

基于超轻量通道注意力的端对端语音增强方法

洪依¹, 孙成立¹, 冷严²

(1. 南昌航空大学信息工程学院, 江西 南昌 330063;
2. 山东师范大学物理与电子科学学院, 山东 济南 250014)

摘要: 全卷积时域音频分离网络(Conv-TasNet)是近年提出的一种主流的端对端语音分离模型。Conv-TasNet利用膨胀卷积扩大感受野,使其在空间上可以融合更多语音特征,极大地提高了网络的语音分离性能,但同时忽略了信息在不同卷积通道间的重要性。基于此,提出一种基于超轻量通道注意力的端对端语音增强方法,该方法结合了Conv-TasNet和通道注意力,并在Conv-TasNet编解码器部分增加一组滤波器来提高网络语音特征提取能力,使卷积神经网络可以更有效地结合空间信息和通道信息来提高语音增强效果。实验验证了所提方法的模型容量在只增加了约0.02%的情况下,语音增强性能获得了有效提升。

关键词: 语音增强; 端到端语音分离网络; 通道注意力

中图分类号: TP183

文献标识码: A

doi: 10.11959/j.issn.2096-6652.202136

End-to-end speech enhancement based on ultra-lightweight channel attention

HONG Yi¹, SUN Chengli¹, LENG Yan²

1. School of Information Engineering, Nanchang Hangkong University, Nanchang 330063, China
2. School of Physics and Electronic, Shandong Normal University, Jinan 250014, China

Abstract: The full convolutional time-domain audio separation network (Conv-TasNet) is a state-of-the-art end-to-end speech separation model which was proposed recently. The Conv-TasNet used dilated convolution to expand the receptive field and fuse more speech features in space, which greatly improved the speech separation performance of the network, but at the same time ignored the importance of information across different convolution channels. An end-to-end speech enhancement method based on ultra-lightweight channel attention was proposed, which effectively combined Conv-TasNet and channel attention. At the same time, a group of filters was added to the Conv-TasNet codec to improve the speech feature extraction ability of the network. This method can make convolutional neural network combine spatial information and channel information more effectively to improve the speech enhancement effect. Experiment shows that the proposed model can effectively improve the performance of speech enhancement when the model capacity is only increased by about 0.02%.

Key words: speech enhancement, end-to-end speech separation network, channel attention

1 引言

语音增强是一种语音前端处理技术,其目的是

去除目标语音中的背景噪声,提高语音的可懂度^[1]。在日常生活中,语音作为一种比较常见的通信交流方式,极易受到各种干扰,从而降低通信的质量和

收稿日期: 2021-07-03; 修回日期: 2021-07-17

通信作者: 孙成立, sun_chengli@163.com

基金项目: 国家自然科学基金资助项目(No.61861033); 江西省自然科学基金重点项目(No.20202ACBL202007); 山东省自然科学基金资助项目(No.ZR2020MF020)

Foundation Items: The National Natural Science Foundation of China (No.61861033), The Key Project of Natural Science Foundation of Jiangxi Province (No.20202ACBL202007), The Natural Science Foundation of Shandong Province (No.ZR2020MF020)

效率。人耳可以相对容易地捕获嘈杂声音中的重要信息,并自动屏蔽各种干扰和噪声,但许多机器学习任务却很难做到这一点。近年来,随着深度学习的引入,基于深度学习的语音增强技术引起广泛的关注^[2-3]。在复杂环境以及数据量大的情况下,传统机器学习方法的信息处理能力十分有限,而基于深度学习的机器学习方法在非线性相关的复杂信号中具有更强的建模能力,更能充分挖掘大数据的优势^[4]。基于卷积神经网络(convolutional neural network, CNN)和循环神经网络(recurrent neural network, RNN)的深度学习模型已经在语音识别、语音分离、语音增强等多个领域取得突破性进展。

早期基于深度学习的语音增强方法大多作用于频域^[5-6],网络的输入是短时傅里叶变换(short-time Fourier transform, STFT)后的幅度谱或功率谱特征,在利用逆傅里叶变换(inverse short-time Fourier transform, ISTFT)重建语音信号时,直接采用混合语音的相位,忽略了相位影响。这是因为早期方法认为相位中只包含很少的有用信息,对增强效果并不重要^[7]。随着近年研究的深入,一些学者认为相位对语音增强也具有重要性。Paliwal K 等人^[8]指出在已知纯净语音相位的情况下,语音感知质量会显著提升。为了充分考虑相位信息,在频域上对相位信息进行补偿和采用复数神经网络等改进方法随之被提出。部分研究学者从基于频域的方法转向基于时域的端对端语音增强方法^[9-11]。相对于基于频域的方法,基于时域的端对端语音增强方法用一维卷积网络代替傅里叶变换来提取特征,卷积网络的参数可以通过机器学习自动更新,一方面提高了特征抽取的有效性,另一方面避免了频域傅里叶变换带来的相位损失和相位失配等问题。

