复杂噪声环境下基于轻量化模型的车内交互语音增强和识别方法

廉筱峪,夏 楠*,戴高乐,杨红琴
(大连工业大学信息科学与工程学院,辽宁大连 116034)

摘 要: 针对车内语音交互在复杂噪声环境下识别率低以及难以在有限计算资源设备上部署问题,本文设计了轻量化的语音增强模型和语音识别模型并进行联合训练. 语音增强模型引入多尺度通道时频注意力模块来提取多尺度时频特征和各个维度上的关键信息. 在语音识别模型中提出了多头逐元素线性注意力,显著降低了注意力模块所需的计算复杂度. 实验表明,在自制数据集上这一联合训练模型表现出良好的噪声鲁棒性.

关键词: 深度学习;语音增强;语音识别;注意力机制;联合训练
基金项目: 教育部产学合作协同育人项目(No.220603231024713)
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112(2024)04-1282-06
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230905

An In-Vehicle Interaction Speech Enhancement and Recognition Method Based on Lightweight Models in Complex Environment

LIAN Xiao-yu, XIA Nan*, DAI Gao-le, YANG Hong-qin
(*School of Information Science and Engineering*, *Dalian Ploytechnic University*, *Dalian*, *Liaoning* 116034, *China*)

Abstract: In order to solve the problem of low recognition rate of in-vehicle voice interaction in complex noise environment and difficult deployment on devices with limited computing resources, this article proposes a lightweight and robust voice recognition method based on joint training framework in the noisy environment. The speech enhancement model introduces a multi-scale channel time-frequency attention module to extract multi-scale time-frequency features and key information in various dimensions. In the speech recognition model, multi-head element-wise linear attention is proposed, which significantly reduces the computational complexity required for the attention module. Experiments show that the joint training model shows good noise robustness on the self-made dataset.

Key words: deep learning; speech enhancement; speech recognition; attention mechanism; joint training
Foundation Item(s): Ministry of Education Industry-University Cooperation and Collaborative Education Project (No.220603231024713)

1 引言

在高级辅助驾驶系统中,语音识别提供了一种无需手动操作的交互方式,使驾驶员可以方便地获取信息并控制车辆. 但在行驶过程中车内存在多种复杂噪声的干扰导致语音识别算法的识别错误率高,同时车载系统在部署算法时需要确保算法足够轻量化,因此如何有效解决这两个问题成为构建车载语音识别系统的关键挑战.

为提高语音识别的噪声鲁棒性,通常情况下会将语音先进行去噪,再将去噪后的语音作为语音识别模型的输入,在实际的车载环境中,各种伴随着汽车启动后的噪声都在干扰着语音识别的效果,这些大量复杂的噪声极大影响了车载语音识别的准确性.

近年来,大多研究人员开始采用深度学习的方法来进行语音增强任务,文献[1]提出了注意力门控膨胀卷积网络,文献[2]使用频谱修复网络来进行语音增强,文献[3]则提出了一种基于协同注意力的神经网络用于语音增强. 上述方法虽然一定程度上提升了语音

收稿日期:2023-09-26;修回日期:2024-01-28;责任编辑:覃怀银
*通讯作者:夏楠

增强效果,但并未对语谱图上的时频特征进行充分利用,且部分方法涉及参数过多,难以适用于计算资源有限的语音增强任务. 同时,端到端语音识别技术简化了传统语音训练识别流程并能获取优异的识别结果. Conformer[4]将卷积和多头自注意力[5]相融合,显著提高了模型的识别能力,但其高额的计算量不适于在计算资源有限的车载平台部署,目前针对注意力复杂度的改进也大多都以损失精度为代价[6,7]. 且为了能同时更新两个模型上的权重,现有主流方法大多对其进行联合训练[8,9].

针对上述模型存在的问题,本文首先提出多尺度通道时频注意力(Multi-scale Channel Time-Frequency Attention, MCT-FA)融入到语音增强模型中,其次使用多头逐元素线性注意力(Multi-Head Element-wise Linear Attention,MHELA)来构建轻量化的自注意力模块,在不损失模型精度的前提下显著降低了多头自注意力的计算量.

