第 42 卷 第 2 期
2023 年 3 月

应用声学
Journal of Applied Acoustics

Vol. 42, No. 2
March, 2023

◇ 研究报告 ◇

深度复卷积递归网络模型的师生学习语声增强方法*

卞金洪[1] 吴瑞琦[1] 周 锋[1†] 赵 力[2]

(1 盐城工学院信息工程学院 盐城 224051)

(2 东南大学信息科学与工程学院 南京 210096)

摘要: 基于深度神经网络的方法已经在语声增强领域得到了广泛的应用,然而若想取得理想的性能,一般需要规模较大且复杂度较高的模型。因此,在计算资源有限的设备或对延时要求高的环境下容易出现部署困难的问题。为了解决此问题,提出了一种基于深度复卷积递归网络的师生学习语声增强方法。在师生深度复卷积递归网络模型结构中间的复长短时记忆递归模块提取实部和虚部特征流,并分别计算帧级师生距离损失以进行知识转移。同时使用多分辨率频谱损失以进一步提升低复杂度学生模型的性能。实验在公开数据集 Voice Bank Demand 和 DNS Challenge 上进行,结果显示所提方法相对于基线学生模型在各项指标上均有明显提升。

关键词: 语声增强;递归神经网络;长短期记忆网络;知识蒸馏

中图法分类号: TN912 **文献标识码:** A **文章编号:** 1000-310X(2023)02-0269-07

DOI: 10.11684/j.issn.1000-310X.2023.02.009

Teacher-student learning for speech enhancement based on deep complex convolution recurrent network

BIAN Jinhong[1] WU Ruiqi[1] ZHOU Feng[1] ZHAO Li[2]

(1 *School of Information Technology, Yancheng Institute of Technology, Yancheng* 224051, *China*)

(2 *School of Information Science and Engineering, Southeast University, Nanjing* 210096, *China*)

Abstract: Deep learning-based methods have been widely used in the field of speech enhancement. However, a model with large scale and high complexity is typically required to achieve the desired performance. Hence, deployment difficulties may occur in devices with limited hardware resources or in applications with strict latency requirements. In order to solve this problem, a teacher-student learning method for speech enhancement based on deep complex convolution recurrent network (DCCRN) is proposed. The real and imaginary feature streams are extracted from the output of the complex long short term memory (Complex LSTM) in the middle of the DCCRN model, and the frame-level teacher-student distance loss is calculated to transfer knowledge. Meanwhile, the multi-resolution short-time Fourier transform loss is used to further improve the performance of the low-complexity student model. The experiment was conducted on the open source dataset Voice Bank Demand and DNS Challenge, and the results show that the proposed method has a significant improvement in various indicators compared with the baseline student model.

Keywords: Speech enhancement; Recurrent neural networks; Long short-term memory networks; Knowledge distillation

2021-12-02 收稿; 2022-03-07 定稿

*国家自然科学基金项目 (61673108), 江苏省高等学校自然科学研究重大项目 (19KJA110002), 江苏省高校自然科学研究面上项目 (19KJB510061), 江苏省自然科学基金项目 (BK20181050), 江苏省产学研指导项目 (BY2020358, BY2020335)

作者简介: 卞金洪 (1966–), 男, 江苏盐城人, 副教授, 研究方向: 语声和图像信号处理。

† 通信作者 E-mail: zfycit@ycit.edu.cn

0 引言

语声增强的研究旨在消除背景噪声，提高语声的质量和可懂度。自 20 世纪 50 年代以来，语声增强算法已经吸引了国内外众多学者的关注[1−2]。语声增强方法在改善人类或机器对语声的理解方面有重要的作用，包括助听器、语声通信和自动语声识别等任务。本文主要关注单通道的语声增强方法。传统的语声增强方法通常基于统计信号处理理论，对带噪语声应用频带抑制增益或滤波器。但这类方法往往基于很多经验性的假设，并且难以应对非平稳的噪声[3]。

