文章编号：1001-3806(2019)01-0109-06

DRNN在激光多普勒测振仪测声系统中的应用

白 涛[1,2]，吴 谨[1*]，李明磊[1,2]，万 磊[1,2]，李丹阳[1,2]
（1. 中国科学院 电子学研究所，北京 100190；2. 中国科学院大学，北京 100049）

摘要：为了降低激光多普勒测振仪在测声过程中给语音信号中引入的噪声，采用深度循环神经网络语音信号去噪的方法，对从激光多普勒测声系统采集回来的语音信号做降噪处理，并进行了理论分析和实验验证。结果表明，利用层数为1层～3层、每层神经元个数为1024的深度循环神经网络，对－6dB～6dB信噪比的语音信号进行处理，随着层数的增加，语音信号的质量在多项评价指标上达到8dB～12dB的提升；深度循环神经网络可以有效对激光多普勒测声系统采集的语音信号进行降噪处理。该研究对提升语音信号的质量有着实际意义。

关键词：激光技术；激光多普勒测振仪；语音信号去噪；深度循环神经网络
中图分类号：TN912；TN247 **文献标志码**：A **doi**：10.7510/jgjs.issn.1001-3806.2019.01.022

Application of DRNN in voice measurement system of laser Doppler vibrometer

BAI Tao[1,2], *WU Jin*[1], *LI Minglei*[1,2], *WAN Lei*[1,2], *LI Danyang*[1,2]
(1. Institue of Electronics, Chinese Academy of Sciences, Beijing 100190, China; 2. Graduate University of Chinese Academy of Sciences, Beijing 100040, China)

Abstract: In order to reduce the noise introduced to speech signal by a laser Doppler vibrometer during the measurement of sound, the method of deep recurrent neural network(DRNN) speech signal denoising was adopted. The speech signal collected from laser Doppler measurement system was denoised. By using the deep recurrent neural network with 1 layer～3 layers and 1024 neurons per layer, the speech signals with signal-to-noise ratio from －6dB to 6dB were processed. After theoretical analysis and experimental verification, the results show that, as the number of layers increases, the quality of speech signals has risen to 8dB～12dB in many evaluation indexes. DRNN can effectively denoise the speech signals collected by laser Doppler acoustic system. The research is of practical significance for improving the quality of speech signals.

Key words: laser technique; laser Doppler vibrometer; speech signal denoising; deep recurrent neural network

引 言

激光多普勒测振仪(laser Doppler vibrometer, LDV)是利用光学多普勒效应来检测物体振动的设备，它可以实现对振动目标非接触、高灵敏度和远作用距离的测量[1]。与此同时，声音信号源附近的物体通过空气耦合，会随着信号源的振动而振动，利用LDV探测说话人附近的物体的振动情况，可以远距离的还原说话人的声音。但是，通过LDV获取的语音信号会被各种各样的噪声所污染，比如激光束照射在粗糙物体表面引起的散斑效应[2]、暗电流噪声和背景噪声等。人们将LDV内部元器件引入的噪声称作LDV的系统噪声，而将LDV的激光束打在振动目标上，最后接收回来的信号中的噪声称之为LDV测声系统的系统噪声，它是在测声过程中各种引入噪声的一个综合作用的结果。本文中主要针对去除LDV测声系统的系统噪声进行了研究，提出针对LDV测声系统的语音信号去噪手段，对提升LDV系统的测声性能，降低从LDV获取的语音信号的噪声，有着重大的意义。

传统的单声道无监督语音信号去噪算法要求噪声比较平稳，以便在非语言段对噪声进行估计，再依据估计出来的噪声对带噪语音段进行处理[3]。但在实际情况中，噪声具有随机性和突变性，使得对噪声的跟踪和估计变得困难。同时，传统的语音增强方法易引入非线性失真[4]。近年来，深度神经网络(deep neural network, DNN)在语音信号处理中有着很多成功的应

作者简介：白 涛(1992-)，男，硕士研究生，现主要从事语音信号处理的研究。
*通讯联系人。E-mail：jwu909@263.net
收稿日期：2018-03-15；收到修改稿日期：2018-04-13

