ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 38,No. 2,Mar. 2023,pp. 262－273
DOI:10. 16337/j. 1004-9037. 2023. 02. 003

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于临界频带的交互性双支路单通道语音增强模型

叶中付[1,2]，赵紫微[1,2]，于润祥[1,2]

(1. 中国科学技术大学电子工程与信息科学系,合肥 230022;2. 语音及语言信息处理国家工程研究中心,合肥 230022)

摘　要：针对目前主流的双支路单通道语音增强方法只关注全频带信息而忽略子频带信息这一问题，设计了一种基于人耳临界频带的交互性双支路模型。主要做法为,在复数谱支路上实施模拟人耳临界频带的划分方法对信号进行分频带处理,提取子带信息;在幅度补偿支路上直接对信号的全频带进行处理,提取全频带信息。复数谱支路负责初步恢复干净语音的幅度和相位,同时,该支路上学到的子带中间特征会被特定的模块传递给幅度补偿支路进行补偿;幅度补偿支路上的输出会对复数谱支路上输出的幅度做进一步的补偿,达到恢复干净语音频谱的目的。实验结果表明,提出的模型在恢复语音质量和可懂度方面优于其他先进的单通道语音增强模型。

关键词：临界频带;交互性;子带;双支路;单通道语音增强

中图分类号：TN912.35;TP183　　**文献标志码**:A

Interactive Dual-Branch Monaural Speech Enhancement Model Based on Critical Frequency Band

YE Zhongfu[1,2], ZHAO Ziwei[1,2], YU Runxiang[1,2]

(1. Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei 230022, China;2. National Engineering Research Center of Speech and Language Information Processing, Hefei 230022, China)

Abstract: Aiming at the problem that the current mainstream dual-branch single-channel speech enhancement methods only pay attention to the full frequency band information while ignoring the subband information, an interactive dual-branch model based on the critical frequency band of the human ear is proposed. The main method is to implement the division method of simulating the critical frequency band of the human ear on the complex spectrum branch to process the signal in frequency division and extract subband information. The whole frequency band of the signal is directly processed on the amplitude compensation branch, and the information of the whole frequency band is extracted. The complex spectrum branch is responsible for initially recovering the amplitude and phase of the clean speech signal. At the same time, the subband intermediate features learned by the branch are transferred to the amplitude compensation branch by specific modules for compensation. The output on the amplitude compensation branch will further compensate the amplitude of the output on the complex spectrum branch to achieve the

基金项目:国家自然科学基金 (61671418)。
收稿日期:2022-10-24;**修订日期**:2023-02-27

purpose of recovering the clean speech spectrum. Experimental results show that the proposed model is superior to other advanced models in restoring speech quality and intelligibility.

Key words: critical frequency band; interactive; subband; dual-branch; monaural speech enhancement

引　　言

单通道语音增强是指从单个麦克风收集到的带噪语音中恢复出干净语音的技术。近年来,深度学习方法广泛应用于语音增强方面,其语音增强性能优于维纳滤波方法[1]、基于子空间的方法[2]和基于统计模型的方法[3]等传统方法。

基于神经网络的语音增强任务可在时域和时频域实现。在时域中,直接学习原始带噪语音波形到干净语音波形的映射[4],实现端到端的语音增强;在时频域中,典型的语音增强方法通常旨在估计掩蔽函数或直接预测干净语音的频谱幅度[5-6],其中在重建时域波形时带噪语音相位保持不变。当文献[7]表明准确的相位谱估计可以显著提高语音质量后,大量基于复数域的方法被提出用于解决相位估计问题。文献[8]提出了将复数理想比率掩蔽(Complex ideal ratio mask,cIRM)用于恢复信号的复时频谱;文献[9]提出了一种基于映射的卷积循环网络(Convolutional recurrent network,CRN),利用复数谱映射网络直接预测干净复数谱的实部和虚部分量;在文献[9]的基础上,文献[10]提出了一种可以模拟复数运算的深度复卷积循环网络(Deep complex convolutional recurrent network, DCCRN),进一步提高了语音增强的效果。

最近,目标解耦方法被提出并广泛应用起来[11-13],主要做法为将原问题解耦为多个子问题,从互补的角度实现分阶段或者分支路的频谱恢复。有研究表明[14],简单地将训练目标更改为cIRM并不能达到同时恢复幅度和相位的预期效果,而双支路或者多阶段的目标解耦模型可以有效缓解幅值与相位之间的补偿问题[15],取得比单支路或者单阶段更好的增强效果。文献[16]提出了一种幅度谱支路和复数谱支路并行的扫视和凝视网络(Glance and gaze network, GaGNet)用于单通道语音增强,同时采用多阶段训练,实现了分阶段的频谱恢复。

到目前为止,基于神经网络的单通道语音增强方法在处理非平稳噪声上已经取得了出色的性能。然而,大部分基于神经网络的语音增强方法总是充分使用全频带信息,而较少关注子频带信息。因此,如何能够将频带的信息充分利用是一个有价值的研究问题。早在20世纪40年代,Fletcher就提出了临界频段的概念,表明人耳对频率的实际感知与真实频率有非线性关系[17],这种非线性关系可以转化为巴克域[18],范围从1到24,对应于听力的前24个临界带。此后,一些研究者将此研究与传统方法结合进行语音增强[19],但是将掩蔽效应与深度学习方法相结合来构造语增强模型的研究却非常少。本文思考是否能结合深度学习方法,构造一种模拟人耳这种特殊声学结构的滤波器组,实现语音增强的目标。

针对单通道语音增强问题,本文提出了一种基于临界频带的交互性双支路模型(Interactive dual branch model based on the critical frequency band, IDBM-CFB)。具体而言,模型分为复数谱支路和幅度补偿支路,复数谱支路旨在通过估计cIRM达到获得干净语音复数谱的目标,幅度补偿支路负责幅度谱细节的填充。两条支路并行处理,相互合作,先将复数谱支路学到的中间特征幅度信息通过特定的模块传递给幅度补偿支路,在最后重构信号频谱时,再用幅度补偿支路的输出对复数谱支路幅度进行补偿,达到重构干净语音频谱的目标。实验结果表明,提出的基于临界频带的交互性双支路单通道语音增强模型在恢复语音质量和可懂度方面取得了很好的效果。

