一种融合相位估计的深度卷积神经网络语音增强方法

袁文浩,梁春燕,夏 斌,孙文珠
(山东理工大学计算机科学与技术学院,山东淄博 255000)

摘 要: 在时频域的语音增强中,幅度估计和相位估计都是影响语音增强性能的重要因素.为了在基于深度学习的语音增强方法中融合对相位的估计,本文将含噪语音短时傅里叶变换(STFT)的实部和虚部特征作为两个通道输入深度卷积神经网络,通过建立一个同步估计纯净语音 STFT 的实部和虚部特征的多任务学习模型,实现了对幅度和相位的同步估计.实验结果表明,相比仅考虑幅度估计的方法,本文方法具有更好的噪声抑制能力,在低信噪比条件下,显著提高了语音增强性能.

关键词: 语音增强;相位估计;幅度估计;深度卷积神经网络
中图分类号: TN912.3 **文献标识码:** A **文章编号:** 0372-2112 (2018)10-2359-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2018.10.008

A Deep Convolutional Neural Network Based Speech Enhancement Approach Incorporating Phase Estimation

YUAN Wen-hao, LIANG Chun-yan, XIA Bin, SUN Wen-zhu
(*College of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong* 255000, *China*)

Abstract: In the speech enhancement of the time-frequency domain, both the amplitude estimation and the phase estimation are the important factors that affect speech enhancement performance. In order to incorporate the phase estimation into the speech enhancement approaches based on deep learning, the real and imaginary part of the short-time Fourier transform (STFT) of noisy speech are treated as two channels and fed into the deep convolutional neural network (DCNN) in this paper. By establishing a multi-task learning model which simultaneously estimates the real and imaginary part of the STFT of clean speech, the synchronous estimation of the amplitude and phase is achieved. Experimental results show that compared with the approaches only considering the amplitude estimation, the proposed approach has better noise suppression ability, and improves speech enhancement performance significantly under the condition of low SNR.

Key words: speech enhancement; phase estimation; amplitude estimation; deep convolutional neural network

1 引言

近年来,深度学习成为机器学习领域的研究热点,深度神经网络(Deep Neural Network, DNN)被广泛应用于语音识别、图像分类等领域中,取得了突破性进展.受 DNN 在语音识别领域的成功案例启发,研究人员开始将 DNN 应用于语音增强中[1].基于深度神经网络的语音增强方法解决了传统基于统计的语音增强方法无法处理高度非平稳噪声的问题[2,3],其本质是通过对大量纯净语音和含噪语音样本数据的学习构造含噪语音训练特征和能够表示纯净语音的训练目标之间的一个复杂非线性函数.其中,训练目标的设计是影响方法语音增强性能的关键因素,依据训练目标的不同,现有的方法可以分为两大类.

(a)基于掩蔽的方法

此类方法针对含噪语音的每个时频单元设计训练目标.文献[4]使用由时频单元的语音幅度和噪声幅度计算得到的 IBM(Ideal Binary Mask)作为训练目标,并

收稿日期:2017-10-21;修回日期:2018-02-20;责任编辑:孙瑶
基金项目:国家自然科学基金(No.61701286, No.11704229);山东省自然科学基金(No.ZR2015FL003, No.ZR2017MF047, No.ZR2017LA011)

通过训练一个二值分类 DNN 来估计 IBM,从而得到对含噪语音各时频单元是语音主导还是噪声主导的一个判断. 文献[5]采用由时频单元的语音能量和噪声能量计算得到的 IRM(Ideal Ratio Mask)代替 IBM 作为训练目标,并通过实验证明了 IRM 相比 IBM 更加适合 DNN 的训练,而且基于 DNN 的语音增强方法相比其他方法明显提高了增强语音的质量和可懂度.

(b)基于映射的方法

此类方法直接针对纯净语音设计训练目标. 徐勇等人将纯净语音的对数功率谱作为训练目标,通过训练 DNN 构造一个含噪语音对数功率谱(Logarithmic Power Spectra,LPS)与纯净语音对数功率谱之间的映射函数[6];并在文献[7]中采用 Global Variance Equalization、Dropout 和 Noise-aware 三种策略来改进 DNN 的训练,进一步提高了模型的语音增强性能. 韩伟等人将听觉掩蔽效应融合到目标语音的幅度谱估计中,通过训练 PM-DNN(Perceptual Masking Deep Neural Network)对感知增益函数和含噪语音幅度谱进行联合优化,取得了优于常见 DNN 语音增强方法的增强效果[8].