全卷积时域音频分离网络(full convolutional time-domain audio separation network, Conv-TasNet)^[12]是一种高效的端对端语音分离网络,其采用多层的全卷积神经网络,由编码器、分离器和解码器构成。其中编码器和解码器的作用分别类似于 STFT 和 ISTFT,编码器将语音线性地转化成相应的表示形式,即提取语音特征;解码器则用于恢复波形,生成增强语音。分离器则主要通过时间卷积网络(temporal convolutional network, TCN)估计不同语音的掩模(mask),从而达到分离语音的目的。

与传统的频域分离方法相比,Conv-TasNet 性能有了大幅度提升,成为一种主流的语音增强方法。一些 Conv-TasNet 的改进模型相继被提出,包括加深编解码器卷积网络层数的模型^[13]、添加多尺度动态加权门控扩展卷积的 FurcaPa^[14]模型和 FurcaPy^[15]模型,极大地推动了时域语音分离网络的研究。Conv-TasNet 在采用时间卷积网络提高对空间信息的利用的同时,忽略了信息在不同通道的重要性,导致部分弱能量的语音被去除,并残留了少量噪声信号。同时,随着深度神经网络的发展,网络模型的部署会受到设备内存的限制,实现高效的内存管理成为深度学习发展的关键问题^[16]。

本文为了提高 Conv-TasNet 的语音增强性能,考虑了卷积通道中不同语音的重要性,并保持模型轻量化的优势,提出一种基于超轻量通道注意力的端对端语音增强方法。该方法在 Conv-TasNet 中引入基于超轻量通道的注意力机制,使模型自动学习更显著的通道信息,同时削弱对不重要的通道信息的关注,避免弱能量语音信号的损失,使卷积网络有效地利用语音的空间信息和通道信息,采用轻量型的分组卷积结构代替标准的卷积网络来加深编解码器结构,提高网络对语音特征的处理能力,并在 LibriMix^[17]语音增强数据集上验证了所提方法的有效性。

本文第2节介绍了 Conv-TasNet 的基本结构,第3节介绍了所提方法中的语音增强网络,第4节对实验结果进行了分析,第5节总结全文。

2 相关工作

2.1 全卷积时域音频分离网络

Conv-TasNet^[12]是一种基于时域的端对端的语音分离模型,其采用编解码结构,基本框架如图1所示,包含编码器、分离器、解码器。

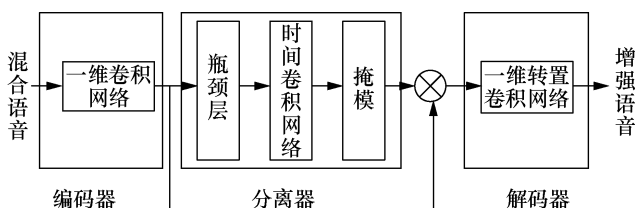


图1 Conv-TasNet 的基本框架

编码器由单个一维卷积网络和激活函数 PReLU^[18]组成,其将长度为 L 的混合语音 $x \in R^{1 \times L}$

波形转换为 N 维特征 $W \in R^{1 \times N}$ ，即时域深度表征特征。转换过程如式 (1) 所示：

$$W = G(xU) \quad (1)$$

其中， $U \in R^{L \times N}$ 为编码器的基本函数， $G()$ 表示激活函数。

解码器与编码器结构类似但作用相反，其由一个一维转置卷积网络组成，作用是将语音信号重构为对应的语音波形，如式 (2) 所示：

$$y = WV \quad (2)$$

其中， $y \in R^{1 \times L}$ 表示重构后的语音信号， $V \in R^{N \times L}$ 表示解码器的基本函数。

分离器生成权重掩模，即屏蔽噪声信号提取干净的语音信号，实现语音分离的功能。其核心部分为 TCN^[19]，TCN 以 CNN 为基础，结合了膨胀卷积^[10]和残差模块^[20]，大大提升了 CNN 处理序列问题的能力，同时避免了 RNN 不可并行处理等问题。采用膨胀卷积可以在卷积层数量有限的基础上最大限度地增加感受野，避免模型过度复杂化，并且不会像池化操作那样损失部分特征信息。随着隐藏层的层数和膨胀因果卷积的扩张率 d 的增加，输出层节点可以看到的输入层的信息也增加了。膨胀因果卷积基本结构如图 2 所示，输出层节点可以在只有 4 层卷积网络的基础上看到之前 15 个输入层的节点信息，很大程度地增加了感受野；而残差模块用于处理网络退化问题并避免了梯度消失和爆炸问题。除此之外，Conv-TasNet 在 TCN 中使用深度可分离卷积网络^[21]代替标准的卷积网络来进一步减少参数数量和运算成本，使模型保持较小的尺寸。同时 Conv-TasNet 采用 CNN 可以很好地避免时域音频分离网络 (time-domain audio separation network, TasNet)^[22]中使用 RNN 的不可并行处理问题以及对混合语音的确切起始点敏感问题。