1 基础理论

1.1 人耳的临界频带

由于人耳的特殊结构,人的听觉系统对声音频率的感知与实际频率的对应关系是一种非线性映射关系。人耳基底膜具有类似于一组听觉滤波器[20]的作用,在0～22 kHz的频率范围内可以划分成24个临界频带,也称为Bark波段。Bark的划分相当于将基底膜分成许多小部分,每个部分对应一个频带[21],同一频带的声音会被叠加在大脑上共同评估。为此,本文将临界频带的划分方法引入到神经网络中,试图构建一种模拟人耳基底膜的听觉过滤系统。在IDBM-CFB中,子带划分方法如表1[18]所示。

表1 在16 kHz采样率和512傅里叶变换点数条件下的频带划分方式

Table 1 Division method of frequency band under the sampling rate of 16 kHz and the number of Fourier transform points of 512

Bark编号	临界频带划分范围/Hz	最接近临界频带的范围/Hz	对应频点范围
1	20～100	31.25～93.75	1～3
2	100～200	125～187.5	4～6
3	200～300	218.75～281.25	7～9
4	300～400	312.5～375	10～12
5	400～510	406.25～500	13～16
6	510～630	531.25～625	17～20
7	630～770	656.25～750	21～24
8	770～920	781.25～906.25	25～29
9	920～1 080	937.5～1 062.5	30～34
10	1 080～1 270	1 093.75～1 250	35～40
11	1 270～1 480	1 281.25～1 468.75	41～47
12	1 480～1 720	1 500～1 718.75	48～55
13	1 720～2 000	1 750～2 000	56～64
14	2 000～2 320	2 031.25～2 312.5	65～74
15	2 320～2 700	2 343.75～2 678.5	75～86
16	2 700～3 150	2 718.75～3 125	87～100
17	3 150～3 700	3 156.25～3 687.5	101～118
18	3 700～4 400	3 718.75～4 375	119～140
19	4 400～5 300	4 406.25～5 281.25	141～169
20	5 300～6 400	5 312.5～6 375	170～204
21	6 400～7 700	6 406.25～7 687.5	205～246
22	7 700～9 500	7 718.75～8 000	247～256
23	9 500～12 000		
24	12 000～15 500		

1.2 双支路语音增强模型重构策略

所提出的模型主要由两个支路构成,即幅度补偿支路和复数谱支路,旨在并行地协同估计干净语音的幅度和相位信息。具体来说,在复数谱支路中,输入是带噪语音的复数谱,模型估计cIRM,用于恢复目标语音的幅度和相位,然后幅度补偿支路利用带噪语音的幅度谱来估计干净语音的幅度补偿掩蔽,用于进一步补偿复数谱支路输出的幅度。设干净语音的频谱为$\boldsymbol{S}$,背景噪声的频谱为$\boldsymbol{N}$,带噪语音的频谱$\boldsymbol{X}$可表示为

$$X(t,f)=S(t,f)+N(t,f) \tag{1}$$

式中:$X(t,f)=X_{\mathrm{r}}(t,f)+\mathrm{j}X_{\mathrm{i}}(t,f)$;$S(t,f)=S_{\mathrm{r}}(t,f)+\mathrm{j}S_{\mathrm{i}}(t,f)$;$N(t,f)=N_{\mathrm{r}}(t,f)+\mathrm{j}N_{\mathrm{i}}(t,f)$;$X(t,f)$、$S(t,f)$和$N(t,f)$分别表示带噪音、干净语音和背景噪声在时频点$(t,f)$处的数值。复数谱支路输出的$\left|\widetilde{\boldsymbol{S}}_1(t,f)\right|\mathrm{e}^{\mathrm{j}\angle\tilde{\theta}_{\mathrm{S}}(t,f)}$表示初步得到的干净语音,用幅度补偿支路的输出$\left|\widetilde{M}(t,f)\right|$与之相乘,得到最终恢复出的干净语音为

$$\widetilde{\boldsymbol{S}}(t,f)=\left|\widetilde{M}(t,f)\right|\left|\widetilde{\boldsymbol{S}}_1(t,f)\right|\mathrm{e}^{\mathrm{j}\angle\tilde{\theta}_{\mathrm{S}}(t,f)} \tag{2}$$

式中$\left|\widetilde{M}(t,f)\right|$表示时频点$(t,f)$处的幅度补偿掩蔽,且$\left|\widetilde{M}(t,f)\right|\in(0,1)$。

恢复过程如图1所示,为了简单明了地说明恢复过程,仅以复数谱中的一个时频点为例。

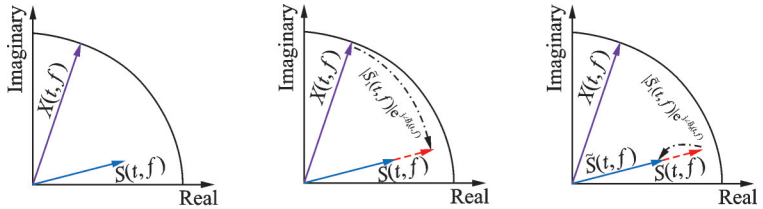

图1 一个时频点的恢复过程示意图

Fig.1 Diagram of the recovery process of one time-frequency point

从图1中可以看出,复数谱支路同时增强了信号的幅度和相位,而由于幅度和相位之间具有补偿问题[15],其并不能完好地恢复出干净语音,因此加入幅度补偿支路的结果能帮助解决这一问题。