上述两种方法,无论是基于掩蔽还是基于映射的方法,最终得到的都是对纯净语音的频域幅度的估计,缺少对相位信息的估计,因此对于时域增强语音的重构需要结合含噪语音的频域相位来完成. Paliwal 等人的近期研究指出:如果能够得到相位谱的准确估计,将会显著提高增强语音的质量[9]. 受此启发,Williamson 等人提出了一种复数域的 Ratio Mask——Complex Ratio Mask,通过使用 DNN 从含噪语音的组合特征中估计 CRM,达到同时增强幅度谱和相位谱的目的[10]. 文献[11]设计了一种相位敏感的训练目标,并采用 LSTM(Long Short-Term Memory)结构的 RNN(Recurrent Neural Network)来建立语音增强模型,实验结果表明训练目标中相位信息的加入提高了模型的语音增强性能. 为了研究相位估计对基于掩蔽的方法的语音分离性能的影响,文献[12]提出了一种基于基音估计的相位估计方法,并将估计得到的相位与由掩蔽方法估计得到的幅度相结合,实验结果表明不管训练目标采用 IBM 还是 IRM,与相位估计的结合都能有效提高增强语音的质量和可懂度.

为了在基于深度神经网络的语音增强方法中更好的实现幅度和相位的同时估计,从而进一步提高语音增强性能,本文提出了一种融合相位估计的深度卷积神经网络(DCNN)语音增强方法. 与已有方法不同,本文直接针对含噪语音和纯净语音 STFT 的实部和虚部设计训练特征和训练目标,通过将含噪语音的实部特征和虚部特征作为两个通道输入深度卷积神经网络进行训练,建立一个能够同步估计纯净语音 STFT 的实部和虚部的多任务模型,从而同时得到纯净语音 STFT 的幅度和相位的估计. 本文的主要内容包括:首先分析了 DCNN 相比 DNN 在语音增强任务中的优势,并设计了适合网络训练的实部和虚部特征,然后给出了具体的网络结构和训练方法,最后通过客观实验对方法的语音增强性能进行了评价.

2 DNN 与 DCNN

DCNN 在图像分类等任务上的成功应用,证明了其在二维图像信号处理上相比 DNN 具有更好的性能[13,14]. 本文将其用于语音增强主要基于以下两方面的考虑.

(1)绝大多数的语音增强方法是通过将含噪语音信号经过短时傅里叶变换转换到时频域上进行的,在时频域的时间和频率两个维度上,语音和噪声信号都具有很强的相关性,这种相关性是语音增强的基础. 同样,在基于深度神经网络的语音增强方法中,为了充分考虑两个维度的相关性,一般采用相邻多帧的频域特征构成具有时间和频率两个维度的特征矩阵作为网络的输入,这种矩阵形式的输入在两个维度上的局部相关性与图像中相邻像素之间的相关性非常类似;矩阵元素之间的相对位置与矩阵元素的数值对于特征的表达起到同等重要的作用. 当网络结构为全连接的 DNN 时,由于其输入层只有一个维度,因此要舍弃位置结构信息将特征矩阵转换为向量作为输入;而当网络结构为 DCNN 时,则可以直接使用特征矩阵作为输入,不破坏矩阵元素之间的相对位置. 得益于 DCNN 在二维平面上的局部连接特性,使其相比 DNN 能够更好地表达网络输入在时间和频率两个维度的内在联系,因而在语音增强时能够更充分地利用语音和噪声信号的时频相关性. 另外,DCNN 通过权值共享极大减少了神经网络需要训练的参数的个数,且具有更好的泛化能力,对未训练噪声理论上应该有更好的处理性能.

(2)由于相位谱缺乏明显的结构性,对相位信息的直接估计是非常困难的,但是信号 STFT 的幅度和相位信息均可以通过信号 STFT 的实部和虚部间接得到,因此可以通过同步估计纯净语音 STFT 的实部和虚部,达到同时估计语音 STFT 的幅度和相位的目的. 而为了采用深度神经网络同步估计纯净语音 STFT 的实部和虚部,需要将含噪语音 STFT 的实部和虚部同时作为网络的输入. 在图像处理中,当网络输入为彩色图像时,可以将图像 RGB 分量的三个矩阵作为 DCNN 的三个通道. 类似的,我们可以将含噪语音 STFT 的实部特征矩阵和虚部特征矩阵作为两个通道输入

DCNN,如图 1 所示.