用,并且适应性好,限制条件少。基于大数据的训练,DNN 可以充分学习噪音和干净语音之间的复杂的非线性关系,它能记住一些噪声模式,因而可以很好地抑制一些非平稳噪声[5]。但参考文献[5]中同时提出,如果将 DNN 网络训练的帧数增加,随着上下文的帧数越来越多,DNN 网络也难以处理。循环神经网络(recurrent neural network,RNN)可以看作是一个有无限层的 DNN,但 RNN 缺少层次信息。为了弥补 DNN 和 RNN 网络结构的不足,HUANG 等人提出了深度循环神经网络(deep recurrent neural network,DRNN)和掩蔽联合优化的语音信号去噪网络结构[6]。在此基础上。HAN 等人提出了联合优化神经网络和约束维纳滤波的语音增强方法[7]。YAO 等人提出了谱减法结合神经网络的语音增强[8]。

针对 LDV 测声系统的语音降噪算法的研究较为初步,LI 等人提出了利用高斯带通滤波器和维纳滤波处理 LDV 获取的语音信号[9]。LÜ 等人提出了基于最小控制递归平均算法估计噪声的维纳滤波抑制噪声[10]。QU 等人提出了一种改进的小波阈值算法应用于 LDV 测声系统[11]。但前述的 3 种方法为通用的语音信号降噪手段,并没有考虑 LDV 测声系统所固有的噪声特性。而 LDV 测声系统的噪声模式较为单一,利用深度神经网络可以很好地提取 LDV 测声系统的噪声模式,并把语音信号从带噪语音中分离开来,从而达到降噪的目的。

本文中研究了利用深度循环网络结构提取 LDV 测声系统的系统噪声特征,同时利用训练好的深度循环神经网络结构,对从 LDV 系统采集回来的语音信号做降噪处理。实验结果表明,DRNN 网络可以有效地对从 LDV 系统采集回来的语音信号做降噪处理,与此同时,可以很好地保留语音信号原有的信息。

1 算法模型

语音信号和 LDV 系统噪声之间的相互作用非常的复杂,但加性噪声是影响听感的主要因素[12],所以在这里将模型简化为加性噪声:

$$y(t) = x(t) + n(t) \tag{1}$$

式中,$y(t)$表示 t 时刻从 LDV 采集的带噪的语音信号,$x(t)$表示 t 时刻播放的语音信号,$n(t)$表示由 LDV 系统在 t 时刻添加的语音信号噪声。

1.1 特征提取

对(1)式两边做离散傅里叶变换,可以表示为:

$$Y(\omega) = X(\omega) + N(\omega) \tag{2}$$

式中,$X(\omega)$表示干净的语音信号,$N(\omega)$和 LDV 测声系统噪声的离散傅里叶变换,ω 表示信号的频率,$Y(\omega)$表示$X(\omega)$和 $N(\omega)$的和。

首选,对信号进行分帧处理,然后计算每帧信号的离散傅里叶变换(discrete Fourier transform,DFT)系数:

$$Y(\omega) = \sum_{t}^{T-1} y(t)h(t)\exp\left(-\mathrm{j}\frac{2\pi \mathrm{d}t}{T}\right),\quad (\omega = 0,1,\cdots,T-1) \tag{3}$$

式中,j 表示虚数单位,$h(t)$表示窗函数,t 表示时间,T 表示周期。

可以用如下公式定义对数功率谱:

$$Y(\omega) = \ln|Y(\omega)|^2,(\omega = 0,1,\cdots,W-1) \tag{4}$$

式中,$W = T/2 + 1$。作者取语音信号的对数功率谱 $X(\omega)$和从 LDV 获取的噪声的对数功率谱 $N(\omega)$,将它们的和 $Y(\omega)$作为神经网络的输入,假设从神经网络预测出的语音信号的对数功率谱为$\hat{X}(\omega)$,由于人耳对相位的微小变化不敏感[13],所以可通过带噪声的语音信号的相位重构去噪后的语音信号 $\hat{x}(t)$:

$$\hat{X}(\omega) = \exp\left\{\frac{\hat{X}(\omega)}{2}\right\}\exp\{\mathrm{j}\angle Y(\omega)\} \tag{5}$$