2 模型构建

2.1 模型整体结构

本文所提出模型的整体结构如图2所示,两个支路都采用编码器-解码器的结构。幅度补偿支路的输入为带噪语音的幅度谱$|\boldsymbol{X}|$,采用6层核大小为(5,2)、步长为(2,1)的卷积层堆叠构成编码器,每层卷积后面都跟着批量归一化层(Batch normalization, BN)和参数校正线性单元(Parametric rectified linear unit, PReLU),通道数为[16, 16, 32, 32, 64, 64];解码器的参数设置与编码器相同,结构呈镜像对称。在复数谱支路中,带噪语音经过短时傅里叶变换后的实部$\boldsymbol{X}_{\mathrm{r}}$和虚部$\boldsymbol{X}_{\mathrm{i}}$作为两个独立的通道,构成了带噪语音的复数谱$\boldsymbol{Y}\in\mathbf{R}^{C_1\times F\times T}$,将$Y$输入到密集连接块得到频谱的高维特征表示$\boldsymbol{D}\in\mathbf{R}^{C_2\times F\times T}$

$$\boldsymbol{D}=\mathrm{Dense}(\boldsymbol{Y}) \tag{3}$$

式中:$C_1=2$、$C_2=32$表示通道数量;$F=256$表示频点个数;T表示特征图的帧数。按照表1中的方法沿频率维对$\boldsymbol{D}$进行分割,可以得到

$$\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_{22}=\mathrm{Chunk}(\boldsymbol{D}) \tag{4}$$

式中:$\boldsymbol{D}_i\in\mathbf{R}^{C_2\times \mathrm{Bark}_i\times T}$,$\mathrm{Bark}_i$表示第$i$个频段的频点个数,且$\sum_{i=1}^{22}\mathrm{Bark}_i=256$。将划分好的22个子带特征向量送入各自对应的由3层复卷积堆叠组成的复编码器中,得到编码后的22个特征向量

$$\boldsymbol{E}_i=\mathrm{Encoder}_i(\boldsymbol{D}_i) \tag{5}$$

式中:$\boldsymbol{E}_i\in\mathbf{R}^{128\times \mathrm{Bark}_i\times T}$,$1\leqslant i\leqslant 22$。具体来说,复编码器的通道数为[32,64,64,128],复卷积层的卷积核大小为(3,2)、步长为(1,1),每层复卷积之后都跟着复数形式的BN层和复数形式的PReLU层。复解码器的参数设置与复编码器相同,结构呈镜像对称。$\boldsymbol{E}_i$在被传递给第i个复解码器的同时,也被送入到了第i个高效通道注意力(Efficient channel attention, ECA)模块[22]中,通过ECA模块,重要通道能被给予更多的关注,从而让模型提取的特征指向性更强。对应的公式为

$$\mathbf{Eca}^i=\mathrm{ECA}_i(\boldsymbol{E}_i) \tag{6}$$

式中:$\mathbf{Eca}^i=(\mathbf{Eca}_{\mathrm{r}}^i,\mathbf{Eca}_{\mathrm{i}}^i)$,$\mathbf{Eca}_{\mathrm{r}}^i\in\mathbf{R}^{64\times \mathrm{Bark}_i\times T}$,$\mathbf{Eca}_{\mathrm{i}}^i\in\mathbf{R}^{64\times \mathrm{Bark}_i\times T}$。进一步地,信息融合与传递模块对22个ECA模块的输出进行整合,并将整合后的信息与幅度补偿支路上编码器的输出相乘。具体操作分为以下3个步骤:

步骤1 将22个复数谱特征向量$\mathbf{Eca}^i$转化为幅值特征向量$\mathbf{Eca}_{\mathrm{M}}^i$并沿频率维度拼接成完整的全频带特征$\mathbf{Eca}_{\mathrm{M}}$,即

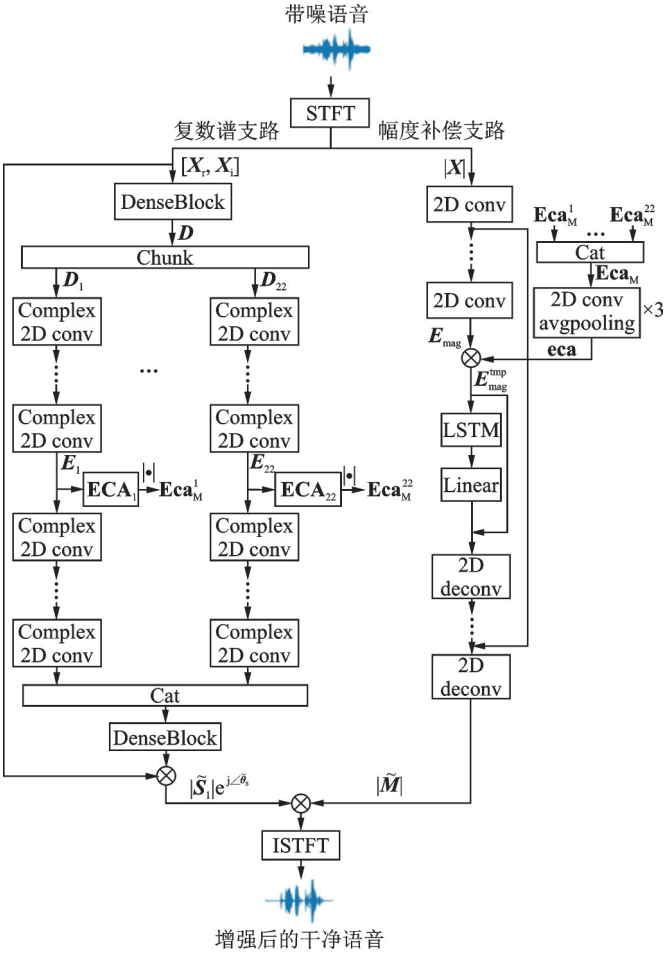

图2 所提模型的整体结构

Fig.2 Overall structure of the proposed model

$$\mathbf{Eca}_{\mathrm{M}}^{i}=\sqrt{\mathbf{Eca}_{\mathrm{r}}^{i}\cdot\mathbf{Eca}_{\mathrm{r}}^{i}+\mathbf{Eca}_{\mathrm{i}}^{i}\cdot\mathbf{Eca}_{\mathrm{i}}^{i}} \tag{7}$$

$$\mathbf{Eca}_{\mathrm{M}}=\mathrm{Cat}\left(\mathbf{Eca}_{\mathrm{M}}^{1},\mathbf{Eca}_{\mathrm{M}}^{2},\cdots,\mathbf{Eca}_{\mathrm{M}}^{22}\right) \tag{8}$$