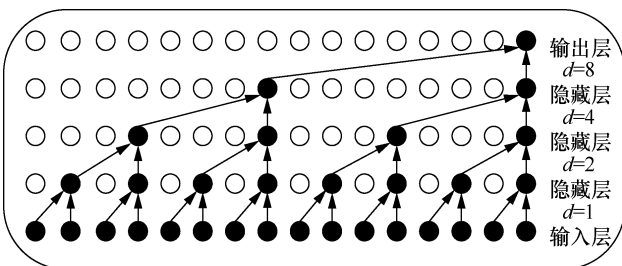


图2 膨胀因果卷积基本结构

2.2 高效通道注意力

高效通道注意力网络 (efficient channel attention network, ECA-Net)^[23] 是一种基于 SENet (squeeze-and-excitation network)^[24] 的改进的通道注意力模型，其在增加极少参数的基础上获得了性能提升，并避免了 SENet^[24] 的降维操作的副作用。SENet 是 2017 年提出的一种新颖的图像识别结构，它提出了一种新的网络改进方向——通道注意力，即主要对特征通道间的相关性进行建模，从而提升网络的准确性。SENet 由压缩 (squeeze)、激励 (excitation)、特征重标定构成，其基本结构如图 3 所示，输入矩阵 X 通过 SENet 转化为 $\bar{X}$ 。

squeeze 部分对应图 3 中的 GAP 操作，采用全局平均池化将特征谱在空间维度上压缩为具有全局信息的 $1 \times 1 \times C$ 的实数数列 (C 为特征通道数)，很好地解决了网络各层次间信息利用率低的问题，处理过程可用式 (3) 表示：

$$Y = F_{sq}(U_C) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W U_C(i, j) \quad (3)$$

其中， H 、 W 分别表示空间维度的高和宽， U_C 和 Y 则分别表示进行 squeeze 处理前后的矩阵。

excitation 部分对应图 3 中的 $F_{ex}(\cdot, W)$ ，主要目的是获取 squeeze 部分的信息，即通道间的依赖关系，并生成每个通道对应的权重值，表示获取信息的重要程度；同时使用两个全连接层将特征通道数 C 缩小为原来的 $1/r$ ，减少计算量，最后通过激活函数 σ 得到权重值。特征重标定部分则通过将 excitation 部分得到的权重值与原始特征逐通道相乘，实现对特征的重标定。

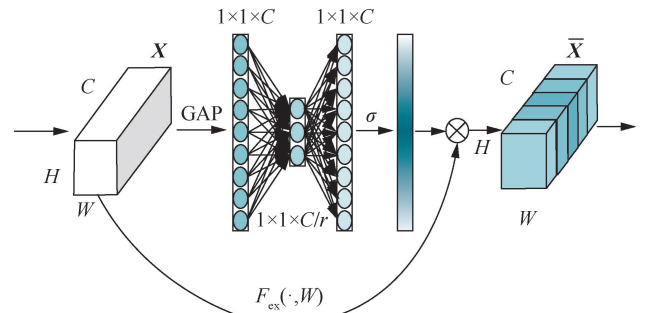


图3 SENet 的基本结构

图 4 描述了本文使用的 ECA-Net 的基本结构，不同颜色表示不同的权重，通过与图 3 比较可以看出，ECA-Net 与 SENet 的主要不同在于舍弃了中间的降维操作，而选择利用各个通道以及其 K 个近邻来获取局部跨通道信息。