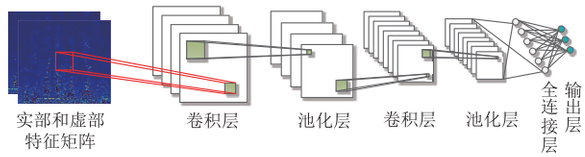

图1 DCNN结构示意图

3 训练特征与训练目标

假设 s 和 d 分别表示纯净语音和加性噪声,则含噪语音

$$y = s + d \tag{1}$$

语音增强就是在已知含噪语音 y 的条件下计算 s 的估计值 $\hat{s}$,假设 y、s 和 $\hat{s}$ 在第 n 帧的 STFT 形式分别为 $Y(n,k)$、$S(n,k)$ 和 $\hat{S}(n,k)$,其中 $k=1,2,\cdots,K$ 是频带序号,设 $S(n,k)$ 的相位为 $\varphi(n,k)$,则 $S(n,k)=|S(n,k)|\mathrm{e}^{\mathrm{j}\cdot\varphi(n,k)}$,其幅度和相位可以分别用实部 $S_r(n,k)$ 和虚部 $S_i(n,k)$ 进行表示

$$|S(n,k)| = \sqrt[2]{S_r(n,k)^2 + S_i(n,k)^2} \tag{2}$$

$$\varphi(n,k) = \tan^{-1}\frac{S_i(n,k)}{S_r(n,k)} \tag{3}$$

因此,可以通过同时估计 $S(n,k)$ 的实部和虚部来估计其幅度和相位. 那么,对第 n 帧的信号而言,STFT 域上的语音增强任务就是最小化如下的误差函数

$$Er = \sum_{k=1}^{K}\left[(\hat{S}_r(n,k) - S_r(n,k))^2 + (\hat{S}_i(n,k) - S_i(n,k))^2\right] \tag{4}$$

令 $S_r(n)$、$S_i(n)$ 和 $\hat{S}_r(n)$、$\hat{S}_i(n)$ 分别表示纯净语音第 n 帧的实部与虚部向量及它们的估计值,该误差函数可以改写为

$$Er = \|\hat{S}_r(n) - S_r(n)\|_2^2 + \|\hat{S}_i(n) - S_i(n)\|_2^2 \tag{5}$$

为了训练能够同步估计 $\hat{S}_r(n)$ 和 $\hat{S}_i(n)$ 的 DCNN,采用 $Y(n,k)$ 的实部和虚部作为网络的输入,$S(n,k)$ 的实部和虚部作为网络的输出. 由于 $Y(n,k)$ 和 $S(n,k)$ 的实部和虚部的取值区间为 $(-\infty,+\infty)$,为了适应网络模型的训练,并保证增强语音重构时的简单性,采用双曲正切函数分别对实部和虚部进行压缩,得到 TR(Tanh-compressed Real component)和 TI(Tanh-compressed Imaginary component)作为网络的输入和输出特征

$$TI_Z(n,k) = \beta\frac{1-\mathrm{e}^{-\alpha\cdot Z_i(n,k)}}{1+\mathrm{e}^{-\alpha\cdot Z_i(n,k)}} \tag{6}$$

$$TR_Z(n,k) = \beta\frac{1-\mathrm{e}^{-\alpha\cdot Z_r(n,k)}}{1+\mathrm{e}^{-\alpha\cdot Z_r(n,k)}} \tag{7}$$

其中,$\alpha=0.5$、$\beta=10$ 是通过实验得到的经验参数,Z 可以为 Y 和 S 分别代表网络的训练特征和训练目标.

基于以上的训练特征和训练目标,本文所提方法的基本思想可以描述为:通过同时训练网络参数集合 λ 和 θ 构造两个高度复杂的非线性函数 f_λ 和 f_θ,使得误差函数最小,如式(8)所示:

$$Er = \|f_\lambda(X(n)) - TR_S(n)\|_2^2 + \|f_\theta(X(n)) - TI_S(n)\|_2^2 \tag{8}$$

从而得到目标输出

$$\hat{TR}_S(n) = f_\lambda(X(n)) \tag{9}$$

$$\hat{TI}_S(n) = f_\theta(X(n)) \tag{10}$$

其中,

$$\begin{aligned}X(n) = [&TR_Y(n-N), TR_Y(n-N+1),\cdots,\\ &TR_Y(n),\cdots,TR_Y(n+N),\\ &TI_Y(n-N), TI_Y(n-N+1),\cdots,\\ &TI_Y(n),\cdots,TI_Y(n+N)]\end{aligned} \tag{11}$$

表示第 n 帧的输入特征,由以第 n 帧为中心的 $(2N+1)$ 帧的 TR_Y 向量和 $(2N+1)$ 帧的 TI_Y 向量构成,$(2N+1)$ 即为输入窗长.

DCNN 采用多任务学习模式同时训练 λ 和 θ,训练过程采用小批量梯度下降法进行,使用的代价函数定义为

$$C(\lambda,\theta) = \frac{1}{M}\sum_{n=1}^{M}\left[\|f_\lambda(X(n)) - TR_S(n)\|_2^2 + \|f_\theta(X(n)) - TI_S(n)\|_2^2\right] \tag{12}$$

其中,M 为网络训练中采用的 mini-batch 的大小.