式中：$\sqrt{\ }$ 表示分别对矩阵 $\boldsymbol{H}$ 中的每一个元素开方；“·”表示对应元素相乘；$\mathbf{Eca}_{\mathrm{M}}^{i}\in\mathbf{R}^{64\times\mathrm{Bark}_i\times T}$；$\mathbf{Eca}_{\mathrm{M}}\in\mathbf{R}^{64\times256\times T}$；$\mathrm{Cat}(\cdot)$表示将张量沿频率维拼接。

步骤2 $\mathrm{Eca}_{\mathrm{M}}$携带了从复数谱支路上学到的22个子带的幅度信息，将其通过3次卷积和平均池化操作，得到整合后的信息 $\mathbf{eca}\in\mathbf{R}^{64\times4\times T}$，即

$$\mathbf{eca}=F_{\theta}(\mathbf{Eca}_{\mathrm{M}}) \tag{9}$$

式中：$F_{\theta}(\cdot)$表示非线性变换；θ表示非线性变换中的可训练参数组。非线性变换包括3次卷积和池化操作，3次卷积的通道数为[64, 64, 64]，卷积核大小为(5, 1)、步长为(2, 1)，池化层核为(2, 1)、步长为(2, 1)。每经过一次卷积层或者池化层，频率维就会减小一半，每层卷积后面都跟着BN层和指数线性单元（Exponential linear unit, ELU）。特别注意的是，最后一层卷积后面采用Sigmoid激活函数代替ELU，使要传递的信息变化范围压缩在0到1之间，为幅度补偿支路提供额外的掩蔽。

步骤3 将eca与幅度补偿支路上编码器的输出相乘，可得

$$\boldsymbol{E}_{\mathrm{mag}}^{\mathrm{tmp}}=\mathbf{eca}\cdot\boldsymbol{E}_{\mathrm{mag}} \tag{10}$$

式中:$\boldsymbol{E}_{\mathrm{mag}}\in\mathrm{R}^{64\times4\times\mathrm{T}}$为幅度补偿支路上编码器的输出;$\boldsymbol{E}_{\mathrm{mag}}^{\mathrm{tmp}}\in\mathrm{R}^{64\times4\times\mathrm{T}}$。

将补偿后的特征$\boldsymbol{E}_{\mathrm{mag}}^{\mathrm{tmp}}$送入长短期记忆网络(Long short-term memory, LSTM)中进行时间上下文分析,然后再进行解码,得到幅度补偿支路输出的幅度补偿掩蔽$\left|\widetilde{\boldsymbol{M}}\right|$为

$$\left|\widetilde{\boldsymbol{M}}\right|=\mathrm{Decoder}_{\mathrm{mag}}\left(\mathrm{Linear}\left(\mathrm{LSTM}\left(\boldsymbol{E}_{\mathrm{mag}}^{\mathrm{tmp}}\right)\right)\right) \tag{11}$$

式中:$\left|\widetilde{\boldsymbol{M}}\right|\in\mathrm{R}^{256\times T}$;Linear(•)表示一次线性全连接操作;$\mathrm{Decoder}_{\mathrm{mag}}(\cdot)$表示幅度补偿支路的解码操作,参数设置与对应的编码器相同,结构呈镜像对称。

在复数谱支路上,经过最后一层密集连接模块,可以得到估计出的cIRM,进而初步得到干净语音的幅度和相位分别为

$$\begin{cases}\widetilde{\boldsymbol{M}}_{\mathrm{mag}}=\sqrt{\widetilde{\boldsymbol{M}}_{\mathrm{r}}\cdot\widetilde{\boldsymbol{M}}_{\mathrm{r}}+\widetilde{\boldsymbol{M}}_{\mathrm{i}}\cdot\widetilde{\boldsymbol{M}}_{\mathrm{i}}}\\ \widetilde{\boldsymbol{M}}_{\mathrm{phase}}=\arctan\left(\widetilde{\boldsymbol{M}}_{\mathrm{r}},\widetilde{\boldsymbol{M}}_{\mathrm{i}}\right)\end{cases} \tag{12}$$

$$\begin{cases}\left|\widetilde{\boldsymbol{S}}_1\right|=\widetilde{\boldsymbol{M}}_{\mathrm{mag}}\cdot\boldsymbol{X}_{\mathrm{mag}}\\ \angle\tilde{\boldsymbol{\theta}}_{\mathrm{S}}=\widetilde{\boldsymbol{M}}_{\mathrm{phase}}+\boldsymbol{X}_{\mathrm{phase}}\end{cases} \tag{13}$$

式中:$\widetilde{\boldsymbol{M}}_{\mathrm{r}}$和$\widetilde{\boldsymbol{M}}_{\mathrm{i}}$分别为估计出cIRM的实部和虚部;$\boldsymbol{X}_{\mathrm{mag}}$和$\boldsymbol{X}_{\mathrm{phase}}$分别为带噪语音的幅度谱和相位谱;$\left|\widetilde{\boldsymbol{S}}_1\right|$为复数谱支路初步恢复出的干净语音的幅度谱;$\angle\tilde{\boldsymbol{\theta}}_{\mathrm{S}}$为复数谱支路恢复出的干净语音的相位。