2　本文方法

2.1　总体结构

假设语音信号与噪声信号相互独立,对于单声道语音增强,时域含噪信号$\boldsymbol{y}(t)$可以被定义为

$$\boldsymbol{y}(t)=\boldsymbol{s}(t)+\boldsymbol{u}(t) \tag{1}$$

其中,$\boldsymbol{s}(t)$和$\boldsymbol{u}(t)$分别代表纯净信号和噪声信号,t表示时间帧的索引值. 进一步利用短时傅里叶变换,将时域信号转换为时频谱图:

$$\boldsymbol{Y}(t,f)=\boldsymbol{S}(t,f)+\boldsymbol{U}(t,f) \tag{2}$$

其中,$\boldsymbol{Y}(t,f)$、$\boldsymbol{S}(t,f)$和$\boldsymbol{U}(t,f)$分别为$\boldsymbol{y}(t)$、$\boldsymbol{s}(t)$和$\boldsymbol{u}(t)$的时频谱图;f表示频率域的索引值. 进一步将含噪语音的对数功率谱作为联合训练网络的输入特征图,模型总体架构如图1所示.

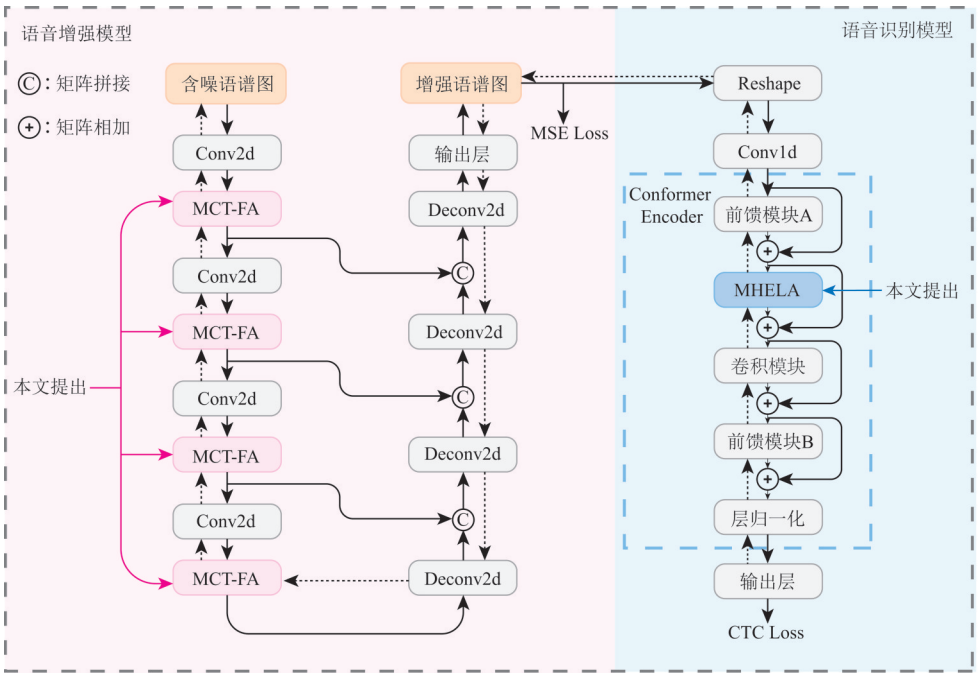


图1　模型总体架构

语音增强模型主要采用编解码架构实现,网络的编码器由二维卷积层与MCT-FA组成,每个编码层输出的特征图在时间维度上保持不变,在频率维度上逐渐减小,每一个卷积编码层提取到的特征与相应反卷积解码层进行上采样后的特征进行拼接融合后送入到下一个解码层中进行逐步还原.

语音识别模型则主要借鉴Conformer[4]编码器架构来进行设计,因为Conformer模块中的多头自注意力计算复杂度较高,不适合部署于计算资源有限的设备中. 本文则提出MHELA来对其进行改进,显著降低了该模块所需的计算量.