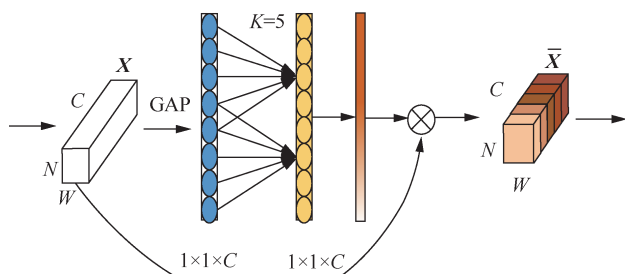


图4 ECA-Net的基本结构

3 基于超轻量通道注意力的端对端语音增强网络

CNN 主要是通过卷积运算来融合来自局部感受野内的空间信息和通道信息,进而构建特征信息。近年来提出的嵌入多尺度信息的 Inception^[25]、考虑空间上下文信息的 Inside-Outside^[26]等基于 CNN 的改进方法均选择从卷积网络的空间维度对网络性能进行提升,Conv-TasNet 在分离器中采用的 TCN 同样对 CNN 的空间层面进行了改进,而没有考虑语音通道间的相关性,导致了语音中部分弱信号的损失。本文提出在通道上对网络做进一步的改进,构建了一个基于超轻量通道注意力的端对端语音增强网络。考虑各通道间相关性,通过引入通道注意力模块完善 Conv-TasNet 对各种信息的有效利用,达到更好的语音增强性能,本文提出的基于超轻量通道注意力的端对端语音增强网络的结构如图 5 所示,其中 ECA-TCN 为结合了本文提出的 ECA-Net 和 TCN 的改进模块,其网络结构如图 6 所示。TCN 由多个一维卷积块堆叠而成,图 6 只标出了其中一个一维卷积块的结构,即图 6 右半部分,其他一维卷积块均做相同改进。

TCN 采用一系列深度可分离网络^[21]的一维卷积块对特征信息进行处理,先对输入的特征进行逐层卷积,即每个通道使用一个滤波器进行卷积运算,之后再通过逐点卷积网络进行拼接。在 Conv-TasNet 分离器中,瓶颈层决定输入通道的数量和后续卷积块的剩余路径,为了使下一部分的 TCN 更高效地利用有用特征,本文提出在瓶颈层后加上 ECA-Net 对编码器提取的语音特征的重要程度进行区分;另外在深度可分离网络部分结合逐层卷积和 ECA-Net,进一步加强对输入信息的重要程度的关注,从而达到提升网络性能的目的。

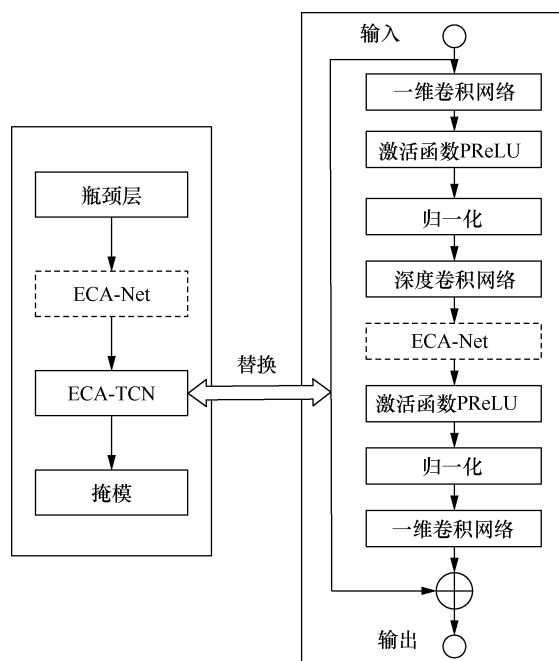


图6 ECA-TCN 的网络结构

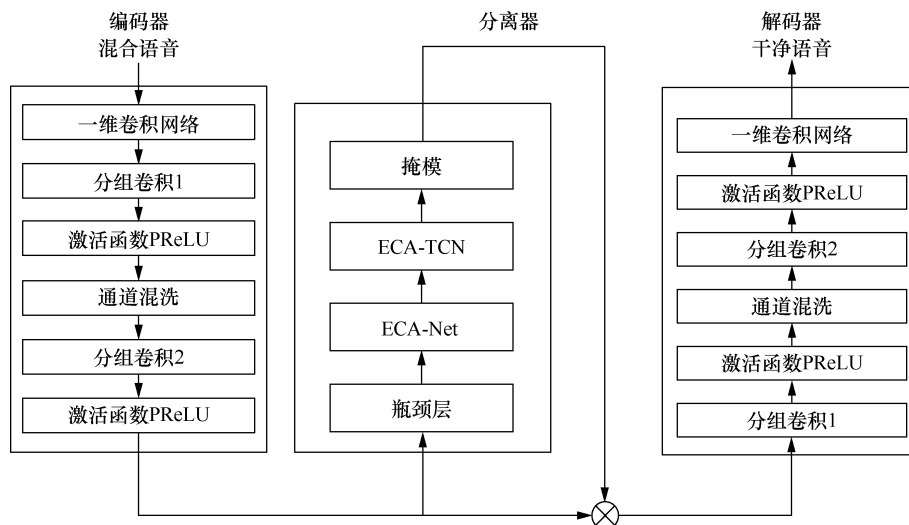


图5 基于超轻量通道注意力的端对端语音增强网络的结构

Wang Q L 等人^[23]比较了 SENet 与其 3 种变体的性能。实验证明, 采用跨通道的信息交互可以取得最好的性能提升, 且不需要对网络进行降维操作。同时认为采取局部的跨通道交互可以在保持极少量参数的情况下提升网络的性能。通道特征 y_i 的权重 ω_i 的计算如式 (4) 所示:

$$\omega_i = \sigma\left(\sum_{j=1}^k \alpha_i^j y_i^j\right), y_i^j \in \Omega_i^k \quad (4)$$