在完成网络训练后进行语音增强时,对第 n 帧的含噪语音 y_n,首先得到训练目标的估计值 $\hat{TR}_S(n)$ 和 $\hat{TI}_S(n)$,然后使用该估计值计算 $\hat{S}(n)$ 的实部和虚部

$$\hat{S}_r(n) = -\frac{1}{\alpha}\log\left(\frac{\beta - \hat{TR}_S(n)}{\beta + \hat{TR}_S(n)}\right) \tag{13}$$

$$\hat{S}_i(n) = -\frac{1}{\alpha}\log\left(\frac{\beta - \hat{TI}_S(n)}{\beta + \hat{TI}_S(n)}\right) \tag{14}$$

最后,通过对 $\hat{S}(n)$ 进行短时傅里叶逆变换(Inverse STFT,ISTFT)得到增强语音的时域重构

$$\hat{s}_n = \mathrm{ISTFT}(\hat{S}_r(n) + j\cdot\hat{S}_i(n)) \tag{15}$$

4 网络结构

在前期研究中,我们设计了一种包含 3 个卷积层和 2 个全连接层的 DCNN 来进行语音增强,并通过实验证明了该网络结构在语音增强中的有效性. 在此基础上,依据本文所采用的训练特征和训练目标,构造如图 2 所示的同时估计纯净语音 STFT 的实部与虚部的 RI-DCNN. 具体的网络结构设计如下.

输入层　网络的输入是分别由多帧 TR_Y 向量和多

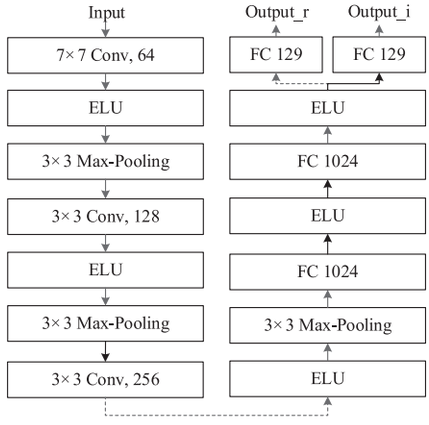

图2 RI-DCNN的结构框图

帧 TI_Y 向量构成的两个特征矩阵.

卷积层 本文所采用的网络共包含 3 个卷积层. 其中,第一层的卷积滤波器大小为 7×7,其余两层的滤波器大小为 3×3,滤波器的个数分别为 64、128、256,步长均设为 1×1.

激活函数层 网络中的激活函数均使用 ELU(Exponential Linear Units).

池化层 在卷积层和激活函数层之后是 3 个池化层,均采用最大池化(Max-Pooling),滤波器大小 3×3,步长为 2×2.

全连接层 2 个全连接层(FC,Fully Connected)的节点数量均为 1024.

输出层 网络的最后是两个 129 个节点的全连接层,分别对应实部和虚部的 129 维目标输出.

5 实验与结果分析

5.1 实验配置

实验所用的纯净语音全部来自 TIMIT 语音数据库[15],所用的噪声数据来自俄亥俄州立大学 Perception and Neurodynamics 实验室的 100 类噪声[16]和 Noisex92 噪声库[17]. 语音和噪声信号的采样频率均转换为 8kHz,短时傅里叶变换的帧长为 32ms(256 点),帧移为 16ms(128 点),相应的 TR 和 TI 特征的维度为 129. 训练集共包含 50000 段含噪语音(约 40 小时),每段含噪语音采用如下方法合成:从 TIMIT 语音库的 Training 集的 4620 段纯净语音中随机选取 1 段,并从 100 类噪声中随机选取 1 类,然后将该类噪声的随机截取片段按照 -10dB、-5dB、0dB、5dB 和 10dB 五种全局信噪比中的随机一种混入语音中.

测试集采用 TIMIT 语音库的 Core test 集的 192 段纯净语音与来自 Noisex92 噪声库的 4 类噪声合成,这 4 类噪声是与训练集噪声完全不同的未知噪声,分别是 Factory2、Buccaneer1、Destroyer engine、HF channel 噪声. 将 192 段语音分别按照 -7dB、0dB 和 7dB 的全局信噪比与 4 类噪声的随机截取片段进行混合,整个测试集共包含 2304(192×3×4)段含噪语音.