2.2　多头逐元素线性注意力

多头自注意力的计算过程可分为三个步骤:相似度计算、Softmax归一化、注意力权重矩阵与$\boldsymbol{V}$矩阵相乘进行加权求和,这三个步骤导致自注意力计算复杂度为序列长度的平方. 此外对输入序列使用线性层权重

$\boldsymbol{W}_h^Q, \boldsymbol{W}_h^K, \boldsymbol{W}_h^V \in \mathbb{R}^{d_m \times d_k}$映射为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$矩阵和输出映射层也占据着不可忽视的计算量.

本文提出MHELA来解决多头自注意力计算复杂度较高的问题. 如图2所示,对于含有H个头的自注意力,首先对输入序列$\boldsymbol{X} \in \mathbb{R}^{T \times d_m}$使用$H$个低秩线性(Low-Rank Linear,LRL)模块得到$\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h \in \mathbb{R}^{T \times d_k}$矩阵,$h \in \{1,2,\cdots,H\}$. LRL模块由含有权重矩阵为$\mathbb{R}^{d_m \times d_e}$和$\mathbb{R}^{d_e \times d_k}$的两个线性层组成,进而对$\boldsymbol{Q}_h$矩阵使用含有权重$\boldsymbol{w}_h^S \in \mathbb{R}^{d_k \times 1}$的线性层对其映射后进行Softmax归一化得到上下文分数$\boldsymbol{s}_h \in \mathbb{R}^{T \times 1}$,该权重$\boldsymbol{w}_h^S$即为图2中的$\boldsymbol{l}_h$向量.

$$\boldsymbol{Q}_h = \boldsymbol{X}\boldsymbol{E}_h^Q\boldsymbol{F}_h^Q, \boldsymbol{K}_h = \boldsymbol{X}\boldsymbol{E}_h^K\boldsymbol{F}_h^K, \boldsymbol{V}_h = \varphi(\boldsymbol{X}\boldsymbol{E}_h^V\boldsymbol{F}_h^V) \tag{3}$$