式中,∠表示求角度。

通过逆傅里叶变换可以得到时域信号 $\hat{x}(t)$:

$$\hat{x}(t) = \frac{1}{T}\sum_{k=0}^{L-1}\hat{X}(k)\exp\left(\mathrm{j}\frac{2\pi kt}{T}\right) \tag{6}$$

式中,k 表示频率。同理可以得到神经网络模型估计得到的系统噪声 $\hat{n}(t)$。

1.2 DRNN 网络结构

RNN 网络结构可以用图 1 表示。

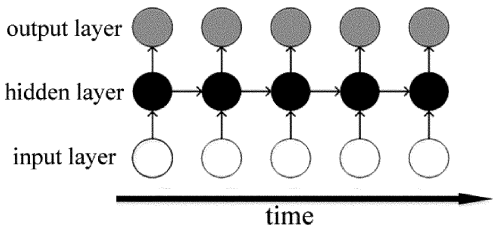

Fig. 1 RNN structure

假设循环神经网络的输入为 x_t,输出为 y_t,隐含层状态为 h_t。则系统可以定义为:

$$h_t = f_h(x_t,h_{t-1}) \tag{7}$$

$$y_t = f_o(h_t) \tag{8}$$

式中,$f_h(\,)$是输入层的激活函数,$f_o(\,)$是输出函数。

RNN 网络参量的更新可以通过以下代价函数完成:

$$J(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T_n} D(y_t^{(n)}, f_o(h_t^{(n)})) \tag{9}$$

式中，$D(\,)$表示求 2 阶方差，上标 n 表示输出层神经元数量，$h_t^{(n)} = f_h(x_t^{(n)}, h_{t-1}^{(n)})$，$h_0^{(n)} = 0$。

而对于深度循环神经网络，网络的结构如图 2 所示。

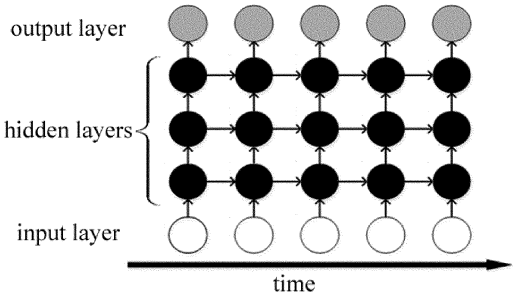

Fig. 2　DRNN structure

隐含层可以被定义为：

$$h_t^{(l)} = f_h^{(l)}(h_t^{(l-1)}, h_{t-1}^{(l)}) = \phi_h(\boldsymbol{W}_l^{\mathrm{T}} h_{t-1}^{(l)} + \boldsymbol{U}_l^{\mathrm{T}} h_t^{(l-1)}) \tag{10}$$

式中，$h_t^{(l)}$ 表示在时刻 t 第 l 层的隐含层的状态，ϕ_h 表示非线性激活函数，在这里使用线性修正单元，$\phi_h = \max(0, x)$，$\boldsymbol{W}_l$ 表示第 l 层的权重矩阵，$\boldsymbol{U}_l$ 表示第 l 层输入的权重矩阵[14]。而当最后一层的隐含层计算完成，可通过：

$$y_t = f_o(h_t, x_t) = \phi_o(\boldsymbol{V}^{\mathrm{T}} h_t) \tag{11}$$

式中，$\boldsymbol{V}^{\mathrm{T}}$ 为输出层的权重矩阵。

1.3　去噪网络结构

图 3 所示为 t 时刻 DRNN 去噪网络的示意图。输入网络的信号 x_t 是从 LDV 采集回来的噪声和语音信号混合，然后取对数谱功率谱。神经网络的输出为两部分 $\hat{y}_{1t}$ 和 $\hat{y}_{2t}$，分别表示对语音信号的预测结果和对噪声的预测结果，再通过第 1.1 中所述的重构方式即可将语音信号和系统噪声还原。其中隐含层的层数为 l，假设原信号为 y_{1t} 和 y_{2t}，通过最小均方误差进行有监督调优：