得益于深度学习的发展，语声增强任务被定义为一个有监督的学习问题。这种数据驱动的方法渐渐成为主流，因为它能够从海量的带噪和干净语声对挖掘有效信息，从而学到强大的噪声抑制能力(特别是对于非平稳噪声)。众多基于深度学习的模型已经在单通道的语声增强任务报告了优良的性能。但若想取得理想的性能，一般需要一个较大规模的深度神经网络 (Deep neural network, DNN) 模型，这也就意味着耗费大量的计算资源和存储空间。因此，在对延迟敏感的应用程序或资源有限的设备(比如耳机、助听器) 上部署此类语声增强算法将会遇到困难。为了实现基于深度学习的语声增强模型的落地部署，有必要研究如何降低模型的存储和计算量。

目前主流的模型压缩方法，比如剪枝、量化和知识蒸馏，均在降低模型的复杂度方面有一定成效[4]。第一类是网络剪枝方法 (Network pruning)，这类方法通过一定的策略选择并删除具有高冗余度的参数，仅保留信息量最大、最重要的参数。同时对剩余的参数进行一定的微调以保证一致性[5]。另一类有效的模型压缩方法是网络量化 (Network quantization)，其通过减少表示每个权重所需的位数来压缩原始网络[6]。而本文主要关注知识蒸馏的方法，其核心思想是将知识从大型教师模型传递给小型学生模型[7]。师生学习多应用于分类任务，在语声增强这样的回归任务上的相关工作并不多见。本文提出了一种用于语声增强模型的师生学习方法，通过拉近师生模型输出的距离，将大规模教师模型的有效信息传递给学生模型。同时，使用多分辨率频谱 (Multi-resolution short-time Fourier transform, MRSTFT) 损失[8]代替原深度复卷积递归网络 (Deep complex convolution recurrent network, DCCRN) 模型使用的尺度不变信噪比 (Scale-invariant source-to-noise ratio, SISNR) 损失，进一步提升低复杂度学生模型的效果。

1 基于师生学习的语声增强模型框架

本文提出一种用于语声增强的师生学习框架，以 DCCRN[9] 作为基线模型进行设计，整体结构如图 1 所示。

本文所采用的师生模型均具有对称式的编码器和解码器结构，而中间设置复数长短期记忆 (Long short-term memory, LSTM) 层。对于基础模型，输入特征选取短时傅里叶变换 (Short-time Fourier transform, STFT) 后的复频谱，而将网络的输出应用 MRSTFT 损失以引导优化。师生学习的位置设置在中间的复 LSTM 层，分别提取教师和学生的实部和虚部特征流以计算师生距离损失。

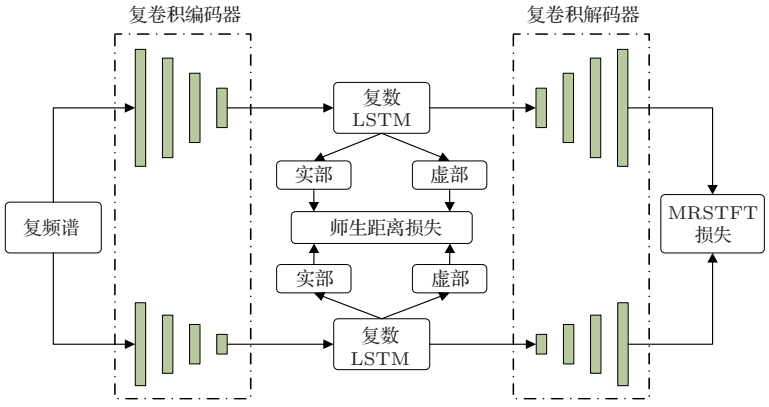

图 1 基于 DCCRN 模型的师生学习框架

Fig. 1 A framework for teacher-student learning based on the DCCRN model

1.1 复卷积递归网络结构

本文所提出的师生语声增强模型的基线模型选用文献[9]中的DCCRN模型，其采用对称式设计的卷积编码器-解码器(Convolutional encoder-decoder, CED)结构，在编码器和解码器中间插入LSTM层用于建模时间依赖关系。具体地说，编码器有6个正向Conv2d卷积块组成，旨在逐步从输入特征中提取高级特征并降低分辨率。而解码器则具有与编码器对称的6个反向Conv2d块，其设置是为了将低分辨率特征重构为输入的原始大小。而中间的LSTM层承担了捕获语声特征长时相关性的任务，并维持了整个框架的因果性。编码器和解码器的Conv2d模块由卷积层/反卷积层构成，并后接批归一化层和激活函数。编码器和解码器的对应层设置跳过连接以促进梯度的流动。