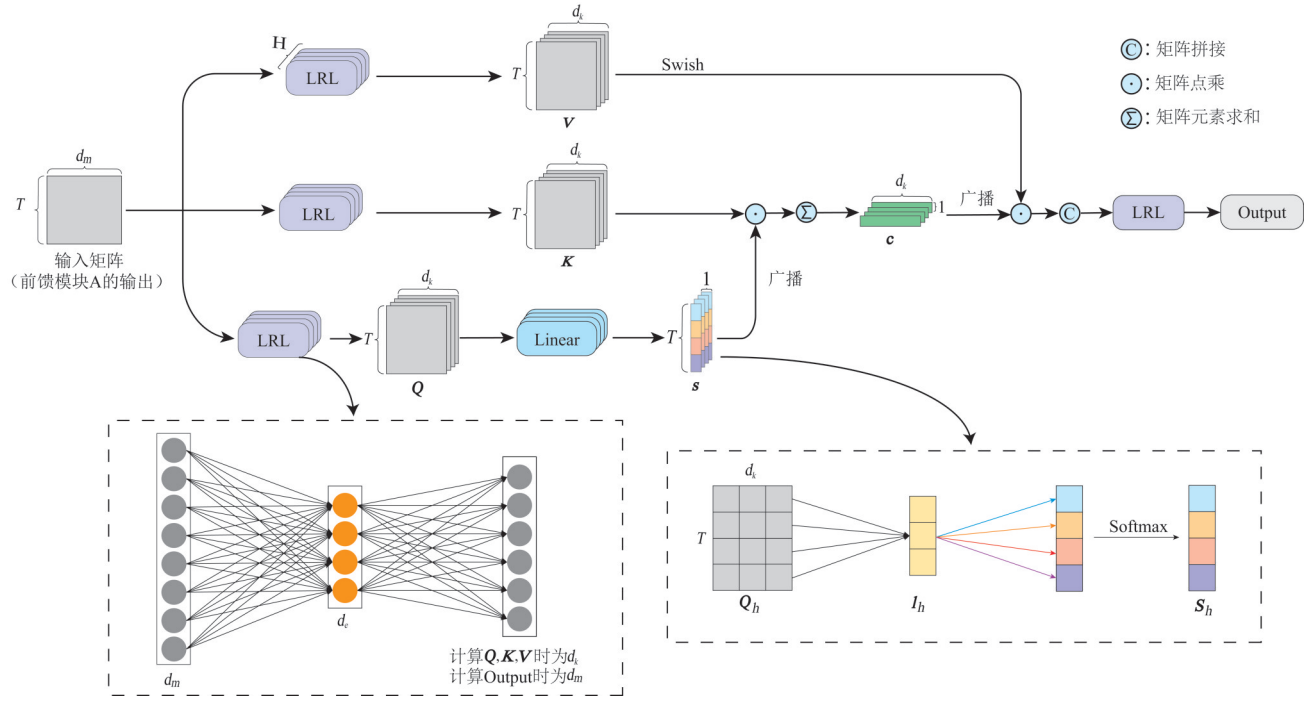


图2 多头逐元素线性注意力

$$\boldsymbol{s}_h = \mathrm{Softmax}\left(\frac{\boldsymbol{Q}_h\boldsymbol{w}_h^S}{\sqrt{d_k}}\right) \tag{4}$$

其中,$\boldsymbol{E}_h^Q \in \mathbb{R}^{d_m \times d_e}, \boldsymbol{F}_h^Q \in \mathbb{R}^{d_e \times d_k}$为$\boldsymbol{W}_h^Q$的权重分解矩阵,$\boldsymbol{E}_h^K \in \mathbb{R}^{d_m \times d_e}, \boldsymbol{F}_h^K \in \mathbb{R}^{d_e \times d_k}$为$\boldsymbol{W}_h^K$的权重分解矩阵,$\boldsymbol{E}_h^V \in \mathbb{R}^{d_m \times d_e}, \boldsymbol{F}_h^V \in \mathbb{R}^{d_e \times d_k}$为$\boldsymbol{W}_h^V$的权重分解矩阵,$T$为时间帧数,$d_m$为隐含层维度,$d_k = d_m/H$为单头注意力维度,$d_e = d_m/\beta$为分解矩阵维度,本文设置$\beta = 32$,$\sqrt{d_k}$为缩放因子,$\varphi$为Swish激活函数.

随后将$\boldsymbol{s}_h$中每一个标量$s_{h,t}$和$\boldsymbol{K}_h$中每一个向量$\boldsymbol{K}_{h,t}$相乘后求和得到上下文向量$\boldsymbol{c}_h \in \mathbb{R}^{1 \times d_k}$:

$$\boldsymbol{c}_h = \sum_{t=1}^{T}\left(s_{h,t}\boldsymbol{K}_{h,t}\right) \tag{5}$$

其中,$\boldsymbol{c}_h$的作用与多头注意力机制中注意力权重矩阵的作用相似. 进一步将每一个值向量与$\boldsymbol{c}_h$进行融合即可得第h个头的注意力. 由于每一个头都在不同的子空间中单独计算自注意力,所以每一个头都包含着不同空间中的特征信息,最终将所有头上所含信息整合后来得到MHELA的输出.

2.3 多尺度通道时频注意力

如图3所示,该模块首先对输入特征图在通道维度进行分组,对每一组特征图分别使用不同扩张率的时域扩张卷积来进行多尺度特征提取,进而对得到的多尺度特征图使用CT-FA来分别对特征图的通道、时间和频率维度添加不同的注意力权重. 对于输入特征图$\boldsymbol{Y} \in \mathbb{R}^{T \times F \times (C/4)}$,通道注意力权重$\boldsymbol{W}_c \in \mathbb{R}^{1 \times 1 \times (C/4)}$的计算方法为

$$\boldsymbol{W}_c = \sigma\left[\mathrm{Conv}_k\left(\mathrm{GAP}(\boldsymbol{Y})\right)\right] \tag{6}$$