5.2 客观性能评价

为了比较不同方法的语音增强性能,本文采用多种客观指标对不同方法增强后的语音进行评价,包括:采用 PESQ(Perceptual Evaluation of Speech Quality)来评价增强语音的质量[18];采用 STOI(Short Time Objective Intelligibility)来评价增强语音的可懂度[19];采用分段信噪比(Segmental SNR,SegSNR)来评价增强语音的信噪比;采用 SDR(Signal-to-Distortion Ratio)来评价增强语音的失真程度[20]. 其中,PESQ 即语音质量感知评估是 ITU-T(国际电信联盟电信标准化部)推荐的语音质量评估指标,其得分范围为 -0.5~4.5,越高的得分表示越高的语音质量;STOI 即短时客观可懂度,则主要衡量语音的可懂度,其得分范围为 0~1,越高的得分表示语音具有越好的可懂度;分段信噪比同样是衡量语音质量的重要指标,它比全局信噪比更接近实际的语音质量,分段信噪比越大,代表主观的语音质量越好;SDR 则能够较好地反映增强语音的失真程度,SDR 越大表明增强语音的失真越小.

为了检验本文所提出的 RI-DCNN 的语音增强性能,我们将其与采用 LPS 作为输入输出特征的基线 DNN(DNN-baseline)和 LPS-DCNN 进行比较[6]. 其中,DNN-baseline 具有 5 个隐层,每个隐层有 1024 个节点,激活函数为 ELU;LPS-DCNN 的输入层是由多帧 LPS 向量构成的特征矩阵,输出层是 129 维的 LPS 向量,其余各层与 RI-DCNN 的结构保持一致. 三种网络均采用微软的 Cognitive Toolkit 进行训练[21],输入窗长均设为 15 帧,mini-batch 的大小均为 128,迭代次数均为 20. 为了使训练过程稳定,网络的输入和输出均进行了 MVN(Mean and Variance Normalization)处理.

使用三种方法对测试集含噪语音进行增强,并计算增强后语音的平均 PESQ、STOI、SegSNR 和 SDR,表 1~表 4 分别给出了在 4 类不同噪声和 3 种不同信噪比下四种指标的平均值,并给出了未处理的含噪语音的四种指标作为对比. 不同噪声条件下的最佳结果均用粗体进行了标记. 可见,在四种不同指标下,RI-DCNN 在大多数噪声条件下都取得了最好的结果,特别是在低信噪比(-7dB)的 HF channel 噪声下,DNN-baseline 和 LPS-DCNN 均出现了不同程度的失效,而 RI-DCNN 仍然取得了较好的语音增强性能,证明 RI-DCNN 具有更好的泛化能力. 另外,LPS-DCNN 相比 DNN-baseline 表现出了更好的语音增强性能,证明了 DCNN 相比 DNN 是更加适用于语音增强任务的网络结构.

表 1　三种方法的平均 PESQ 得分

噪声类型	信噪比(dB)	含噪语音	DNN-baseline	LPS-DCNN	RI-DCNN
Factory2	-7	1.62	2.09	**2.19**	2.15
	0	2.08	2.63	**2.74**	2.63
	7	2.54	3.04	**3.15**	2.91
Buccaneer1	-7	1.29	1.60	1.61	**1.68**
	0	1.63	2.16	2.14	**2.23**
	7	2.08	**2.65**	2.64	2.62
Destroyer engine	-7	1.49	1.07	1.82	**1.96**
	0	1.81	1.84	**2.38**	2.37
	7	2.21	2.55	**2.85**	2.75
HF channel	-7	1.30	0.86	1.20	**1.74**
	0	1.57	1.60	1.81	**2.29**
	7	1.97	2.22	2.43	**2.68**

表 2　三种方法的平均 STOI 得分

噪声类型	信噪比(dB)	含噪语音	DNN-baseline	LPS-DCNN	RI-DCNN
Factory2	-7	0.61	0.70	0.72	**0.73**
	0	0.76	0.84	**0.86**	0.85
	7	0.87	0.91	**0.92**	0.91
Buccaneer1	-7	0.48	0.56	0.57	**0.59**
	0	0.63	0.74	0.75	**0.76**
	7	0.79	0.86	**0.87**	0.86
Destroyer engine	-7	0.53	0.44	0.58	**0.63**
	0	0.69	0.70	0.79	**0.80**
	7	0.84	0.86	**0.90**	0.89
HF channel	-7	0.52	0.46	0.34	**0.63**
	0	0.69	0.71	0.60	**0.80**
	7	0.84	0.85	0.84	**0.88**

表 3　三种方法的平均 SegSNR(dB)

噪声类型	信噪比(dB)	含噪语音	DNN-baseline	LPS-DCNN	RI-DCNN
Factory2	-7	-7.69	-1.38	-0.68	**-0.04**
	0	-4.49	0.36	**1.00**	0.94
	7	-0.24	1.77	**2.16**	1.72
Buccaneer1	-7	-7.92	-2.18	-1.83	**-0.57**
	0	-4.91	-0.61	-0.09	**0.46**
	7	-0.77	0.87	1.29	**1.34**
Destroyer engine	-7	-7.91	**-1.46**	-2.41	-1.99
	0	-4.84	-0.33	-0.36	**-0.01**
	7	-0.67	1.02	1.20	**1.27**
HF channel	-7	-7.96	-1.19	-4.35	**-0.03**
	0	-4.91	-0.52	-1.63	**1.10**
	7	-0.81	0.77	0.89	**1.88**