与一般的CED结构不同，DCCRN模型将所有的卷积层、批归一化层和LSTM层均重构为模拟复数运算的形式，因此其处理复频谱的特征更为有利。复卷积模块包含复数Conv2d块、复批归一化层和实值PReLU激活函数。复Conv2d块由4个传统Conv2d块联合计算构成。假设复数卷积滤波器W定义为$W_{\mathrm{r}}+\mathrm{j}W_{\mathrm{i}}$，其中实值矩阵$W_{\mathrm{r}}$和$W_{\mathrm{i}}$分别表示复卷积核的实部和虚部。因此复卷积层的输出特征F_{out}由复数信息流和复卷积滤波器的模拟卷积运算得出：

$$\begin{aligned}F_{\mathrm{out}} = {} & (X_{\mathrm{r}} * W_{\mathrm{r}} - X_{\mathrm{i}} * W_{\mathrm{i}}) \\ & + \mathrm{j}\,(X_{\mathrm{r}} * W_{\mathrm{i}} + X_{\mathrm{i}} * W_{\mathrm{r}})\,. \end{aligned} \tag{1}$$

类似的，给定输入特征图的实部X_{r}和虚部X_{i}，复数LSTM层的计算过程为

$$\begin{aligned} & F_{\mathrm{rr}} = \mathrm{LSTM}_{\mathrm{r}}(X_{\mathrm{r}}), \quad F_{\mathrm{ir}} = \mathrm{LSTM}_{\mathrm{r}}(X_{\mathrm{i}}), \\ & F_{\mathrm{ri}} = \mathrm{LSTM}_{\mathrm{i}}(X_{\mathrm{r}}), \quad F_{\mathrm{ii}} = \mathrm{LSTM}_{\mathrm{i}}(X_{\mathrm{i}}), \\ & F_{\mathrm{out}} = (F_{\mathrm{rr}} - F_{\mathrm{ii}}) + \mathrm{j}\,(F_{\mathrm{ri}} - F_{\mathrm{ir}})\,, \end{aligned} \tag{2}$$

其中，F_{rr}、F_{ir}、F_{ri}、F_{ii}表示模拟复数运算的交叉计算中间量。一个复数LSTM模块包含两个传统的LSTM层$\mathrm{LSTM}_{\mathrm{r}}$和$\mathrm{LSTM}_{\mathrm{i}}$，分别表征实部和虚部部分。

DCCRN模型的输入特征为语声STFT后的复频谱，而训练目标则为极坐标下的复掩蔽。假设原带噪语声的频谱极坐标形式为$Y = Y_{\mathrm{mag}} \cdot \mathrm{e}^{\mathrm{j}Y_{\mathrm{phase}}}$，而通过DCCRN模型预测的复掩蔽为$\widetilde{M} = \widetilde{M}_{\mathrm{r}} + \mathrm{j}\widetilde{M}_{\mathrm{i}}$，其中$\widetilde{M}_{\mathrm{r}}$和$\widetilde{M}_{\mathrm{i}}$分别表示复掩蔽的实部和虚部。则可以重建极坐标形式下的复掩蔽为

$$\begin{cases} \widetilde{M}_{\mathrm{mag}} = \sqrt{\widetilde{M}_{\mathrm{r}}^2 + \widetilde{M}_{\mathrm{i}}^2}, \\ \widetilde{M}_{\mathrm{phase}} = \arctan\left(\widetilde{M}_{\mathrm{i}}, \widetilde{M}_{\mathrm{r}}\right), \end{cases} \tag{3}$$

其中，$\widetilde{M}_{\mathrm{mag}}$表示估计的干净语声幅度谱，$\widetilde{M}_{\mathrm{phase}}$表示估计的相位谱，arctan表示反正切函数。继而，可以通过预测的复掩蔽还原干净语声频谱$\tilde{S}$：

$$\tilde{S} = Y_{\mathrm{mag}} \cdot \widetilde{M}_{\mathrm{mag}} \cdot \mathrm{e}^{\mathrm{j}\left(Y_{\mathrm{phase}} + \widetilde{M}_{\mathrm{phase}}\right)}. \tag{4}$$