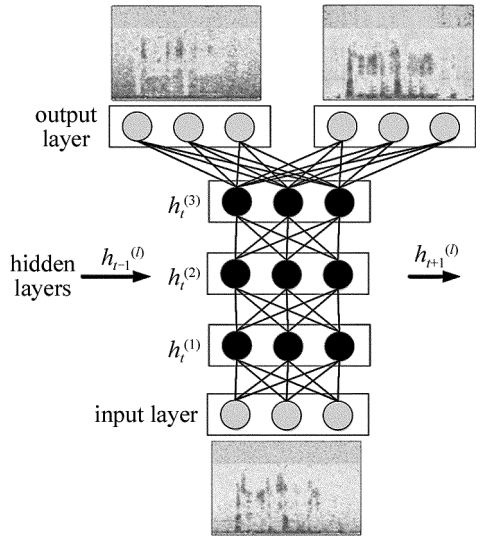

Fig. 3　Denoised network structure

$$J_{\mathrm{MSE}} = \sum_{t=1}^{T} (\| \hat{y}_{1t} - y_{1t} \|_2^2 + \| \hat{y}_{2t} - y_{2t} \|_2^2) \tag{12}$$

与此同时，引入正则化项 γ，使得预测的语音信号与干净的语音信号尽可能相似的同时，与 LDV 系统噪声差异尽可能大[6]：

$$J_{\mathrm{MSE}} = \sum_{t=1}^{T} (\| \hat{y}_{1t} - y_{1t} \|_2^2 + \| \hat{y}_{2t} - y_{2t} \|_2^2 - \gamma \| \hat{y}_{1t} - y_{2t} \|_2^2 - \gamma \| \hat{y}_{2t} - y_{1t} \|_2^2) \tag{13}$$

使用基于时间的反向传播算法（back propagation through time，BPTT）优化网络参量[15]。在测试阶段，将从 LDV 测声系统获取的语音信号的对数功率谱输入网络，然后对波形进行重构，即可获得由 DRNN 网络预测出的语音和噪声信号。

2　实验过程

2.1　实验装置

图 4 所示为 LDV 测声系统的原理图。为了使实验过程尽量贴近实际应用，振动目标采用生活中常见的纸盒。首先用扬声器驱动纸盒振动，LDV 将激光打在纸盒上，然后接收反射光。LDV 的原理如图 4 左侧所示。采用波长为 1550nm 的单频窄线宽激光器，激光器发出的激光束经过光纤耦合器（optic cable，OC）分为两束[16]：一束是探测光，首先经过环路器，然后通过光束聚焦镜（beam focusing，BF）聚焦在振动目标表面，声源通过空气使得振动目标表面产生振动，从而使得探测光在振动目标表面产生多普勒频移，再经过振动目标散射的回波由光束聚焦镜收集，通过环路器作为信号光，最后输入六端口混频器；另一束是本振光，输入六端口混频器。六端口混频器输出 4 路混频光分

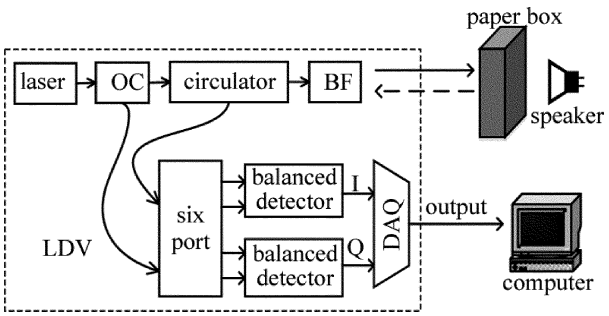

Fig. 4　LDV sound measurement system

别进入平衡外差探测器产生 I/Q 两路信号。两路信号经过放大器放大,然后通过数据采集(data acquisition, DAQ)卡(采样频率 100kHz)同时采集[17]。

为了使得实验数据的采集和分析过程更加便捷,作者搭建了基于 LABVIEW 的数据实时采集和分析系统。该系统在采集数据的同时,可以实时地采集数据做一些简单的处理,如带通滤波、平滑等,为采集数据和分析实验带来极大的方便。

2.2 数据集构建

用于训练神经网络的数据分为两部分,语音信号和 LDV 系统噪声。语音信号取自标准男声和女声朗读,总共 2h 的声音素材。在振动目标附近没有声源振动时,即可获得 LDV 测声系统的系统噪声。利用从 LDV 系统获取的系统噪声和干净的语料可以构建大量的平行语料:

$$Y = X + \alpha N(k), (k = \beta, \beta + 1, \cdots, T, \cdots, \beta - 1) \quad (14)$$