其中, σ 表示 Sigmoid 函数, y_i^j 为第 i 个通道的第 j 个近邻通道的特征, α_i^j 为学习参数, Ω_i^k 为 y_i 的 k 个近邻通道的集合。

让所有通道共享相同的学习参数 α^j 以进一步提高模型性能。具体表现如式 (5) 所示:

$$\omega_i = \sigma\left(\sum_{j=1}^k \alpha^j y_i^j\right), y_i^j \in \Omega_i^k \quad (5)$$

为了更加简洁直接地实现上述过程, Wang Q L 等人^[23]提出了高效通道注意力 (efficient channel attention, ECA) 模块, 通过核为 k 的一维卷积网络实现, 则通道特征 y 的权重 ω 可由式 (6) 计算得出。其中 1D 表示一维卷积, 卷积核大小为 k 。

$$\omega = \sigma(1D_k(y)) \quad (6)$$

由于 k 表示跨通道交互的覆盖范围, 当通道数或网络结构改变时, k 的大小会极大地影响网络性能, 因此为了使 ECA 模块被更广泛地应用于各种网络, 且避免因手动调节 k 的大小而造成的计算消耗, ECA-Net 采用了一种自适应选择 k 的大小的方法。其定义如式 (7) 所示:

$$k = \psi(C) = \left\lfloor \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right\rfloor_{\text{odd}} \quad (7)$$

其中, C 为通道数, $|x|_{\text{odd}}$ 表示取离 x 最近的奇数, γ 和 b 为两个参数, 在原论文中分别设置为 2 和 1。由上式可知 k 的大小与通道数 C 成正比。

Deep-Encoder-Decoder-Conv-TasNet^[13] 提出的在编解码部分添加一组小参数量滤波器的改进方法对 Conv-TasNet 语音分离性能有较大提升, 证明了编解码器部分对网络性能提升的潜力。但引入过多的 CNN 会带来过多的参数量, 提高网络复杂度。本文采用 ShuffleNet^[27] 结构, 用分组卷积^[28]代替普通的一维卷积网络对网络的编解码器进行改进, 提升网络对语音特征的表现能力, 同时采用通道混洗模块解决分组卷积各组信息不流通问题, 分组卷积

与通道混洗示意图如图 7 所示。首先进行分组卷积, 将输入特征谱分为 n 组, 图 7 中不同颜色代表不同的组, 再通过通道混洗将各组信息打散分配到各个组中, 使得第二个分组卷积可以得到每个组的输入信息, 提高了模型对信息的利用率。

4 实验结果与分析

4.1 实验环境

本文将混合了纯净语音 LibriSpeech^[29]库和含有真实环境噪声的 WHAM! (WSJ0 Hipster Ambient Mixtures)^[30]库的混合语音库 LibriMix 增强语音数据集作为实验数据。LibriMix 是^[17]一个开源语音数据库, 包含带噪和纯净多说话人混合语音。本文采用了其中 100 h 的语音样本, 样本采样频率为 8 kHz。实验采用的 LibriSpeech 数据集见表 1。

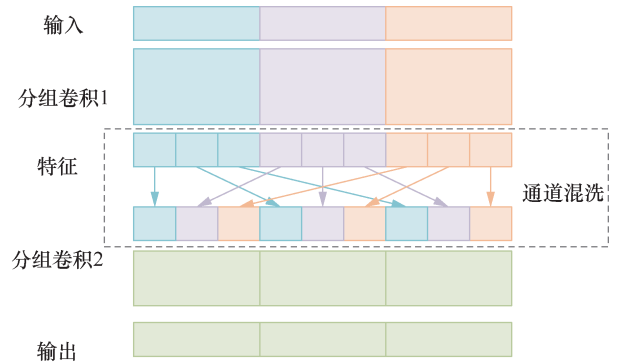


图7 分组卷积与通道混洗示意图

表1 实验采用的 LibriSpeech 数据集

子集	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/ 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

本文使用端对端语音分离较通用的信号失真比 (signal to distortion ratio, SDR) 以及其变体尺度不变信号失真比 (scale-invariant SDR, SI-SDR)^[31]来评测模型性能, 其中 SI-SDR 定义如下:

$$\text{SI-SDR} = 10 \lg \left(\frac{\|S_{\text{tar}}\|^2}{\|e_{\text{noi}}\|^2} \right) \quad (8)$$

$$S_{\text{tar}} = \frac{\langle \hat{s}, s \rangle s}{\|s\|^2} \quad (9)$$

$$e_{\text{noi}} = \hat{s} - S_{\text{tar}} \quad (10)$$

其中, $\hat{s}$ 和 s 分别表示待评测信号和原始信号。为

了全面地表现网络性能,本文同时采用主观语音质量评估 PESQ^[32]和短时客观可懂度 STOI^[33]两种常见的语音评价指标对网络性能进行评测,其中 PESQ 的范围是 $-0.5 \sim 4.5$, STOI 的范围是 $0 \sim 1$,两者的数值越大,系统性能越佳。