表 4　三种方法的平均 SDR(dB)

噪声类型	信噪比(dB)	含噪语音	DNN-baseline	LPS-DCNN	RI-DCNN
Factory2	-7	-6.42	4.71	4.91	**6.34**
	0	0.44	9.25	9.46	**9.65**
	7	7.40	11.75	**11.97**	11.16
Buccaneer1	-7	-6.39	2.13	2.20	**4.15**
	0	0.20	7.01	7.16	**8.05**
	7	7.12	10.55	**10.61**	10.48
Destroyer engine	-7	-6.39	-1.60	0.46	**2.27**
	0	0.27	5.38	7.02	**7.85**
	7	7.23	10.07	**10.79**	10.60
HF channel	-7	-6.44	-0.79	-12.45	**4.80**
	0	0.23	4.66	-4.25	**8.65**
	7	7.10	9.40	7.46	**10.66**

为了验证相位信息对语音增强性能的影响,我们将 DNN-baseline 和 LPS-DCNN 估计得到的幅度谱分别与纯净语音的真实相位进行合成,然后分别计算由此得到的增强语音的平均 PESQ、STOI、SegSNR 和 SDR,并定义两种方法下真实相位相比含噪语音相位带来的 PESQ 提升分别为 $\Delta PESQ_{Baseline}$ 和 $\Delta PESQ_{LPS}$,STOI 提升为 $\Delta STOI_{Baseline}$ 和 $\Delta STOI_{LPS}$,SegSNR 提升为 $\Delta SegSNR_{Baseline}$ 和 $\Delta SegSNR_{LPS}$,SDR 提升为 $\Delta SDR_{Baseline}$ 和 ΔSDR_{LPS}. 另外,为了反映融合相位信息的 RI-DCNN 与单纯估计幅度谱的 LPS-DCNN 在语音增强性能上的差异,定义 RI-DCNN 相比 LPS-DCNN 的 PESQ 提升分别为 $\Delta PESQ_{RI}$,STOI 提升为 $\Delta STOI_{RI}$,SegSNR 提升为 $\Delta SegSNR_{RI}$,SDR 提升为 ΔSDR_{RI}. 图 3 分别给出了不同信噪比下三种指标的上述四种不同提升值.

可见,在四种指标中,RI-DCNN 相比 LPS-DCNN 带来的提升都是随着信噪比的升高而减小;同时,在 STOI、SegSNR 和 SDR 三种指标中,无论幅度谱的估计方法是 DNN-baseline 还是 LPS-DCNN,真实相位相比含噪语音相位带来的提升同样是随着信噪比的升高而减小. 分析其原因,是因为信噪比越高,含噪语音的相位越接近于纯净语音的真实相位,直接使用含噪语音相位作为增强语音相位带来的误差也就越小;反之,信噪比越低,含噪语音与纯净语音的相位差越大,由相位信息带来的误差也就越大. 本文提出的 RI-DCNN 通过融合对相位信息的估计,减小了低信噪比下由相位误差带来的增强语音重构误差,有效提高了低信噪比下的语音增强性能.

为了更加直观的比较 RI-DCNN 与其他方法的语音增强性能,我们分别采用三种方法对一段含有 Destroyer engine 噪声信噪比为 0dB 的含噪语音进行语音增强,然

后对增强语音的语谱图进行比较. 图 4(*a*)和图 4(*b*)分别给出了含噪语音与其相应的纯净语音的语谱图,图 4(*c*)~(*e*)则分别给出了采用 DNN-baseline、LPS-DCNN 以及 RI-DCNN 增强后语音的语谱图. 可以看到,采用 RI-DCNN 增强后的语音残留噪声成分最少,语音的纯净度最高,语谱图与纯净语音的语谱图最接近. 另外,LPS-DCNN 相比 DNN-baseline 在语谱图上表现出了更好的噪声抑制能力.