其中,σ为Sigmoid激活函数,GAP代表全局平均池化,Conv_k代表一维卷积,卷积核大小k随通道数变化,将$\boldsymbol{W}_c$与$\boldsymbol{Y}$进行融合后得到通道注意力特征图$\boldsymbol{Y}_c \in \mathbb{R}^{T \times F \times (C/4)}$,进一步对特征图中每一个通道$\boldsymbol{Y}_c^i \in \mathbb{R}^{T \times F \times 1}, i \in \{1,2,\cdots,(C/4)\}$并行执行时频注意力:

$$\boldsymbol{W}_t^i = \sigma\left[\mathrm{Conv}_7\left(\mathrm{GAP}\left(\boldsymbol{Y}_c^i\right)\right)\right] \tag{7}$$

$$\boldsymbol{W}_f^i=\sigma\left[\mathrm{Conv}_3\left(\mathrm{GAP}\left(\boldsymbol{Y}_c^i\right)\right)\right] \tag{8}$$

其中，$\boldsymbol{W}_t^i\in\mathbb{R}^{T\times1\times1}$，$\boldsymbol{W}_f^i\in\mathbb{R}^{1\times F\times1}$代表第$i$个通道特征图的时间和频率注意力权重，将其与$\boldsymbol{Y}_c^i$融合后分别得到时间和频率注意力特征图，随后进行拼接通过二维卷积层得到第i个通道的时频注意力特征图，最后将每组经过CT-FA后的特征图进行拼接后进行通道重排得到多尺度融合的特征图.