1.2 MRSTFT损失

本文使用频谱图量级上的多分辨率损失[8]作为整体网络的损失函数。假定y和$\hat{y}$分别代表干净语声信号和增强语声信号，那么单一尺度的STFT谱尺度损失$L_{\mathrm{stft}}(y,\hat{y})$可以定义为谱收敛(spectral convergence, sc)损失和幅度谱(magnitude, mag)损失的叠加：

$$L_{\mathrm{stft}}(y,\hat{y}) = L_{\mathrm{sc}}(y,\hat{y}) + L_{\mathrm{mag}}(y,\hat{y})\,, \tag{5}$$

其中，谱收敛损失和幅度谱损失分别定义为

$$\begin{aligned} L_{\mathrm{sc}}(y,\hat{y}) &= \frac{\big\|\,|\mathrm{STFT}(y)| - |\mathrm{STFT}(\hat{y})|\,\big\|_F}{\big\|\,|\mathrm{STFT}(y)|\,\big\|_F}, \\ L_{\mathrm{mag}}(y,\hat{y}) &= \frac{1}{N}\left\|\lg|\mathrm{STFT}(y)| - \lg|\mathrm{STFT}(\hat{y})|\right\|_1, \end{aligned} \tag{6}$$

其中，$\|\cdot\|_F$和$\|\cdot\|_1$分别表示Frobenius和L_1规范化，$|\mathrm{STFT}(\cdot)|$和N分别是STFT幅度谱和其中的元素数量。而MRSTFT损失L_{MRSTFT}则是具有不同分析参数(即快速傅里叶变换大小、窗长和帧移)的单尺度STFT损失的叠加：

$$L_{\mathrm{MRSTFT}} = \frac{1}{M}\sum_{m=1}^{M} L_{\mathrm{STFT}}^{(m)}(y,\hat{y}), \tag{7}$$

其中，M是不同分析参数的频谱损失数目，每个$L_{\mathrm{STFT}}^{(m)}$应用不同尺度的快速傅里叶变换大小、窗长和帧移。其中快速傅里叶变换大小取自$\{512, 1024, 2048\}$，窗长取自$\{240, 600, 1200\}$，帧移取自$\{50, 120, 240\}$。

1.3 用于知识转移的师生学习方法损失

师生学习是一种有效的知识转移方法，首先预训练大规模的教师网络，然后通过师生学习教导小规模的学生模型做出与教师网络相同的推断。本文基于DCCRN模型实施了师生学习方法，表1中展示了教师模型和学生模型的超参数设置。其中卷积层参数从左到右依次是卷积核大小(时间维度×频率维度)、卷积步长(stride)和通道数。每个复LSTM层具有两个传统LSTM层，分别对应实数部

分和虚数部分，隐层节点数已在表1中给出。教师模型和学生模型的主要差异在于卷积层的通道数，注意到教师模型和学生模型在复LSTM模块具有相同的隐层节点数，这建立了教师和学生模型间的沟通，使得知识的转移能够通过拉近师生特征距离来实现。教师模型总计具有2.8 M的参数，而学生模型仅有0.23 M的参数，为教师模型的8.2%。

表1 师生模型的超参数设置

Table 1 Hyperparameter settings for teacher and student models

层名称	超参数	
	学生模型	教师模型
Conv_1	2×5, (1,2), 8	2×5, (1,2), 32
Conv_2	2×5, (1,2), 16	2×5, (1,2), 64
Conv_3	2×5, (1,2), 32	2×5, (1,2), 128
Conv_4	2×5, (1,2), 64	2×5, (1,2), 256
Conv_5	2×5, (1,2), 64	2×5, (1,2), 256
Conv_6	2×5, (1,2), 64	2×5, (1,2), 256
LSTM_1(×2)	64	64
LSTM_2(×2)	64	64
Deconv_1	2×5, (1,2), 64	2×5, (1,2), 256
Deconv_2	2×5, (1,2), 64	2×5, (1,2), 256
Deconv_3	2×5, (1,2), 32	2×5, (1,2), 128
Deconv_4	2×5, (1,2), 16	2×5, (1,2), 64
Deconv_5	2×5, (1,2), 8	2×5, (1,2), 32
Deconv_6	2×5, (1,2), 2	2×5, (1,2), 2