最后,按照式(2)将两个支路的输出做结合,即可恢复出干净语音频谱。

2.2 密集连接模块

如图3所示,密集连接层使用3个非线性变换函数,下一层的输入不仅来自上一层,还来自之前所有层的输出。带噪语音的实部和虚部频谱图被视为两个不同的输入通道。在网络第l层的输出$\boldsymbol{x}_l$可表示为

$$\boldsymbol{x}_l=H_l\left(\left[\boldsymbol{x}_{l-1},\boldsymbol{x}_{l-2},\cdots,\boldsymbol{x}_0\right]\right) \tag{14}$$

式中:H_l表示非线性变换,它是一个合并操作,包括二维卷积层,BN层和PReLU,$1\leqslant l\leqslant3$。卷积核大小为3×2、步长为1,每个H_l通过因果形式的补零操作确保输入和输出特征的长度和宽度保持不变,通道数为[16, 32, 32]。

经过3次非线性运算后,密集连接模块将通道从二维扩展到了32维,特征图的时间和频率维度不改变。每个通道都会表示特定的特征信息,从而达到扩大感受野和获得频谱高维特征表示的目的。在复数谱支路上,分别在支路的入口和出口处用到密集连接层,出口处与入口处的参数设置相同,结构呈对称关系。

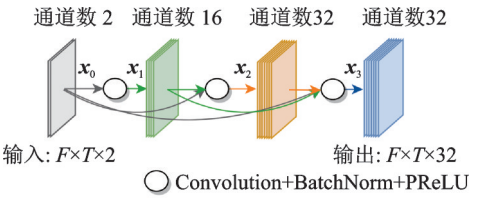

图3 密集连接模块结构

Fig.3 Structure of dense block

2.3 高效通道注意力模块

如图4所示,与传统结构不同的是,本文对实部和虚部都设置了对应的通道注意力模块。在经过一次频率维全局平均池化后,特征向量的实部和虚部分别通过考虑每k个相邻通道来实现局部跨通道交互,利用核大小为k的一维卷积和Sigmoid函数得到每个通道的权重,其中k值由自适应函数$f(k)$确定,即

$$f(k)=\left|\frac{\log_2 C+1}{2}\right|_{\text{odd}} \tag{15}$$

式中：$|x|_{\text{odd}}$表示离x最近的奇数；C为通道数；k的大小表示了局部跨通道交互的覆盖范围，即有多少个相邻通道参与了一个通道注意力的预测。最后，将所得到的权重与原特征相乘得到更新后的特征。

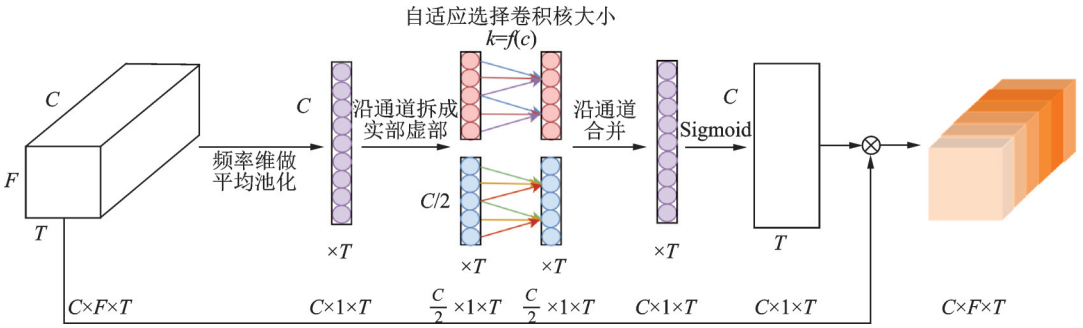

图4 高效通道注意力模块结构

Fig.4 Structure of ECA

将ECA模块加在复数谱支路和幅度补偿支路之间，对增加模型的复杂度方面几乎可以忽略不计，但却能使中间信息的特征指向性更强，更有效地将复数谱支路上学到的22个局部幅度谱信息传递给幅度补偿支路。

2.4 损失函数

为了缓解幅度和相位之间的补偿问题，同时优化信号的幅度谱损失和复数谱损失，即

$$\begin{cases} L^{\text{Mag}}=\left\| \sqrt{\left|\widetilde{S}_{\text{r}}\right|^2+\left|\widetilde{S}_{\text{i}}\right|^2}-\sqrt{\left|S_{\text{r}}\right|^2+\left|S_{\text{i}}\right|^2} \right\|^2 \\ L^{\text{RI}}=\left\|\widetilde{S}_{\text{r}}-S_{\text{r}}\right\|^2+\left\|\widetilde{S}_{\text{i}}-S_{\text{i}}\right\|^2 \\ L^{\text{Full}}=\left(L^{\text{Mag}}+L^{\text{RI}}\right)/2 \end{cases} \tag{16}$$

式中：L^{Mag}和L^{RI}分别表示幅度谱损失和复数谱损失；S_{r}和S_{i}分别表示干净语音频谱的实部和虚部分量；$\widetilde{S}_{\text{r}}$和$\widetilde{S}_{\text{i}}$分别表示增强后的语音频谱的实部和虚部分量。通过减少式(16)中幅度谱损失和复数谱损失来提高语音增强质量。

3 数据集和实验

3.1 实验数据集

为了验证IDBM-CFB的有效性，选用WSJ0-SI84语料库[23]和MUSAN噪声集[24]进行实验。WSJ0-SI84语料库包含83个说话人(42男41女)的7 138条话语，MUSAN噪声集包含930条噪声。从中选出的训练集、验证集和测试集的构成情况如表2所示。

将训练集的6 684条话语与从820条噪声中随机抽取的3条噪声进行混合，生成20 052条语音样本进行训练；将训练集的606条话语与从70条噪声中随机抽取的6条噪声进行混合，生成3 636条语音样本进行验证；将测试集的454条话语与从40条噪声中随机抽取的5条噪声进行混合，生成2 270条语音样本进行测试。

表2 实验数据集的构成情况

Table 2 Composition of experimental datasets

数据集	WSJ0-SI84 语料库	MUSAN 噪声集	每条语音随机 混合噪声数量	混合信噪比/dB
训练集	39男和38女,共6 684条话语	820	3	(−5～0),间隔为1
验证集	3男3女(来自训练集),共606条话语	70	6	(−3～3),间隔为3
测试集	3男3女,共454条话语	40	5	(−6～6),间隔为3