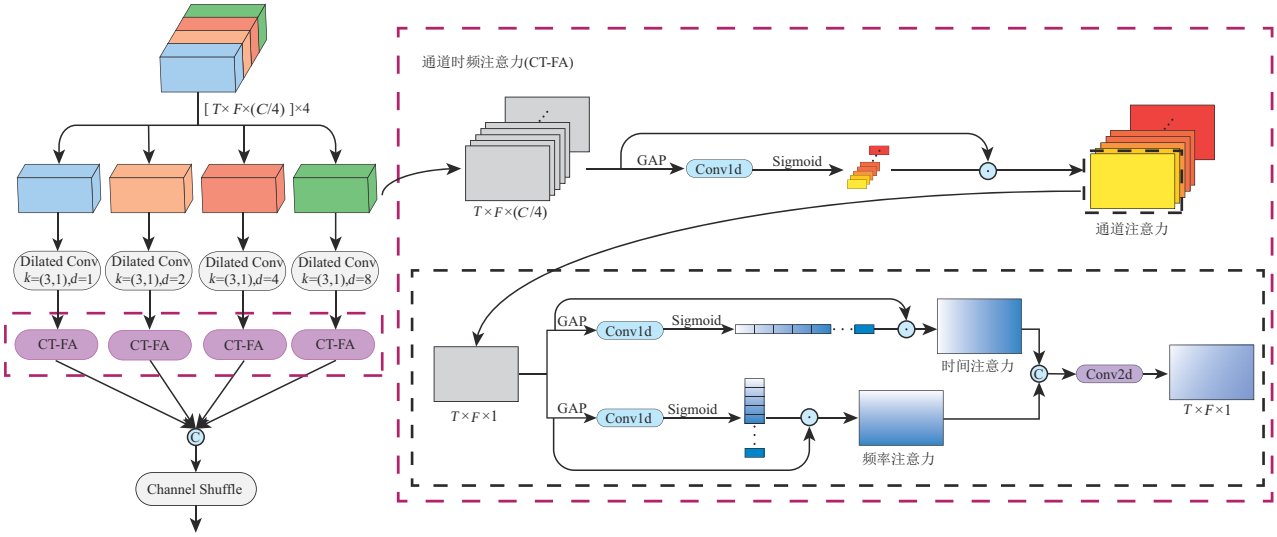


图3　多尺度通道时频注意力

3　实验结果与分析

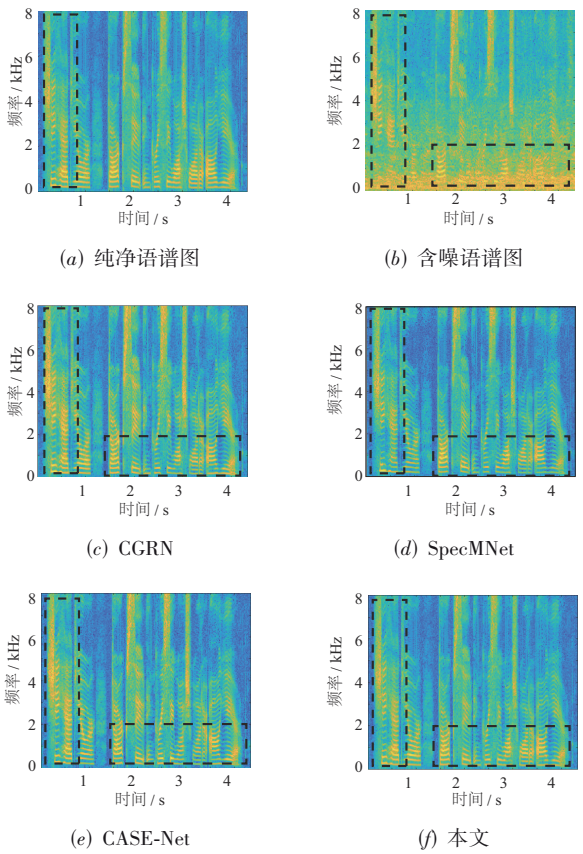
3.1　实验准备

本文根据车载实际环境自制了中文车载语音数据集，该自制数据集包含录音文本共200条. 其中指令文本64条，完整语句136条，基本包含了车载语音场景中需要用到的所有对话信息. 所有文本由195名20～60岁的不同年龄段的人进行录制而成，共包含31 543条语音，总时长大约25 h，采样率设置为16 000 Hz，在本实验中，将该数据集按照8∶1∶1的比例划分为训练集、验证集、测试集. 噪声数据集采用In-Vehicle Noise Dataset. 该数据集为乘客在数十个车辆上录制的车内噪声片段，数据集时长为5.08 h，基本涵盖了行驶时车外环境噪声与车内可能产生的所有噪声，随机从噪声数据集中截取与纯净音频时长相等的噪声片段以不同的信噪比混合为含噪语音.

本文实验在Ubantu16.04系统上运行，采用Python编程语言，语音的傅里叶变换点数设置为512，设定汉明窗口大小为25 ms，帧移为10 ms.

3.2　实验分析

在图4中我们可视化了不同语音增强网络的增强语谱图效果，进行实验的网络包括CGRN[1]、SpecMNet[2]、CASE-Net[3]. 为了方便公平比较，所有网络语音预处理方式与本文相同且维度与本文保持一致. 可以发现本文方法对语谱图上语音谐波恢复得最为明



(a) 纯净语谱图　(b) 含噪语谱图

(c) CGRN　(d) SpecMNet

(e) CASE-Net　(f) 本文

图4　不同网络下的增强语谱图比较

显,对噪声成分的抑制更显著.

同时为验证本文提出的联合训练模型的抗噪性能,将本文模型与主流的语音识别模型来进行字错误率、注意力复杂度及参数量上的对比,包括对比模型包括Conformer[4]、LAC[6]、和GLAC[7]. 我们对不同语音识别模型的代码自行实现编写后加入到程序中,表1为在不同信噪比下的语音识别模型效果对比. 可以看出,本文提出的联合训练模型在不同信噪比下均获得了明显的字错误率上的降低,表2为在不同序列长度下本文模型与其他模型在GPU和CPU上对单条语音推理延迟的对比,结果显示本文模型所需推理延迟均低于其他模型.