在预训练教师模型后，让学生模型模仿教师的输出。师生学习的过程通过在学生模型原损失的基础上附加师生距离损失实现，学生模型的损失$\mathcal{L}_{\text{stu}}$为

$$\mathcal{L}_{\text{stu}} = \mathcal{L}_{\text{MRSTFT}} + \beta\mathcal{L}_{\text{distance}}, \tag{8}$$

其中，$\mathcal{L}_{\text{MRSTFT}}$是原模型的MRSTFT损失，$\mathcal{L}_{\text{distance}}$是教师和学生模型输出差异的测度。这里$\beta \geqslant 0$为用于衡量两种损失的权重参数，本文中设置为1。师生学习的位置设置在编码器和解码器中间的复LSTM模块。由于DCCRN模型本身的对称性，在靠近中部的高层次特征具有从靠近两端的低层次特征中学习有效信息的能力。并且复LSTM模块的参数量占总体的30%，承担了语声帧间相关性的分析任务，因此在该处实施知识转移能够更好地传递有效信息。由于复LSTM模块的计算分为实部和虚部两个流向，因此对实部和虚部的输出分别计算距离损失并叠加形成总的距离损失$\mathcal{L}_{\text{distance}}$：

$$\mathcal{L}_{\text{distance}} = \sum_{l=1}^{L}\sum_{t=1}^{T}\sum_{f=1}^{F}\Big[\left(O_{l,t,f}^{\text{tea\_real}} - O_{l,t,f}^{\text{stu\_real}}\right)^2 + \left(O_{l,t,f}^{\text{tea\_imag}} - O_{l,t,f}^{\text{stu\_imag}}\right)^2\Big], \tag{9}$$

其中，$O_{l,t,f}^{\text{stu_real}}$和$O_{l,t,f}^{\text{stu_imag}}$表示学生模型实部和虚部的输出，$O_{l,t,f}^{\text{tea_real}}$和$O_{l,t,f}^{\text{tea_real}}$表示教师模型实部和虚部的输出，$L$为复LSTM模块总数，$T$为输入语声总帧数，$F$为特征维度。对每一帧的输出单独处理而不预先进行压平，因为希望每一帧的数据对于知识转移有独特的贡献。最终师生学习通过原模型损失和师生距离损失的联合优化进行。

2 实验设置

2.1 实验数据

本文分别选择在小型公开数据集Voice Bank Demand[10]和大型公开数据集DNS Challenge[11]上进行对比实验。

在Voice Bank Demand数据集中，干净的语声数据来自Voice Bank语料库中的30名说话人，其中28人包含在训练集中，2人包含在测试集中。每个说话人提供约400句话。对于训练集，将10种噪声(babble、cafeteria、car、kitchen、meeting、metro、restaurant、ssn、station、traffic)随机与干净语声剪辑在4种信噪比(15 dB、10 dB、5 dB和0 dB)下叠加生成11572条带噪-干净语声对。因此，总共考虑了40种不同的噪声条件。而测试集的建立则使用了Demand数据库中剩余的5种噪声(bus、cafe、living、psquare、office)和4种不同的信噪比设置(17.5 dB，12.5 dB，7.5 dB和2.5 dB)，这使得测试集共有20种不同的组合。注意到，由于测试集和训练集使用了不同的说话人和噪声环境，二者是互不交叉的。

DNS Challenge数据集包含来自2150个说话人的500 h干净语料和总计约180 h的65000条噪声剪辑。随机切分语料库成训练集和验证集各60000条和1000条语声。训练集和验证集中的带噪语声是通过从语声集和噪声集中随机选择片段，并在$-5 \sim 15$ dB之间的随机信噪比下进行混合来生成的。总计使用了100 h的语声数据用于训练和验证。测试集使用DNS Challenge官方提供的无混响测试集进行客观评分的比较。

所有的语声数据使用16 kHz采样。使用32 ms的汉宁窗并设置帧移为50%。STFT点数设置为512点，输入复频谱特征为257维。使用Adam优化器对网络进行训练，学习率设置为0.0006，总训练轮数(epoch)为20，批处理大小为16，在一个小批次中，所有样本被零填充以具有和最长样本相同的时间步。