3.2 参数设置

实验数据集都被降采样到16 kHz,最大话语长度为3 s,采用512点的离散傅里叶变换,帧长为400点,帧移为100点,加汉宁窗。提出的模型和所有其他基线模型都使用Adam优化器,初始学习率为5e−4,每训练一轮学习率就下降为原来的97%。考虑到幂律压缩在去混响和去噪任务中的有效性[25-26],本文对所有基线模型都使用了幂律压缩,在保持相位不变的情况下对频谱幅度进行压缩,压缩系数为0.5,以更好地衰减背景噪声。所有卷积层的补零均为因果补零,保证时间轴的输入输出维度不变。此外,文中所有的复数型模型结构都借鉴于文献[27]。

3.3 评价指标

实验采用感知语音质量评价(Perceptual evaluation of speech quality, PESQ)[28]、短时客观可懂度(Short-time objective intelligibility, STOI)[29]、信号失真比(Source-to-distortion ratio, SDR)和3个平均主观意见分(Mean opinion score, MOS)[30](即CSIG、CBAK和COVL)指标来评估增强后的语音质量,其中CSIG、CBAK和COVL分别用于测量信号失真、背景噪声质量和整体音频质量评价。PESQ的取值范围为−0.5到4.5,STOI的取值范围为0到1,3个MOS评分范围为1到5,上述所有指标的值越高,表示语音质量越好。

3.4 基线模型

实验选用两个先进的编码器-解码器模型(CRN[9]和DCCRN[10])、两个先进的目标解耦模型(GaGNet[16]和基于两级网络的复数谱映射(Complex spectral mapping based two-stage network, CTS-Net[31]))和3个对照模型(Dual branch fullband(DB-Full)、Dual branch subband1(DB-Sub1)和Dual branch subband2(DB-Sub2))与所提的IDBM-CFB进行比较,所有模型均是因果模型。

CRN和DCCRN:CRN是基于映射的模型,它利用信号复数谱映射网络直接预测干净信号复数谱的实部和虚部。CRN包含一个编码器和两个解码器,其中编码器联合处理实部和虚部,两个解码器分别负责处理实部和虚部,在编码器和解码器之间有LSTM层学习信号的短时上下文信息。DCCRN是对CRN的改进模型,采用模拟复数运算的复数型网络结构。

GaGNet:GaGNet是一种用于单通道语音增强的复数域多阶段双支路学习框架,包括频谱特征提取模块和堆叠的扫视-凝视模块。在每个扫视-凝视模块中,模型将频谱优化任务分成幅度谱支路和复数谱支路两条路径,两条路径均提取全频带信息,协同促进频谱估计,同时,该模型还用到了多阶段学习策略,通过反复展开扫视-凝视模块,优化最终结果。

CTS-Net:CTS-Net是一个两阶段复数谱映射网络,在第一阶段,采用幅度粗估计网络估计幅度谱,然后与原始噪声相位耦合初步得到信号复数谱;在第二阶段,复数谱细化网络以原始和上一阶段得到的复数谱作为输入,在有效修复频谱的同时,进一步抑制噪声分量,优化最终结果。

DB-Full、DB-Sub1和DB-Sub2: DB-Full是全频带双支路模型,与IDBM-CFB相比,在复数谱支路

中,不拆分频率,且两个支路中间没有任何信息传递;DB-Sub1是子带双支路模型,与IDBM-CFB相比,两个支路中间没有任何信息传递;DB-Sub2与IDBM-CFB相比,两个支路中间有信息融合与传递模块,但是中间信息没有经过22个ECA模块。

3.5 实验结果和分析

针对单通道语音增强实验,表3列出了不同模型下PESQ和STOI的结果,表4列出了不同模型下SDR和CSIG的结果,表5列出了不同模型下CBAK和COVL的结果。

表3 不同模型下PESQ和STOI的评估结果

Table 3 Evaluation results of PESQ and STOI under different models

模型	参数/10^4	PESQ						STOI/%					
		−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave	−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave
Unprocessed		1.14	1.17	1.21	1.29	1.40	1.24	72.55	77.63	82.05	86.36	89.51	81.62
CRN	610	1.58	1.78	2.00	2.22	2.42	2.00	85.86	89.62	92.07	94.22	95.45	91.44
DCCRN	360	1.68	1.88	2.09	2.30	2.50	2.09	86.23	89.94	92.44	94.43	95.76	91.76
GaGNet	594	**1.88**	**2.10**	2.31	2.53	2.70	2.30	88.10	91.00	93.10	94.90	95.85	92.59
CTS-Net	435	1.83	2.04	2.24	2.40	2.54	2.21	**88.47**	**91.58**	**93.47**	**95.00**	95.99	**92.90**
DB-Full	**147**	1.73	1.95	2.17	2.40	2.59	2.17	86.43	90.04	92.35	94.50	95.69	91.80
DB-Sub1	333	1.78	2.00	2.25	2.49	2.68	2.24	86.67	90.20	92.48	94.56	95.74	91.93
DB-Sub2	339	1.81	2.04	2.28	2.52	2.73	2.27	87.09	90.56	92.80	94.77	95.89	92.22
IDBM-CFB	339	1.84	2.07	**2.32**	**2.56**	**2.77**	**2.31**	87.58	90.93	93.10	**95.00**	**96.10**	92.54