表1 不同信噪比下的语音识别模型字错误率对比(T为时间帧数)

对比方法	注意力复杂度	参数量/M	信噪比/dB					
			−5	0	5	10	20	平均
Conformer	$O(T^2)$	30.71	21.02	10.13	6.36	2.18	0.76	8.09
LAC	$O(T)$	22.83	21.58	10.64	6.98	2.52	0.95	8.53
GLAC	$O(T)$	19.83	20.96	10.82	6.25	2.51	0.61	8.23
本文	$O(T)$	9.73	**8.54**	**4.33**	**3.53**	**1.52**	**0.73**	**3.73**

表2 不同时间帧数下GPU/CPU推理时长对比 单位:ms

对比方法	时间帧数/个			
	256	512	768	1 024
Conformer	10.51/105.03	18.70/210.37	29.68/328.56	45.74/478.95
LAC	9.12/115.34	12.13/161.86	14.75/201.14	17.39/291.40
GLAC	7.63/90.09	10.32/124.73	13.65/147.96	15.39/254.39
本文方法	**3.91/41.34**	**5.82/55.68**	**8.28/93.48**	**10.98/159.44**

4 总结

在本文中我们分别构建了轻量化的语音增强模型和语音识别模型来解决车内语音交互问题. 实验结果表明,本文的联合训练模型在相较于其他主流语音识别模型在不同信噪比情况下的字错误率平均降低4.55%,同时拥有更低的参数量及更快的CPU/GPU推理速度. 未来的研究方向包括进一步优化网络各模块的结构,并在模型参数与识别效果之间取得更好的平衡.

参考文献

[1] 袁文浩, 胡少东, 时云龙, 等. 一种用于语音增强的卷积门控循环网络[J]. 电子学报, 2022, 50(12): 2945-2956.
YUAN W H, HU S D, SHI Y L, et al. A convolutional gated recurrent network for speech enhancement[J]. Acta Electronica Sinica, 2020, 48(7): 1276-1283. (in Chinese)

[2] FAN C H, ZHANG H M, YI J Y, et al. SpecMNet: Spectrum mend network for monaural speech enhancement[J]. Applied Acoustics, 2022, 194: 108792.

[3] XU X M, TU W P, YANG Y H. CASE-Net: Integrating local and non-local attention operations for speech enhancement[J]. Speech Communication, 2023, 148: 31-39.

[4] GULATI A, QIN J, C C CHIUet al. Conformer: Convolution-augmented Transformer for speech recognition[C]//Interspeech 2020. Singapore: ISCA, 2020: 5036-5040.

[5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 5999-6009.

[6] LI S Q, XU M L, ZHANG X L, et al. Efficient conformer-based speech recognition with linear attention[C]//Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. New York: IEEE, 2021: 448-453.

[7] 李宜亭, 屈丹, 杨绪魁, 等. 一种改进的线性注意力机制语音识别方法[J]. 信号处理, 2023, 39(3): 516-525.
LI Y T, QU D, YANG X K, et al. Speech recognition model based on improved linear attention mechanism[J]. Journal of Signal Processing, 2023, 39(3): 516-525. (in Chinese)

[8] FAN C H, DING M M, YI J Y, et al. Two-stage deep spectrum fusion for noise-robust end-to-end speech recognition [J]. Applied Acoustics, 2023, 212: 109547.

[9] ZHU Q S, ZHANG J, ZHANG Z Q, et al. A joint speech enhancement and self-supervised representation learning framework for noise-robust speech recognition[J]. ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 1927-1939.

作者简介

廉筱峪　男,2001年7月出生于辽宁省抚顺市. 现为大连工业大学硕士研究生. 主要研究方向为语音信号处理.
E-mail: 709242393@qq.com

夏　楠　男,1983年5月出生于辽宁省大连市. 2013年在大连理工大学获工学博士学位,其后在国家无线电监测中心从事无线电监测定位研究工作,高级工程师,现为大连工业大学信息科学与工程学院副教授. 主要研究方向为阵列信号处理、语音信号处理等.
E-mail: xia_nan0520@aliyun.com

戴高乐　男,2002年11月出生于浙江省宁波市,现为大连工业大学本科生. 主要研究方向为语音信号处理.
E-mail: 2050491891@qq.com

杨红琴　女,2000年12月出生于云南省昆明市,现为大连工业大学硕士研究生. 主要研究方向为语音信号处理、语音情感识别.
E-mail: 1909847594@qq.com