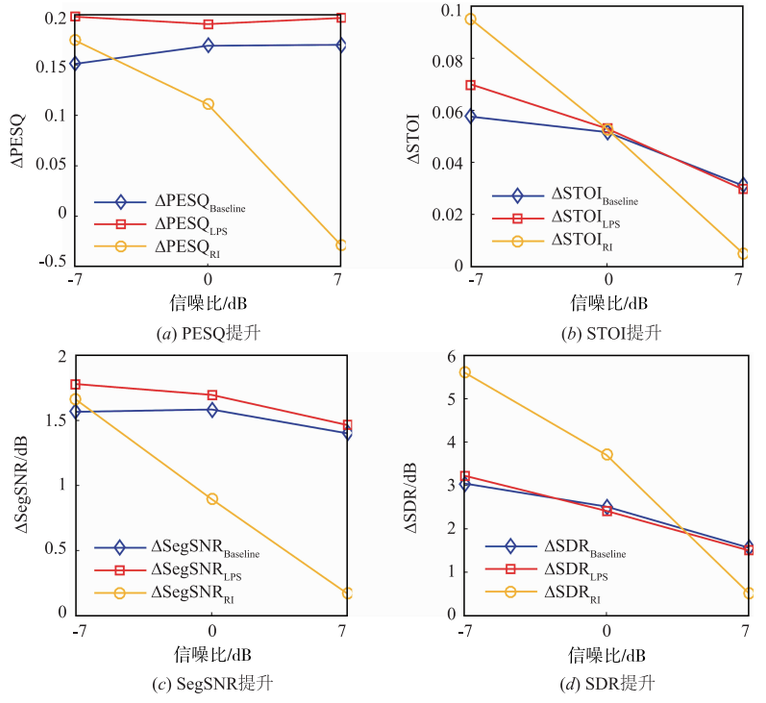

图3 不同信噪比下的提升值

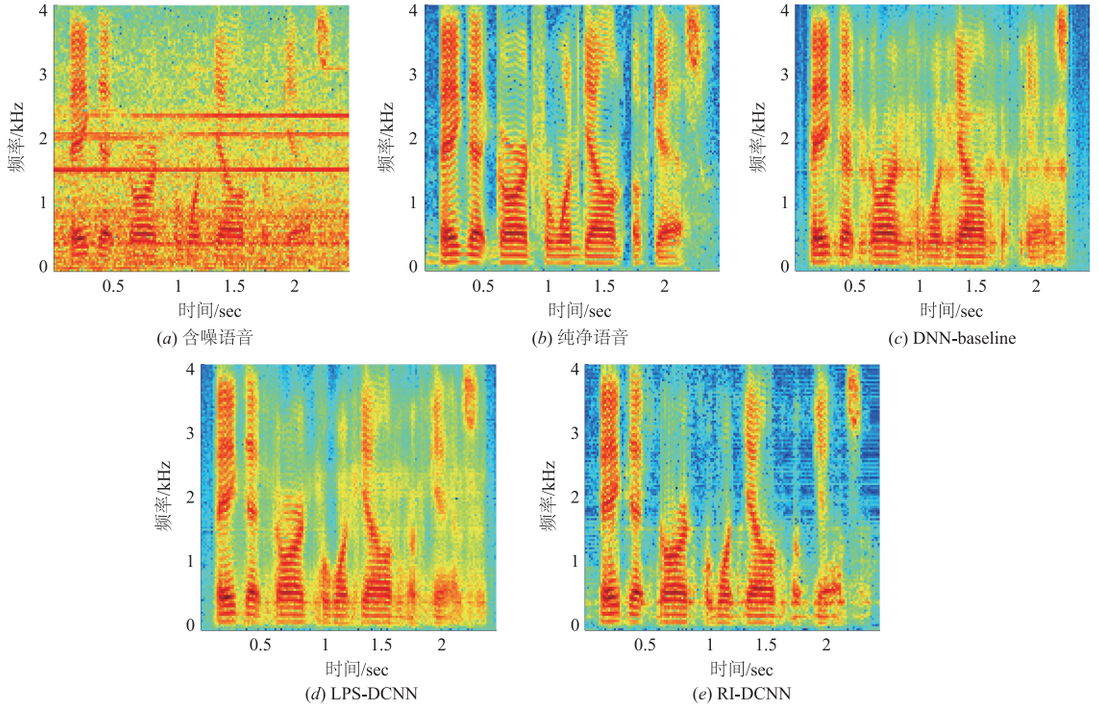

图4 0dB的Destroyer engine噪声下的增强语音语谱图示例

6 总结

相位信息在语音增强中的重要性,越来越受到研究人员的关注. 考虑到纯净语音 STFT 的幅度和相位均可以通过其实部和虚部转换得到,通过针对含噪语音和纯净语音 STFT 的实部和虚部设计适合模型训练的 TR 和 TI 特征,并采用能够更好表达特征二维相关性的深度卷积神经网络进行建模,本文提出了一种融合相位估计的深度卷积神经网络语音增强方法,借由对纯净语音 STFT 的实部和虚部的同步估计,实现了对纯净语音 STFT 的幅度和相位的同步估计. 实验结果表明,与单纯进行幅度估计的 DNN 或 DCNN 相比,本文所提方法能够更好的抑制背景噪声,有效减小了低信噪比条件下由相位误差带来增强语音重构误差,显著提高了低信噪比下的语音增强性能.