式中,X 表示干净语音信号的对数功率谱,N 表示 LDV 测声系统噪声的对数功率谱,Y 表示混合信号的对数功率谱,β 是一个随机因子。通过调整噪声的能量因子 α,达到可以控制输入神经网络信号的信噪比的目的,通过调整 α 的值,作者构建了 -6dB, -4dB, -2dB,0dB,2dB,4dB,6dB 信噪比的带噪语音信号,将带噪的语音信号降采样到 16kHz,每个文件的时长为 10s,做统一的滤波和归一化处理,然后将数据按 8:2 的个数比分为训练集和测试集两部分。为了测试网络结构在实际应用中的去噪性能,作者用扬声器播放标准男声和女声朗读,驱动振动目标振动,再通过 LDV 采集振动信息,通过 LDV 还原振动目标的振动信息后,即可获得带噪的语音信号,将它作为网络的另一部分测试样本。

2.3 数据处理

如图 5 所示,在训练阶段,先对训练样本进行特征提取,在这里用对数功率谱。将每个帧长设置为 1024,帧移是 512,然后计算重叠帧的 DFT 系数[11]。

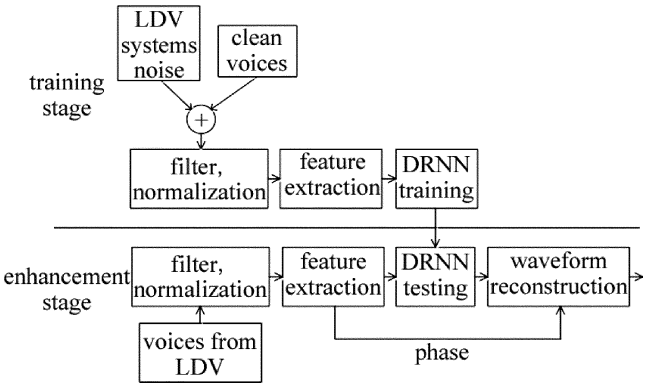

Fig. 5 Flow of data processing

为了测试训练好的网络的性能,作者首先用干净的语音信号与从 LDV 采集回来的系统噪声做直接相加,然后做滤波、归一化处理,进行特征提取,接着送入网络进行测试。为了测试训练好的网络在实际应用中的性能,将从 LDV 获取的带噪语音信号送入网络中进行测试。

3 实验分析

利用训练好的神经网络对从 LDV 测声系统获得的语音信号做降噪处理。如图 6 所示,从图 6a ~ 图 6c 依次为:干净的语音信号、从 LDV 测声系统采集回的带噪语音信号,以及处理后的结果的语谱图。