2.2 评价指标

为了评估各模型的增强效果，选择以下客观语声评估指标用于性能的评判。WB-PESQ：ITU-T P.862.2推荐的语声质量感知评估方法，本文选用它的宽带版本[12]；STOI：语声短时客观可懂度评估方法[13]，其得分范围为0 ∼ 1之间，越高的得分意味着越好的语声可懂度；CSIG：信号失真的平均意见得分(Mean opinion score, MOS)预测[14]；CBAK：背景噪声干扰侵入下的MOS评分预测[14]；COVL：整体语声质量的MOS评分预测[14]。

3 实验结果评估与分析结论

3.1 Voice Bank Demand 数据集实验结果分析

为了评估所提算法的性能，选择了一些在Voice Bank Demand数据集上公开结果的算法进行性能上的对比，包括基于先验信噪比估计的维纳滤波算法[15]，基于时域U-Net结构的生成对抗网络(Speech Enhancement Generative Adversarial Network)[16]，使用深度特征损失(Deep feature loss)训练的时域膨胀卷积网络[17]，用于声频源分离的端到端算法Wave-U-Net[18]，基于离散余弦变换的DNN语声增强方法[19]，以及利用多种客观评价指标训练生成对抗网络生成器的算法MetricGAN[20]。而本文用于师生学习的教师模型和学生模型单独训练的版本分别为DCCRN-T和DCCRN-S。其中，DCCRN-O-S 模型由文献[9] 中的SISNR损失进行训练所得，而其余DCCRN模型均采用MRSTFT损失进行训练。

表2中展示了本文算法与其他算法的客观指标对比。为了比较所提算法与现有算法在计算复杂度上的差异，表2中给出了算法因果性(Cau.)和模型参数量(M单位为百万，K单位为千)的说明。其中，模型的因果性决定了其是否能够进行实时的部署。而对于因果的模型，其参数量的大小反映了模型的空间复杂度。注意到，采用MRSTFT损失训练的学生模型相比于使用SISNR的DCCRN-O-S模型仅在CBAK一项指标上略有降低，其余指标均有一定提升，这说明MRSTFT损失能够利用多个尺度的频域信息，更好地引导学生模型优化。而本文所提算法DCCRN-TS基于师生学习的方法，利用预训练的教师模型知识引导学生模型取得更好的增强效果。相比学生模型DCCRN-S，所提出的DCCRN-TS模型在各项指标上均有提升，并显著缩短了与教师模型间的差距。值得关注的是，本文所提出的师生学习方法并不会在模型的推断阶段增加

表2　所提出的模型与其他算法在Voice Bank Demand数据集上的客观语声质量评估结果

Table 2　Results of objective speech quality evaluation of the proposed model with other algorithms on Voice Bank Demand dataset. The "–" indicates that the data is not given in the original text

模型	Cau.	Param.(M)	PESQ	STOI	CSIG	CBAK	COVL
Noisy	–	–	1.97	0.92	3.35	2.44	2.63
Wiener	✓	5.07 (K)	2.22	–	3.23	2.68	2.67
SEGAN	×	97.47	2.16	0.93	3.48	2.94	2.80
DFL	×	–	–	–	3.86	3.33	3.22
Wave-U-Net	×	10.00	2.40	–	3.52	3.24	2.96
DCT	✓	3.45	2.7	–	3.90	3.29	3.29
MetricGAN	×	–	2.86	0.94	3.99	3.18	3.42
DCCRN-T	✓	2.81	2.87	0.94	4.276	3.275	3.586
DCCRN-O-S	✓	0.23	2.70	0.93	3.709	3.302	3.187
DCCRN-S	✓	0.23	2.74	0.94	4.175	3.295	3.465
DCCRN-TS	✓	0.23	2.79	0.94	4.205	3.302	3.510