参考文献

[1] 刘文举,聂帅,梁山,等. 基于深度学习语音分离技术的研究现状与进展[J]. 自动化学报,2016,42(6):819-833.
LIU Wen-Ju, NIE Shuai, LIANG Shan, ZHANG Xue-Liang. Deep learning based speech separation technology and its developments[J]. Acta Automatica Sinica, 2016, 42(6):819-833. (in Chinese)

[2] 孟宪波,鲍长春. 基于最小控制 GARCH 模型的噪声估计算法[J]. 电子学报,2016,44(3):747-752.
MENG Xian-bo, BAO Chang-chun. Noise estimate algorithm based on minima controlled GARCH model[J]. Acta Electronica Sinica,2016,44(3):747-752. (in Chinese)

[3] 何玉文,鲍长春,夏丙寅,等. 基于 AR-HMM 在线能量调整的语音增强方法[J]. 电子学报,2014,42(10):1991-1997.
HE Yu-wen,BAO Chang-chun,XIA Bing-yin,et al. Online energy adjustment using AR-HMM for speech enhancement[J]. Acta Electronica Sinica,2014,42(10):1991-1997. (in Chinese)

[4] WANG Y,WANG D L. Towards scaling up classification-based speech separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1381-1390.

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

[6] XU Y, DU J, DAI L R, et al. An experimental study on speech enhancement based on deep neural networks[J]. IEEE Signal Processing Letters,2014,21(1):65-68.

[7] 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.

[8] 韩伟,张雄伟,闵刚,等. 基于感知掩蔽深度神经网络的单通道语音增强方法[J]. 自动化学报,2017,43(2):248-258.
HAN Wei,ZHANG Xiong-Wei,MIN Gang,et al. A single-channel speech enhancement approach based on perceptual masking deep neural network[J]. Acta Automatica Sinica, 2017,43(2):248-258. (in Chinese)

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

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

[11] WENINGER F, ERDOGAN H, WATANABE S, et al. Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR[A]. Proceedings of International Conference on Latent Variable Analysis and Signal Separation[C]. Liberec:Springer International Publishing,2015. 91-99.

[12] MAYER F, WILLIAMSON D S, MOWLAEE P, et al. Impact of phase estimation on single-channel speech separation based on time-frequency masking[J]. The Journal of the Acoustical Society of America,2017,141(6):4668-4679.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[A]. Proceedings of the International Conference on Neural Information Processing Systems[C]. Nevada: Curran Associates Inc,2012. 1097-1105.

[14] 柯圣财,赵永威,李弼程,彭天强. 基于卷积神经网络和监督核哈希的图像检索方法[J]. 电子学报,2017,45(1):157-163.
KE Sheng-cai, ZHAO Yong-wei, LI Bi-cheng, PENG Tian-qiang. Image retrieval based on convolutional neural network and kernel-based supervised hashing[J]. Acta Electronica Sinica,2017,45(1):157-163. (in Chinese)

[15] GAROFOLO J S, LAMEL L F, FISHER W M, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus[CD]. Philadelphia:Linguistic Data Consortium,1993.

[16] HU G. "100 Nonspeech Environmental Sounds, 2004"[OL]. http://web. cse. ohio-state. edu/pnl/corpus/HuNonspeech/HuCorpus. html,2004.

[17] VARGA A, STEENEKEN H J M. Assessment for automatic speech recognition:II. NOISEX-92:A database and an experiment to study the effect of additive noise on

speech recognition systems[J]. Speech Communication, 1993,12(3):247 – 251.
[18] 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[A]. Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C]. Utah: IEEE, 2001. 749 – 752.
[19] 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, and Language Processing, 2011, 19(7): 2125 – 2136.
[20] VINCENT E, GRIBONVAL R, FEVOTTE C. Performance measurement in blind audio source separation[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2006, 14(4): 1462 – 1469.
[21] YU D, EVERSOLE A, SELTZER M, et al. An Introduction to Computational Networks and the Computational Network Toolkit[R]. Tech. Rep. MSR, Microsoft Research, 2014.

作者简介

袁文浩(通信作者) 男. 1985 年出生, 山东寿光人. 2013 年毕业于华东理工大学获博士学位, 现为山东理工大学计算机科学与技术学院讲师. 主要研究方向为语音信号处理, 语音增强.
E-mail: why_sdut@126.com

梁春燕 女. 1986 年出生, 山东淄博人. 2014 年毕业于中国科学院声学研究所获博士学位, 现为山东理工大学计算机科学与技术学院讲师. 主要研究方向为语音信号处理, 说话人识别.
E-mail: liangchunyan@sdut.edu.cn