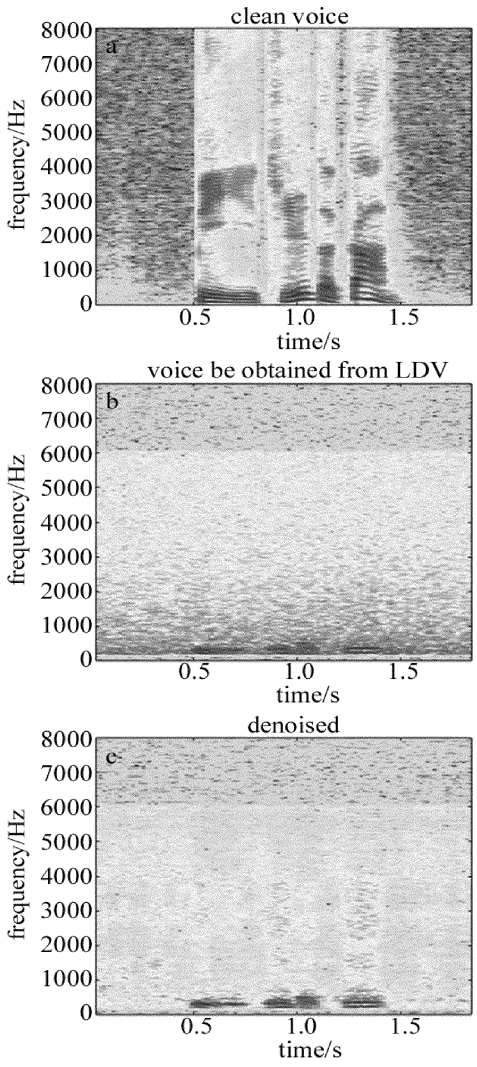

Fig. 6 Clean voice, the voice obtained from LDV and the denoised results

从图 6 中可知,从 LDV 系统获取的语音信号在整个频带内都被 LDV 系统噪声所污染,但通过 DRNN 网

络的降噪处理后,语音信号得到了明显的改善。

为了量化地评估作者所构建的 DRNN 网络的去噪能力,引入信号失真比率(sources-to-distortion ratio, SDR)、信号干扰比率(sources-to-interferences ratio, SIR)、信号人工比率(sources-to-artifacts,SAR),作为衡量语音信号去噪效果的评价指标[18],上述指标均为无量纲值。

为了测试不同的网络层数对去噪结果的影响,构建了不同参量结构的 DRNN 网络。图 7 所示为隐含层层数 l 为 1,2,3、隐含层神经元个数为 1024 的网络结构参量下的测试集的平均测试结果。图中用不同的柱状体表示对不同信噪比信号的处理结果,对比可以看出,网络层数的增加可以明显地提升网络的去噪性能。由此可知,网络层数的增加可以使得网络拟合 LDV 系统噪声和语音信号的能力更强,所以比单一层数的 RNN 网络有着更好的降噪效果。