表4 不同模型下SDR和CSIG的评估结果

Table 4 Evaluation results of SDR and CSIG under different models

模型	SDR/dB						CSIG					
	−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave	−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave
Unprocessed	−5.84	−2.91	0.06	3.04	6.03	0.08	2.08	2.27	2.48	2.69	2.93	2.49
CRN	7.74	9.99	11.97	13.79	15.46	11.79	3.13	3.38	3.60	3.81	4.00	3.58
DCCRN	10.15	12.15	13.94	15.61	**17.11**	13.79	2.90	3.17	3.41	3.65	3.86	3.40
GaGNet	10.74	12.37	13.75	15.15	16.28	13.66	3.43	3.66	3.87	4.05	4.21	3.84
CTS-Net	**11.02**	**12.89**	**14.41**	**15.71**	16.89	**14.18**	3.34	3.58	3.77	3.94	4.08	3.74
DB-Full	8.95	11.22	12.89	14.63	16.07	12.75	3.34	3.58	3.80	4.00	4.16	3.77
DB-Sub1	9.24	11.37	13.07	14.75	16.25	12.93	3.37	3.61	3.84	4.04	4.21	3.81
DB-Sub2	9.14	11.38	13.15	14.91	16.37	12.99	3.40	3.64	3.86	4.05	4.24	3.84
IDBM-CFB	9.30	11.60	13.31	15.02	16.64	13.17	**3.44**	**3.67**	**3.90**	**4.09**	**4.28**	**3.88**

观察表3、4和表5,分别从分带、信息融合与传递模块和ECA模块3方面评估所提模型的有效性。

(1)分带:对比DB-Full和DB-Sub1的实验结果,可以发现,分频带后的DB-Sub1结果比全频带的DB-Full结果都有提高,PESQ提高了0.07,STOI提高了0.13%,SDR提高了0.18 dB,CSIG提高了0.04,CBAK提高了0.05,COVL提高了0.05,表明这种按照临界频带进行分频的思想是有效的。

(2)信息融合与传递模块:对比DB-Sub1和DB-Sub2的实验结果,添加信息融合与传递模块后的DB-Sub2结果比无信息融合与传递模块的DB-Sub1结果都有进一步提高,PESQ提高了0.03,STOI提

表5 不同模型下CBAK和COVL的评估结果

Table 5 Evaluation results of CBAK and COVL under different models

模型	CBAK						COVL					
	−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave	−6 dB	−3 dB	0 dB	3 dB	6 dB	Ave
Unprocessed	1.55	1.69	1.86	2.07	2.31	1.90	1.52	1.63	1.78	1.94	2.12	1.80
CRN	2.31	2.55	2.79	3.05	3.28	2.80	2.32	2.56	2.79	3.02	3.22	2.78
DCCRN	2.34	2.58	2.82	3.07	3.30	2.82	2.25	2.50	2.74	2.97	3.19	2.73
GaGNet	**2.66**	**2.87**	**3.08**	3.26	3.41	**3.06**	**2.63**	2.86	3.09	3.30	3.48	3.07
CTS-Net	2.65	2.84	3.02	3.17	3.30	3.00	2.57	2.80	3.00	3.17	3.32	2.97
DB-Full	2.43	2.68	2.92	3.18	3.40	2.92	2.52	2.76	2.99	3.22	3.41	2.98
DB-Sub1	2.47	2.71	2.98	3.23	3.46	2.97	2.56	2.80	3.05	3.28	3.47	3.03
DB-Sub2	2.48	2.73	2.99	3.25	3.49	2.99	2.58	2.83	3.07	3.31	3.51	3.06
IDBM-CFB	2.51	2.76	3.02	**3.28**	**3.52**	3.02	2.62	**2.87**	**3.12**	**3.35**	**3.56**	**3.10**

高了0.29%,SDR提高了0.06 dB,CSIG提高了0.03,CBAK提高了0.02,COVL提高了0.03,表明在两条支路之间加入信息融合与传递模块来增加两条支路的交互性是有效的。

(3)ECA模块:对比DB-Sub2和IDBM-CFB的实验结果,对每个子带学到的信息先用ECA模块整合再做融合的结果比简单地将22个子带学到的信息融合的结果有进一步的提高,PESQ提高了0.04,STOI提高了0.32%,SDR提高了0.18 dB,CSIG提高了0.04,CBAK提高了0.03,COVL提高了0.04。这表明采用轻量级的ECA模块帮助子带信息进行整合是有效的。

通过以上分析可以看出,虽然直接分带计算提升的语音质量有限,但是通过加入了信息融合与传递模块和ECA模块,在参数量增加不大的情况下,两个模块结合分带思想,较好地提升了模型的整体性能。因此,分带、信息融合与传递模块和ECA模块是本文所提出的IDBM-CFB提升单通道语音增强性能的关键。

最后,对比IDBM-CFB、CRN和DCCRN,可以看出,IDBM-CFB有更优秀的语音增强性。对比IDBM-CFB和GaGNet,IDBM-CFB的参数量减少了255万,而其在除SDR以外的所有客观和主观指标上达到GaGNet;对比IDBM-CFB和CTS-Net,IDBM-CFB的参数量减少了96万,而其在除STOI和SDR以外的所有客观和主观指标上超过CTS-Net。总的来说,在6个评价指标中,IDBM-CFB有4个指标均与上述基线模型相当或更优,而其中一个或两个指标的不足,对比模型参数量的减少是值得的。

4 结束语

针对单通道语音增强问题,本文将语音心理声学上临界频带的概念与深度学习相结合,引入分带、信息融合与传递模块和高效通道注意力模块,使复数谱支路和幅度补偿支路相互合作,提出了基于临界频带的交互性双支路模型(IDBM-CFB)。在WSJ0-SI84语料库和MUSAN噪声集进行了训练、验证和测试实验。实验结果表明,IDBM-CFB能够以更少的参数量,在大部分客观和主观评价指标上达到或超过对比基线模型,提升了单通道语音增强性能。

参考文献:

[1] XIA Bingyin, BAO Changchun. Wiener filtering-based speech enhancement with weighted denoising autoencoder and noise

classification[J]. Speech Communication, 2014, 60(1): 13-29.

[2] HERMUS K, WAMBACQ P, VAN HAMME V. A review of signal subspace speech enhancement and its application to noise robust speech recognition[J]. EURASIP Journal on Advances in Signal Processing, 2006, 2007(1): 1-15.