额外的负担，因为知识的转移过程完全在训练阶段进行，用于推断的基础模型并未改变。因此，本文算法能够无负担地提升学生模型的增强效果。在表 2 中可以观察到，除了 MetricGAN 模型外，本文所提算法相比其他算法在各项指标上均具有优势地位。MetricGAN 模型由于其结构是非因果的，并不适用于实时应用。而本文模型仅具有 0.23 M 的参数量，且结构为因果的，利于现实场景下的实时应用部署。为了进一步证明算法在延时方面的改进，分别用 I7 8700 型号的 CPU 测量了学生模型和教师模型处理 6.25 ms 一帧数据所花费的时间，教师模型单帧执行时间为 3.38 ms，而学生模型则是 2.02 ms，这说明所提师生学习方法在参数量级和延时方面实现了对教师模型的压缩，同时维持了较好的增强效果。

图 2 中展示了测试集中一条带噪语声的处理效果，可以观察到经过师生学习的引导，学生模型对于噪声的滤除更加干净。

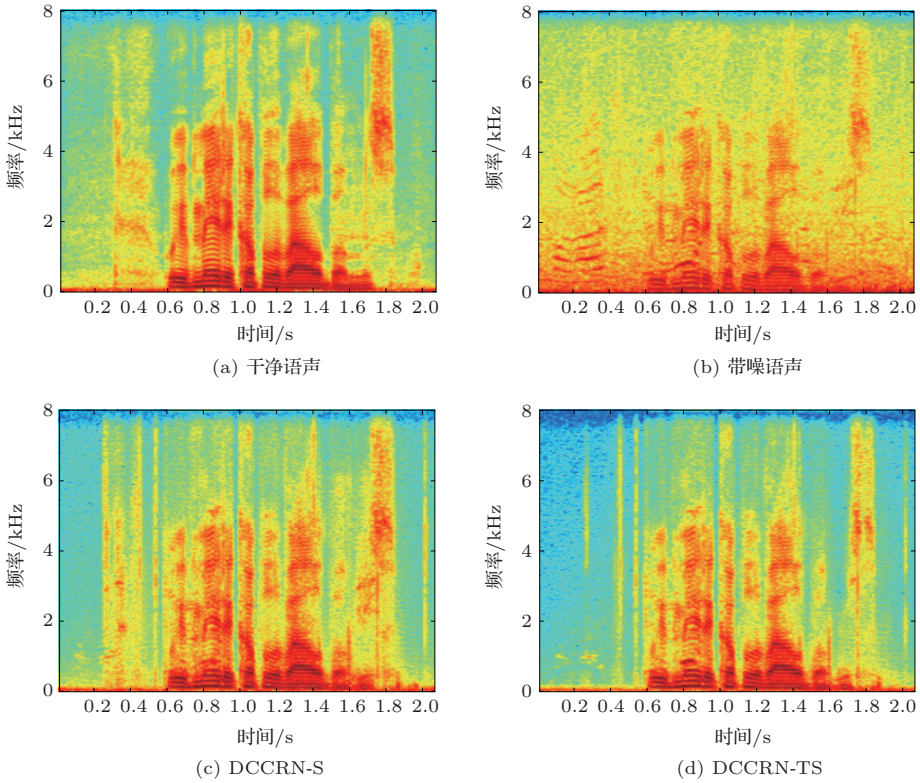

图 2　测试集语声 p232_170 的语谱图结果

Fig. 2　Speech spectrogram results for test set speech p232_170

3.2　DNS Challenge 数据集实验结果分析

为了更好地证明所提算法的性能，在更大规模的数据集 DNS Challenge 上进行了对比实验。由于本文算法着眼于降低模型的参数量，选择因果低复杂度的模型 NSNet [21]、RNNoise [22] 和本文算法进行对比。各算法的评估指标对比如表 3 所示。

从实验结果看，采用 MRSTFT 损失的 DCCRN-S 相比原版的基线模型在 PESQ 指标上有略微提升，而在此基础上进行师生学习引导优化的 DCCRN-TS 模型取得了进一步的提升，这证明在大规模数据集 DNS Challenge 上本文所提出的师生学习方法仍然有效。但注意到经过师生学习后，与教师模型的

表 3　所提出的模型与其他算法在 DNS Challenge 数据集上的客观语声质量评估结果

Table 3　Results of objective speech quality evaluation of the proposed model with other algorithms on DNS Challenge dataset