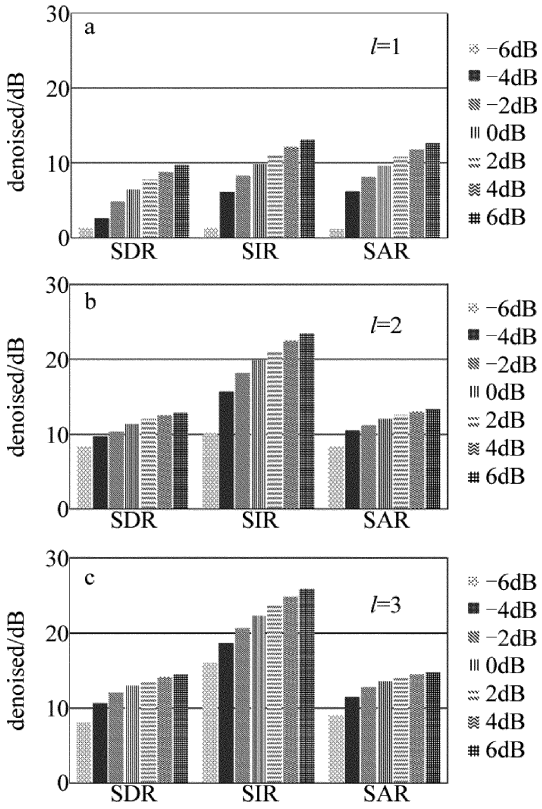

Fig. 7　Influence of network layers on the denoised results

与此同时,作者用非负矩阵分解去噪[19]的方法和

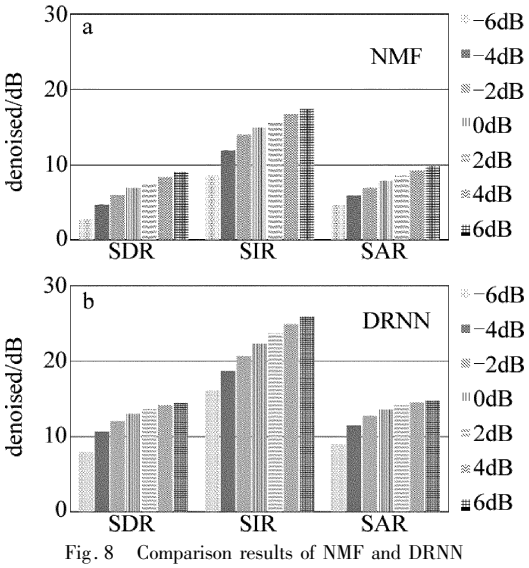

Fig. 8　Comparison results of NMF and DRNN

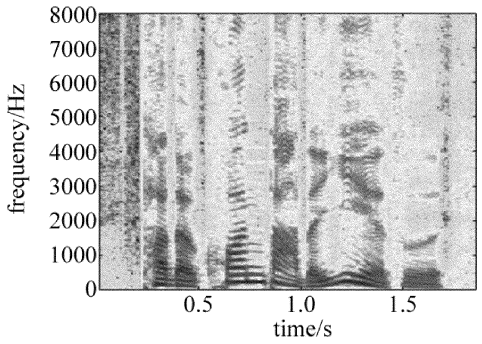

Fig. 9　Spectrogram of clean voice playing through a loudspeaker

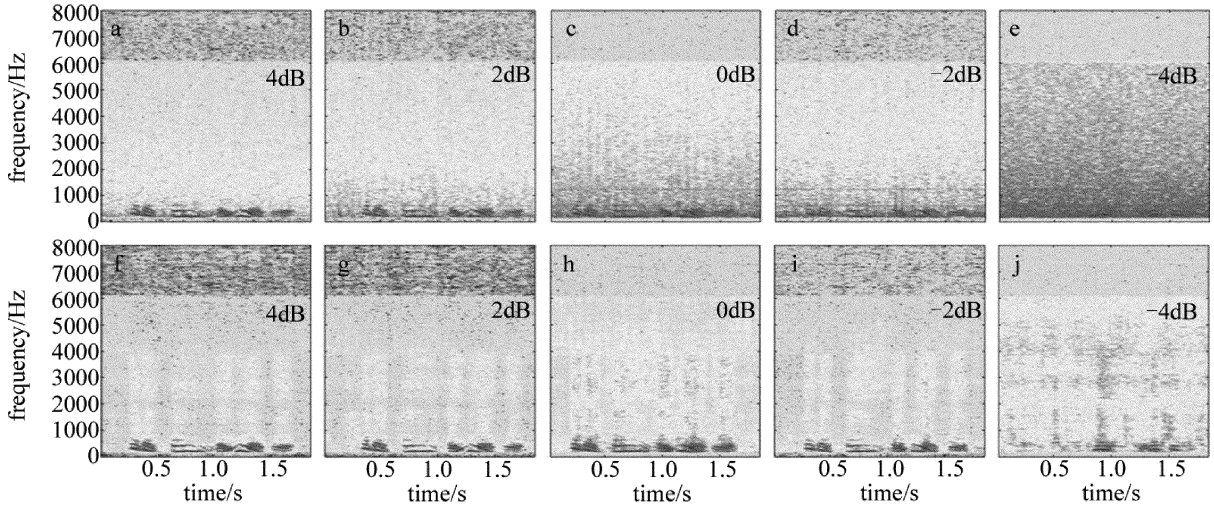

Fig. 10　The obtained voice from LDV and denoised results

本文中所用的方法做对照实验,如图 8 所示。图 8a 为使用非负矩阵分解的方法做不同信噪比 LDV 语音信号去噪的实验所得的结果,而图 8b 为使用网络层数为 3 的 DRNN 网络对 LDV 语音信号去噪的结果。可以看出,本文中所用的方法较非负矩阵分解去噪的方法有较大的提升。

为了验证 DRNN 网络对不同信噪比的从 LDV 获取的带噪语音信号的处理结果,利用扬声器控制播放的音量,以获取不同信噪比的带噪语音信号。图 9 所示为通过扬声器播放的干净的语音信号的语谱图。

利用 LDV 测声系统获取到的该语音信号和处理后的结果如图 10 所示。图 10a ~ 图 10e 的 5 幅图是从 LDV 测声系统获取的不同信噪比的带噪语音信号,图 10f ~ 图 10j 的 5 幅图是利用 DRNN 网络对其做降噪处理后的结果。每一列代表对不同信噪比的信号以及相应的处理结果。从图 10 中可以看出,作者所构建的 DRNN 网络在处理 4dB ~ -2dB 从 LDV 获取的语音信号时,可以得到很好的处理结果,语谱图中语音信号的结构也很清晰,但当信号的信噪比降到 -4dB,虽然背景噪声去除了很多,但是语音信号有明显的失真。