本文采用 5 s 的混合语音,在 Asteroid 上迭代 90 次,批处理大小(batch size)为 2,初始学习率为 5×10^{-4} ,其他参数设置均与原始 Conv-TasNet 相同,其中当验证集的准确率在连续 3 个周期未上升时,学习率减半,优化器为 Adam。

4.2 实验结果和分析

4.2.1 ECA-Net 与 SENet 性能比较

ECA-Net 与 SENet 均为当前主流的通道注意力模型,ECA-Net 在 SENet 的基础上舍弃了降维操作,选择局部跨通道信息交互的方法。本文在 Conv-TasNet 的相同位置依次插入 ECA-Net 与 SENet 注意力结构,并分别使用 SDR、SI-SDR、PESQ

以及 STOI 对网络性能进行评估,对比结果见表 2。

从表 2 可以看出,相对于插入 SENet 结构,在插入超轻量的 ECA-Net 结构之后,Conv-TasNet 各评价指标相较于 Conv-TasNet 均有提高,证明本文提出的改进方法可以更高效地学习 CNN 的通道交互信息,而用于增加编解码器复杂度的 Deep-ECA-Net 可以进一步提高语音增强性能。在模型参数上,插入 ECA-Net 结构后的网络参数增量为 0.001 M,几乎可以忽略,而插入 SENet 结构后,网络参数由原始的 4.984 M 增加到 5.773 M,增加了约 15.8%。

在性能比较上,本文随机选取了一个样本并绘制了其波形图和频谱图,具体如图 8、图 9 所示。

由图 8 可以看出,Conv-TasNet 以及 Deep-ECA-Net 均有很强去噪效果,但对比图 8 中红色框标记部位可以看出,原始的 Conv-TasNet 由于没有考虑语音通道间的相关性,在去除环境噪声的同时也去除了一些能量较弱的信号,而本文提出的

表 2 在 Conv-TasNet 相同位置添加不同通道注意力网络的性能

方法	SDR/dB	SI-SDR/dB	PESQ	STOI	网络参数量/M
Conv-TasNet	14.668	13.895	2.835	0.916	4.984
Conv-TasNet+SENet	14.700	13.959	2.826	0.916	5.773
Conv-TasNet+ECA-Net	14.827	14.094	2.870	0.918	4.985
Conv-TasNet+Deep-ECA-Net	14.888	14.095	2.903	0.920	6.173