模型	Cau.	Param.(M)	PESQ	STOI
Noisy	–	–	1.58	0.91
NSNet	✓	1.30	2.15	0.94
RNNoise	✓	0.06	1.97	0.93
DCCRN-T	✓	2.81	2.70	0.96
DCCRN-O-S	✓	0.23	2.39	0.94
DCCRN-S	✓	0.23	2.40	0.94
DCCRN-TS	✓	0.23	2.44	0.94

差距仍比较大，这是由于师生模型原本的差距相比小数据集扩大了，使得师生的引导相对困难。而与同样低复杂度的实时算法NSNet和RNNoise相比，本文所提出的模型在维持低参数量的同时取得了更好的指标结果。

4 结论

在本文中，针对现有基于深度学习的语声增强模型参数规模大、计算复杂度高的问题，基于DCCRN结构构建了师生学习框架。对复LSTM模块的实部和虚部特征流分别计算帧级特征损失以拉近教师和学生模型的距离。同时，以MRSTFT损失作为学生模型的基础损失以提升学生模型的增强效果。实验结果表明，相对于基线的学生模型，所提方法在各项指标上均有优势。通过师生学习引导训练的学生模型能在低参数量下取得与大规模模型相近的性能，在公开数据集上取得了具有竞争力的结果。未来将继续研究师生学习在各种网络结构上的应用。

参 考 文 献

[1] Benesty J, Makino S, Chen J. Speech enhancement[M]. Germany: Springer Science & Business Media, 2006.

[2] Loizou P C. Speech enhancement: theory and practice[M]. America: Chemical Rubber Company Press, 2017.

[3] Xu Y, Du J, Dai L R, 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.

[4] Han S, Mao H, Dally W J. Deep compression: compressing deep neural network with pruning, trained quantization and huffman coding[J]. arXiv Preprint, arXiv: 1510.00149.

[5] LeCun Y, Denker J S, Solla S A. Optimal brain damage[C]// Advances in Neural Information Processing Systems, 1990: 598–605.

[6] Cheng Y, Wang D, Zhou P, et al. A survey of model compression and acceleration for deep neural networks[J]. arXiv Preprint, arXiv: 1710.09282.

[7] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network[J]. arXiv Preprint, arXiv: 1503.02531.

[8] Yamamoto R, Song E, Kim J M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[C]// ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 6199–6203.

[9] Hu Y, Liu Y, Lyu, S, et al. DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement[C]// Proc. Interspeech 2020: 2472–2476.

[10] Veaux C, Yamagishi J, King S. The voice bank corpus: design, collection and data analysis of a large regional accent speech database[C]// 2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013: 1–4.

[11] Reddy C K A, Gopal V, Cutler R, et al. The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective testing framework, and challenge results[C]// Proc. Interspeech 2020: 2492–2496.

[12] ITU R I T U T P. 862.2: wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs[Z]. ITU-Telecommunication Standardization Sector, 2007.

[13] Taal C H, Hendriks R C, Heusdens R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125–2136.

[14] Hu Y, Loizou P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2008, 16(1): 229–238.

[15] Scalart P, Filho J V. Speech enhancement based on a priori signal to noise estimation[C]// 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, 1996, 2: 629–632.

[16] Pascual S, Bonafonte A, Serrà J. SEGAN: speech enhancement generative adversarial network[C]// Proc. Interspeech 2017: 3642–3646.

[17] Germain F G, Chen Q, Koltun V. Speech denoising with deep feature losses[J]. arXiv Preprint, arXiv: 1806.10522.

[18] Macartney C, Weyde T. Improved speech enhancement with the wave-u-net[J]. arXiv Preprint, arXiv: 1811.11307.

[19] Geng C, Wang L. End-to-end speech enhancement based on discrete cosine transform[C]// 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2020: 379–383.

[20] Fu S W, Liao C F, Tsao Y, et al. Metricgan: generative adversarial networks based black-box metric scores optimization for speech enhancement[C]// International Conference on Machine Learning, 2019: 2031–2041.

[21] Xia Y, Braun S, Reddy C K A, et al. Weighted speech distortion losses for neural-network-based real-time speech enhancement[C]// ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 871–875.

[22] Valin J. A hybrid DSP/deep learning approach to real-time full-band speech enhancement[C]// 2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP), 2018: 1–5.