4 结 论

本文中构建基于语音信号降噪的 DRNN 网络,利用 LDV 测声系统采集 LDV 的系统噪声和从 LDV 获取的带噪语音信号,构建用于训练和测试网络的数据集。通过控制扬声器的分贝,以及 DRNN 网络结构和训练参量做了多组对照实验。实验结果表明,DRNN 网络可以对从 LDV 测声系统获取的语音信号做有效的降噪处理,对提升 LDV 测声系统的性能有着实际意义。

参 考 文 献

[1] LI F F, WU J, ZHAO Zh L, *et al*. Air coupled vibration detection of all-fiber laser Doppler vibrometer[J]. High Power Laser and Particle Beams, 2012, 24(11): 2549-2554(in Chinese).

[2] YU G, WANG Sh G, YU J H. Technology of digital speckle pattern interferometry and its applications[J]. Laser Technology, 2002, 26(3): 237-240(in Chinese).

[3] JING X J. Research and implementation of speech enhancement algorithm[D]. Hangzhou: Zhejiang University, 2005: 1-55(in Chinese).

[4] YUE D G, XIE Zh W. A new method to evaluate nonlinear distortion [J]. Technical Acoustics, 2007, 26(1): 84-89(in Chinese).

[5] 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, 2013, 23(1): 7-19.

[6] HUANG P S, MIN J K, MARK H J, *et al*. Joint optimization of masks and deep recurrent neural networks for monaural source separation[J]. IEEE-ACM Transactions on Audio Speech and Language Processing, 2015, 23(12): 2136-2147.

[7] HAN W, ZHANG X W, ZHOU X Y, *et al*. Joint optimization of deep neural networks and constrained Wiener filter for single channel speech enhancement[J]. Application Research of Computers, 2017, 34(3): 706-713(in Chinese).

[8] YAO Y, WANG Q J, ZHOU W, *et al*. Research on speech enhancement based on spectral subtraction and neural network[J]. Electronic Measurement Technology, 2017, 40(7): 74-79(in Chinese).

[9] LI W H, LIU M, ZHU Z G, *et al*. LDV remote voice acquisition and enhancement[J]. International Conference on Pattern Recognition, 2006, 20(24): 262-265.

[10] LÜ T,ZHANG H Y, GUO J, *et al*. Acquisition and enhancement of remote voice based on laser coherent method[J]. Optics and Precision Engineering, 2017, 25(3): 569-575(in Chinese).

[11] QU Zh, ZHANG B H. An improved wavelet threshold algorithm applied in laser interception[J]. Laser Technology, 2014, 38(2): 218-224(in Chinese).

[12] XU Y. Research on deep neural network based speech enhancement [D]. Hefei: University of Science and Technology of China, 2015: 55-75(in Chinese).

[13] CHENG Y P, BU F L. Experiment study on phase perception in speech[J]. Acta Acustica, 2003, 28(1): 7-11(in Chinese).

[14] HERMANS M,SCHRAUWEN B. Training and analyzing deep recurrent neural networks[R]. Lake Tahoe,USA: Proceedings of International Conference on Learning Representations(NIPS), 2013: 190-198.

[15] WERBOS P J. Backpropagation through time: what it does and how to do it[J]. Proceedings of the IEEE, 1990, 78(10): 1550-1560.

[16] HUO L, ZENG X D, AN Sh Y, *et al*. Vibration measurement and analysis by means of laser Doppler heterodyne principle[J]. Laser Technology, 2011, 35(5): 600-602(in Chinese).

[17] LIANG N. Research on laser Doppler vibrometer with homodyne detection[D]. Beijing: The University of Chinese Academy of Sciences, 2014: 13-23(in Chinese).

[18] VINCENT E, GRIBONVAL R, FEVOTTE C. Performance measurement in blind audio source separation[J]. IEEE Trans on Audio Speech & Language Processing, 2006, 14(4):1462-1469.

[19] ZHANG L W, JIA Ch, ZHANG X W, *et al*. Speech enhancement based on convolutive nonnegative matrix factorization with sparseness constraints[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 259-264(in Chinese).