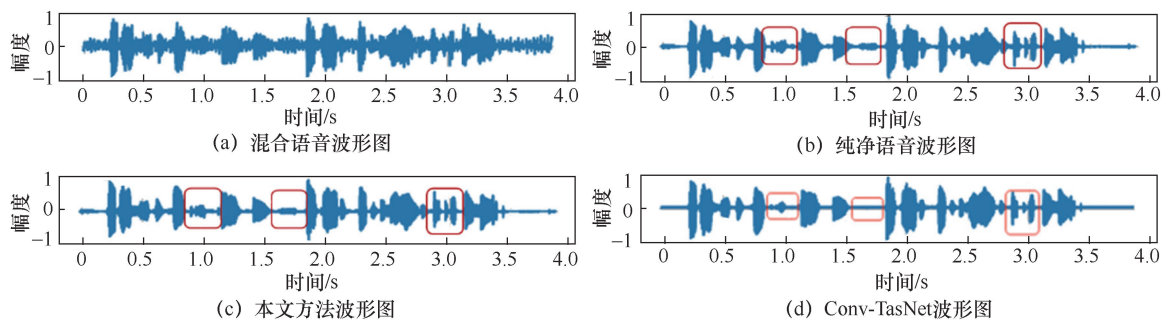


图 8 样本波形图

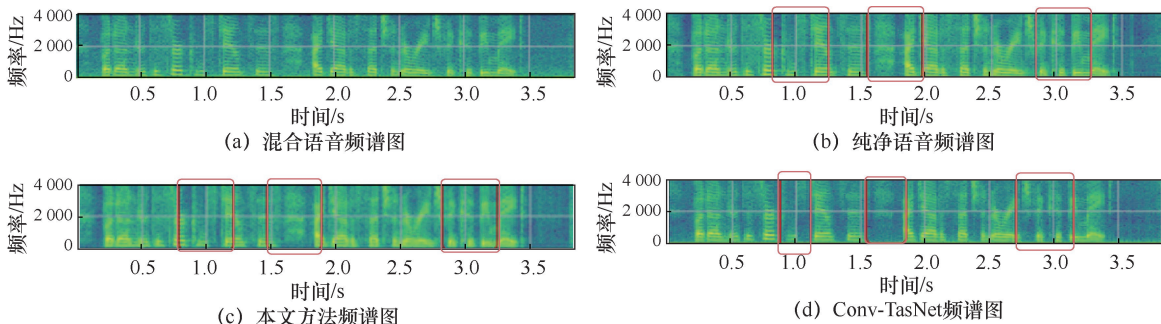


图 9 样本频谱图

Deep-ECA-Net 由于加入了通道注意力模块, 对语音信号以及噪声信号进行了权重均衡, 在保证模型去噪性能的同时不会导致语音信息丢失, 使得增强语音更加接近纯净语音, 模型性能得到提升。

4.2.2 通道注意力位置对增强性能的影响

Conv-TasNet 由编解码器和分离器组成, ECA-Net 结构被插入 Conv-TasNet 不同位置时的性能必然不同。为了使 ECA-Net 能最大限度地提升 Conv-TasNet 的语音增强性能, 本文测试了将超轻量的 ECA-Net 结构插入 Conv-TasNet 的 3 种位置时对增强性能的影响: ①插入编解码器, ②插入分离器, ③同时插入编解码器和分离器, 并分别命名为 ECA-Net1、ECA-Net2、ECA-Net3, 结果见表 3。

表 3 ECA-Net 处于 Conv-TasNet 不同位置时的网络性能

方法	SDR/dB	SI-SDR/dB	PESQ	STOI
Conv-TasNet	14.668	13.895	2.835	0.916
ECA-Net1	14.667	13.898	2.836	0.916
ECA-Net2	14.827	14.094	2.870	0.918
ECA-Net3	14.812	14.030	2.866	0.917

由表 3 可以看出, 在 Conv-TasNet 的分离器部分插入 ECA-Net 的 ECA-Net2、ECA-Net3 均可以获得较好的增强结果, 而 ECA-Net1 的 SDR 指标甚至低于原始 Conv-TasNet, 其他几项指标只获得微小的提升。TCN 采用一维卷积块组成的多层深度可分离网络对提取的语音特征进行处理, 随着网络层数的加深, 感受野变大, 网络获取的底层信息更丰富, 同时有效地结合通道注意力使得网络可以利用不同语音在不同通道的重要性, 从而获得更有效的语音特征信息, 同时削弱 Conv-TasNet 对噪声信号的关注, 提高网络语音增强的性能。因此本文在 Conv-TasNet 的分离器插入 ECA-Net, 使得语音增强模型可以最高效地利用该模块获取有效的特征信息, 提高网络性能。

4.2.3 实验结果

本文提出了一种基于超轻量通道注意力的端对端语音增强方法, 在添加通道注意力的基础上扩张编解码结构, 提高语音特征表达能力, 弥补原始 Conv-TasNet 语音信息利用不充分的不足, 提升模型语音增强性能。

5 结束语

本文结合卷积网络针对空间信息和通道交互信息的利用构建一个基于超轻量通道注意力的端

对端语音增强网络, 并在 LibriMix 增强语音数据集上对网络性能进行测试。实验证明, 该网络不仅避免了常规注意力模型 SENet 中降维操作的副作用, 而且可以使网络在增加非常少量参数的基础上更高效地捕获跨通道的交互信息, 提升网络语音增强的性能。但是由于端对端的结构是逐点计算的, 网络的 batch size 不可设置得过大, 对硬件设备要求高, 且运用了较多的分组卷积, 使得网络训练速度较慢, 后续将研究通过改进网络结构来提升网络训练速度, 提高模型的实时性。

参考文献:

- [1] LOIZOU P. Speech enhancement: theory and practice[M]. Boca Raton: CRC Press, 2007.
- [2] ZHOU N, DU J, TU Y H, et al. A speech enhancement neural network architecture with SNR-progressive multi-target learning for robust speech recognition[C]//Proceedings of the 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Piscataway: IEEE Press, 2019: 873-877.
- [3] WU B, YU M, CHEN L W, et al. Distortionless multi-channel target speech enhancement for overlapped speech recognition[J]. arXiv preprints, 2020, arXiv: 2007.01566.
- [4] 张钹. 人工智能进入后深度学习时代[J]. 智能科学与技术学报, 2019, 1(1): 4-6.
ZHANG B. Artificial intelligence is entering the post deep-learning era[J]. Chinese Journal of Intelligent Science and Technology, 2019, 1(1): 4-6.
- [5] GRZYWALSKI T, DRGAS S. Using recurrences in time and frequency within U-Net architecture for speech enhancement[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 6970-6974.
- [6] TAN K, WANG D L. A convolutional recurrent neural network for real-time speech enhancement[C]//Proceedings of the INTERSPEECH 2018. [S.l.:s.n.], 2018: 3229-3233.
- [7] WANG D, LIM J. The unimportance of phase in speech enhancement[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1982, 30(4): 679-681.
- [8] PALIWAL K, WÓJCICKI K, SHANNON B. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(4): 465-494.
- [9] LU X G, TSAO Y, MATSUDA S, et al. Speech enhancement based on deep denoising autoencoder[C]//Proceedings of the INTERSPEECH 2018. [S.l.:s.n.], 2018: 436-440.
- [10] DEN OORD A V, DIELEMAN S, ZEN H, et al. WaveNet: a generative model for raw audio[J]. arXiv preprint, 2016, arXiv:1609.03499.
- [11] PANDEY A, WANG D L. A new framework for CNN-based speech enhancement in the time domain[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(7): 1179-1188.
- [12] LUO Y, MESGARANI N. Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2019, 27(8): 1256-1266.

- [13] BAI S J, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv preprint, 2018, arXiv:1803.01271.
- [14] SHI Z Q, LIN H B, LIU L, et al. Deep attention gated dilated temporal convolutional networks with intra-parallel convolutional modules for end-to-end monaural speech separation[C]//Proceedings of the INTERSPEECH 2019. [S.l.:s.n.], 2019: 3183-3187.
- [15] SHI Z Q, LIN H B, LIU L, et al. End-to-end monaural speech separation with multi-scale dynamic weighted gated dilated convolutional pyramid network[C]//Proceedings of the INTERSPEECH 2019. [S.l.:s.n.], 2019: 4614-4618.
- [16] 马玮良, 彭轩, 熊倩, 等. 深度学习中的内存管理问题研究综述[J]. 大数据, 2020, 6(4): 56-68.
- MA W L, PENG X, XIONG Q, et al. Memory management in deep learning: a survey[J]. Big Data Research, 2020, 6(4): 56-68.
- [17] COSENTINO J, PARIENTE M, CORNELL S, et al. LibriMix: an open-source dataset for generalizable speech separation[J]. arXiv preprint, 2020, arXiv:2005.11262.
- [18] HE K M, ZHANG X Y, REN S Q, et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Press, 2015: 1026-1034.
- [19] LEA C, VIDAL R, REITER A, et al. Temporal convolutional networks: a unified approach to action segmentation[M]//Lecture notes in computer science. Cham: Springer International Publishing, 2016.
- [20] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 770-778.
- [21] HOWARD A G, ZHU M L, CHEN B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint, 2017, arXiv:1704.04861.
- [22] LUO Y, MESGARANI N. TaSNet: time-domain audio separation network for real-time, single-channel speech separation[C]//Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2018: 696-700.
- [23] WANG Q L, WU B G, ZHU P F, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2020: 11531-11539.
- [24] HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(8): 2011-2023.
- [25] SZEGEDY C, LIU W JIA Y Q, et al. Going deeper with convolutions[C]//Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014.
- [26] BELL S, ZITNICK C L, BALA K, et al. Inside-Outside net: detecting objects in context with skip pooling and recurrent neural networks[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2016: 2874-2883.
- [27] ZHANG X Y, ZHOU X Y, LIN M X, et al. ShuffleNet: an extremely efficient convolutional neural network for mobile devices[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 6848-6856.
- [28] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.
- [29] PANAYOTOV V, CHEN G G, POVEY D, et al. LibriSpeech: an ASR corpus based on public domain audio books[C]//Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2015: 5206-5210.
- [30] WICHERN G, ANTOGNINI J, FLYNN M, et al. WHAM!: extending speech separation to noisy environments[C]//Proceedings of the INTERSPEECH 2019. [S.l.:s.n.], 2019.
- [31] ROUX J L, WISDOM S, ERDOGAN H, et al. SDR – half-baked or well done?[C]//Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway: IEEE Press, 2019: 626-630.
- [32] RIX ANTONY W, BEERENDS JOHN G, HOLLIER MICHAEL P, et al. Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs[C]//Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Piscataway: IEEE Press, 2002: 749-752.
- [33] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[J]. IEEE Transactions on Audio Speech & Language Processing, 2011, 19(7): 2125-2136.

[作者简介]



洪依 (1997—), 女, 南昌航空大学信息工程学院硕士生, 主要研究方向为信号处理、语音增强、回声消除等。



孙成立 (1975—), 男, 博士后, 南昌航空大学信息工程学院教授, 主要研究方向为人工智能、语音信号处理、语音识别、语音增强等。



冷严 (1982—), 女, 博士, 山东师范大学物理与电子科学学院副教授, 主要研究方向为音频信息处理、音频分类与检测。