[3] LIN L, AMBIKAIRAJAH E, HOLMES W H. Speech enhancement for nonstationary noise environment[C]//Proceedings of Asia-Pacific Conference on Circuits and Systems. [S.l.]: IEEE, 2002: 177-180.

[4] HSIEH T A, WANG H M, LU X, et al. WaveCRN: An efficient convolutional recurrent neural network for end-to-end speech enhancement[J]. IEEE Signal Processing Letters, 2020, 27: 2149-2153.

[5] WANG Yuxuan, NARAYANAN A, WANG Deliang. On training targets for supervised speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1849-1858.

[6] XU Yong, DU Jun, DAI Lirong, et al. A regression approach to speech enhancement based on deep neural networks[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(1): 7-19.

[7] PALIWAL K, WOJCICKI K, SHANNON B. The importance of phase in speech enhancement[J]. Speech Communication, 2011, 53(4): 465-494.

[8] WILLIAMSON D S, WANG Yuxuan, WANG Deliang. Complex ratio masking for monaural speech separation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(3): 483-492.

[9] TAN K, WANG Deliang. Complex spectral mapping with a convolutional recurrent network for monaural speech enhancement [C]//Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2019: 6865-6869.

[10] HU Yanxin, LIU Yun, LV Shubo, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[C]//Proceedings of Interspeech. [S.l.]: [s.n.], 2020: 2472-2476.

[11] YU Runxiang, ZHAO Ziwei, YE Zhongfu. PFRNet: Dual-branch progressive fusion rectification network for monaural speech enhancement[J]. IEEE Signal Processing Letters, 2022, 29: 2358-2362.

[12] YU Guochen, LI Andong, ZHENG Chengshi, et al. Dual-branch attention-in-attention transformer for single-channel speech enhancement[C]//Proceedings of 2022 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2022: 7847-7851.

[13] YU Guochen, LI Andong, WANG Hui, et al. DBT-Net: Dual-branch federative magnitude and phase estimation with attention-in-attention transformer for monaural speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 2629-2644.

[14] YIN Dacheng, LUO Chong, XIONG Zhiwei, et al. PHASEN: A phase-and-harmonics-aware speech enhancement network [C]//Proceedings of Conference on Artificial Intelligence. [S.l.]: AAAI, 2020: 9458-9465.

[15] WANG Zhongqiu, WICHERN G, ROUX J L. On the compensation between magnitude and phase in speech separation[J]. IEEE Signal Processing Letters, 2021, 28: 2018-2022.

[16] LI Andong, ZHENG Chengshi, ZHANG Lu, et al. Glance and gaze: A collaborative learning framework for single channel speech enhancement[J]. Applied Acoustics, 2022, 187: 1-9.

[17] FLETCHER H. Auditory patterns[J]. Reviews of Modern Physics, 1940, 12(1): 47-65.

[18] ZWICKER E. Subdivision of the audible frequency range into critical bands (frequenzgruppen)[J]. The Journal of the Acoustical Society of America, 1961, 33(2): 248.

[19] TSOUKALAS D E, MOURJOPOULOS J N, KOKKINAKIS G. Speech enhancement based on audible noise suppression [J]. IEEE Transactions on Speech and Audio Processing, 1997, 5(6): 497-514.

[20] MUNKONG R, JUANG B H. Auditory perception and cognition[J]. IEEE Signal Processing Magazine, 2002, 25(3): 98-117.

[21] MOORE B. Parallels between frequency selectivity measured psychophysically and in cochlear mechanics[J]. Scand Audio Suppl, 1986, 25: 139-152.

[22] WANG Qilong, WU Banggu, ZHU Pengfei, et al. ECA-Net: Effificient channel attention for deep convolutio-nal neural networks[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2020: 11534-

11542.

[23] PAUL D B, BAKER J M. The design for the wall street journal-based CSR corpus[C]//Proceedings of the Workshop on Speech and Natural Language. [S.l.]: Association for Computational Linguistics, 1992: 357-362.

[24] SNYDER D, CHEN G, POVEY D. Musan: A music, speech, and noise corpus[EB/OL]. (2015-10-28). https://arXiv preprint arXiv:1510.08484.

[25] LI Andong, LIU Wenzhe, LUO Xiaoxue, et al. A simultaneous denoising and dereverberation framework with target decoupling[C]//Proceedings of Interspeech. [S.l.]: [s.n.], 2021: 2801-2805.

[26] LI Andong, ZHENG Chengshi, PENG Runhua, et al. On the importance of power compression and phase estimation in monaural speech dereverberation[J]. JASA Express Letters, 2021. DOI:10.1121/10.0003321.

[27] TRABELSI C, BILANIUK O, ZHANG Y, et al. Deep complex networks[EB/OL]. (2018-02-25). https://doi.org/10.48550/arXiv.1705.09792.

[28] RIX A W, BEERENDS J G, HOLLIER M P, et al. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs[C]//Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE, 2001: 749-752.

[29] TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time-frequency weighted noisy speech[C]//Proceedings of 2011 IEEE Transactions on Audio, Speech, and Language Processing. [S.l.]: IEEE, 2011: 2125-2136.

[30] HU Yi, LOIZOU P C. Evaluation of objective quality measures for speech enhancement[C]//Proceedings of 2008 IEEE Transactions on Audio, Speech, and Language Processing. [S.l.]: IEEE, 2008: 229-238.

[31] LI Andong, LIU Wenzhe, ZHENG Chengshi, et al. Two heads are better than one: A two-stage complex spectral mapping approach for monaural speech enhancement[J]. IEEE-ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1829-1843.

作者简介:

叶中付(1959-),**通信作者**,男,教授,博士生导师,研究方向:阵列信号处理、语音及语言信息处理和图像处理,E-mail: yezf@ustc.edu.cn。

赵紫微(1998-),女,硕士研究生,研究方向:语音信号处理。

于润祥(1998-),男,硕士研究生,研究方向:语音信号处理。

(编辑:王静)