的方法无需添加去相关算法，因为不会产生通道之间的非唯一性问题。通过传统自适应算法和其他深度学习算法的比较，CBAM-CRN 方法在时域波形图、语谱图以及语音质量客观评价指标上都验证了有效性。本文提出的 CBAM-CRN 方法在这三个方面凸显了一定的优势，也证明了注意力机制的强大竞争力，表明结合通道域和空间域的注意力机制提升了网络模型的性能，在回声消除任务中有助于提升听觉质量，并且降低近端语音信号的失真度。

5. 结论

本文主要研究的问题是在多麦克风音视频会议远场情景下消除免提通话的声学回声，提出了基于 CBAM-CRN 的多通道回声消除方法。具体工作包括：为了使得深层网络捕获多尺度重要特征，在 CRN 模型的基础上，结合了通道注意力和空间注意力联合的 CBAM 网络以及稀疏自适应归一化(SSN)，采用时频掩蔽算法训练模型，提出了一种基于 CBAM-CRN 的时频掩蔽多通道回声消除方法。此外，介绍了将这种联合的 CBAM 注意力模块嵌入 CRN 网络的内部结构，详细阐述了该方法的原理框图和具体算法步骤。为了验证这种新型的模型能够提取目标语音的关键特征，利用强大的计算机建立模型进行实验，最后通过比较传统自适应滤波器算法和其他深度学习算法的实验结果，分析实验图像和数据结果，发现深度神经网络处理多通道回声问题的效果显著提升，得出 CBAM-CRN 方法提高了抑制回声的效果。虽然添加了注意力机制的 CBAM-CRN 提高了通道和空间的利用率，可以提高回声消除效果。但是，语音的高频部分容易损失。因此，后续的工作应该探索继续优化该模型。未来的研究方向应该探索如何将深度学习技术与其他信号处理技术相结合，以解决多通道回声消除技术中的挑战。

致　谢

本研究得到了南昌航空大学现信中心高性能计算服务的支持。此外，我要感谢在本文撰写时所有付出的人。首先，我要感谢我的导师孙成立教授。孙老师在学术上教会我如何阅读大量的英文文献和理解研究课题的程序，给我提供了很多学习资料和科研帮助，在他的指导下我学会了独立思考研究内容，汲取了很多前沿理论并且找到了自己的研究方向，感谢孙老师辛苦地指导本论文的撰写与修改，再次向孙老师表达诚挚的谢意！

其次，我要感谢实验室的小伙伴们，感谢我的师兄师姐、同届和师弟师妹们给我学习上的帮助和支持，在我遇到难题的时候帮我耐心解答。我也要感谢我的家人，感谢你们一直支持我，在生活中包容我理解我，希望你们永远健康平安，感谢你们的培养和无私奉献！

基金项目

国家自然科学基金(61861033)；江西省赣鄱俊才支持项目(20232BCJ22050)；江西省教育厅科技项目(DA202104170)；江西省自然科学基金重点项目(20202ACBL202007)；南昌航空大学博士启动基金(EA201904283)。

参考文献

[1] Sondhi, M.M. and Morgan, D.R. (1995) Stereophonic Acoustic Echo Cancellation—An Overview of the Fundamental Problem. *IEEE Signal Processing Letters*, **2**, 148-151. https://doi.org/10.1109/97.404129

[2] Widrow, B. and Hoff, M.E. (1960) Adaptive Switching Circuits. *Neurocomputing*, **4**, 126-134.

[3] Soo, J.S. and Pang, K.K. (1990) Multidelay Block Frequency Domain Adaptive Filter. *IEEE Transactions on Acoustics Speech & Signal Processing*, **38**, 373-376. https://doi.org/10.1109/29.103078

[4] Gilloire, A., Petillon, T. and Theodoridis, S. (1992) Acoustic Echo Cancellation Using Fast RLS Adaptive Filters with Reduced Complexity. 2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Toronto, 6-11 June 2021, 7138-7142.

[5] Westhausen, N.L. and Meyer, B.T. (2021) Acoustic Echo Cancellation with the Dual-Signal Transformation LSTM Network. *ICASSP* 2021-2021 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Toronto, 6-11 June 2021, 7138-7142. https://doi.org/10.1109/ICASSP39728.2021.9413510

[6] Zhang, H. and Wang, D. (2018) Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios. *Training*, **161**, 322. https://doi.org/10.21437/Interspeech.2018-1484

[7] Zhang, Y., *et al*. (2020) Generative Adversarial Network Based Acoustic Echo Cancellation. *Interspeech*, 3945-3949. https://doi.org/10.21437/Interspeech.2020-1454

[8] Kim, J.-H. and Chang, J.-H. (2020) Attention Wave-U-Net for Acoustic Echo Cancellation. *Interspeech*, 3969-3973. https://doi.org/10.21437/Interspeech.2020-3200

[9] Zhang, H. and Wang, D.L. (2021) A Deep Learning Approach to Multi-Channel and Multi-Microphone Acoustic Echo Cancellation. *Interspeech*, 1139-1143. https://doi.org/10.21437/Interspeech.2021-1508

[10] 程琳娟, 彭任华, 郑成诗, 等. 两阶段复数谱卷积循环网络立体声回声消除[J]. 声学学报, 2023, 48(1): 199-214.

[11] Ruiz, S., van Waterschoot, T. and Moonen, M. (2022) Cascade Multi-Channel Noise Reduction and Acoustic Feedback Cancellation. 2022 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Singapore, 23-27 May 2022, 676-680. https://doi.org/10.1109/ICASSP43922.2022.9747291

[12] Panayotov, V., *et al*. (2015) Librispeech: An ASR Corpus Based on Public Domain Audio Books. 2015 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), South Brisbane, 19-24 April 2015, 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964

[13] Wichern, G., *et al*. (2019) WHAM!: Extending Speech Separation to Noisy Environments. *Interspeech*, 1368-1372. https://doi.org/10.21437/Interspeech.2019-2821

[14] Breining, C., *et al*. (1999) Acoustic Echo Control. An Application of Very-High-Order Adaptive Filters. *IEEE Signal Processing Magazine*, **16**, 42-69. https://doi.org/10.1109/79.774933

[15] Rix, A.W., *et al*. (2001) Perceptual Evaluation of Speech Quality (PESQ)—A New Method for Speech Quality As-

sessment of Telephone Networks and Codecs. 2001 *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, 7-11 May 2001, 749-752.

[16] Taal, C.H., *et al*. (2010) A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech. 2010 *IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, 14-19 March 2010, 4214-4217. https://doi.org/10.1109/ICASSP.2010.5495701

[17] Le Roux, J., *et al*. (2019) SDR–Half-Baked or Well Done? 2019 *IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), Brighton, 12-17 May 2019, 626-630. https://doi.org/10.1109/ICASSP.2019.8683855

[18] Lea, C., *et al*. (2016) Temporal Convolutional Networks: A Unified Approach to Action Segmentation. In: Hua, G. and Jégou, H., Eds., *Computer Vision—ECCV* 2016 *Workshops. Lecture Notes in Computer Science*, Vol. 9915, Springer, Cham, 47-54.