文章编号：2095-2163(2019)05-0146-04 中图分类号：TN912.35 文献标志码：A

基于深度学习的语音增强方法研究

刘 鹏

（山西工程技术学院 信息工程与大数据科学系，山西 阳泉 045000）

摘 要： 针对基于深度学习的语音增强方法展开研究，系统阐述了基于深度学习的语音增强方法提出的背景、模型原理和实施过程。在TensorFlow平台上搭建了基于DNN的深度学习语音增强模型进行了实验，验证了基于DNN的语音增强方法，提高了增强语音的可懂度。

关键词： 深度学习；语音增强；DNN；语音可懂度

Research on speech enhancement method based on deep learning

LIU Peng

(Department of Information Engineering and Big Data Science, Shanxi Institute of Technology, Yangquan Shanxi 045000, China)

【Abstract】 The background, model principle and implementation process of speech enhancement based on deep learning are systematically expounded. A DNN-based deep learning speech enhancement model is built on the TensorFlow platform to conduct experiments, and it is verified that the speech enhancement method based on DNN improves the intelligibility of enhanced speech.

【Key words】 deep learning; speech enhancement; DNN; speech intelligibility

0 引 言

语音是人与人之间沟通交流的主要媒介，然而在现实生活中语音不可避免地会受到外界噪声的干扰，影响人们对语音的正确理解，特别是对于那些基于语音技术的实际应用领域。比如，自动语音识别技术（Automatic Speech Recognition，ASR）和人工耳蜗技术（Cochlear Implant，CI）等，噪声干扰严重制约了相关技术的发展。因此，研究如何从带噪语音中估计出纯净语音即显得尤为必要。

迄今为止，学者们提出了很多噪声去除和语音增强的方法，比如维纳滤波法（Wiener Filtering）、谱减法（Spectral Subtraction Method）、信号子空间方法（Signal Subspace Approach）和最小均方误差方法（Minimum Mean Square Error，MMSE）。然而，这些方法主要集中在研究语音与噪声的统计特性差异上，需要保证语音和噪声信号不存在相关关系，而且在降噪过程中会出现“音乐噪音”（music noise），导致语音失真[1]。此外，对于在语音增强中遇到的快速变化的噪声（如机关枪）和负谱估计等问题，传统的语音增强方法处理效果不佳[2]。

Rumelhart等3位学者在1988年发表的创新著作“Learning representations by back-propagating errors”中提出了多层神经网络，不仅可以用相对简单的方法进行有效的训练，而且隐藏层可以用来克服感知器在学习复杂模式时的弱点[3]。Hinton等学者[4]在2006年发表了一篇题为“A Fast Learning Algorithm for Deep Belief Nets”的突破性论文，使得深度学习技术得以兴起。这篇论文不仅首次提出了深度学习的概念，还展示了采用无监督方法进行逐层训练的有效性，并在此基础上进行了监督微调（fine-tuning），实现了MNIST字符识别数据集的最新结果。此后，Bengio等学者[5]随即发表了另一篇开创性的论文，即：Greedy Layer-wise Training of Deep Networks，揭示了为什么多层深度学习网络能够分层学习特性，而浅神经网络或支持向量机（SVM）则不能。该论文解释说明了使用DBNs、RBMs和自动编码器（AutoEncoder）的无监督方法进行预训练（pre-training）不仅可以初始化权值以获得最优解，而且提供了良好的可被学习的数据表示形式。Bengio等人在其论文“Scaling Algorithms Towards AI”中通过CNN、RBM、DBN等架构以及无监督的预训练和微调等技术重申了进行深度学习的优势，并引发了新一轮深度学习的研发热潮[6]。

近年来，随着基于深度学习的语音处理技术的逐步成功，不断有学者提出了基于深度学习的语音增强框架，期望从带噪语音噪声特征中预测出纯净语音特征来实现语音的降噪处理[7-11]。

作者简介：刘 鹏（1986-），男，硕士，助教/工程师，主要研究方向：语音处理、机器学习。

收稿日期：2019-07-18

1　语音增强和深度学习的概述

1.1　语音增强的过程和目标

语音增强是利用各种算法（包括传统的音频信号处理技术和现如今的深度学习技术）来提高退化语音信号（degraded speech signal）的质量（语音的听觉舒适度）或可懂度（语音的可理解性）[1]。其中，降噪语音增强是语音增强领域中最重要的研究方向，被广泛应用于手机、VoIP、电话会议系统、语音识别、助听器等领域。

1.2　语音增强的方法概述

传统的语音增强降噪算法可分为 3 类：滤波技术（Filtering Techniques）、频谱恢复（Spectral Restoration）和基于语音模型（Speech-Model-Based）的方法[1]。其中，滤波技术主要包括有维纳滤波法（WF）、谱减法（SSM）和信号子空间方法（SSA）。频谱恢复技术主要有最小均方误差短时谱振幅估计器方法（Minimum Mean-Square-Error Short-Time Spectral Amplitude Estimator，MMSE-STSA）。

1.3　深度学习的基本概念

深度学习是机器学习研究的一种形式，将其引入是为了使机器学习更接近研究的最初目标之一：人工智能。深度学习使计算机能够从经验数据中学习，并根据概念的层次来理解世界。由于计算机从经验数据中收集知识，因此不需要人工指定计算机所需的所有知识。概念的层次结构允许计算机从简单的概念中通过构建复杂的概念来学习，这使得层次结构图可有许多层。深度学习允许由多个处理层组成的计算模型中学习具有多个抽象级别的数据表示。这些方法极大地提高了语音识别、视觉目标识别、目标检测以及药物发现和基因组学等许多领域的技术水平。深度学习通过使用反向传播算法（Back-propagation algorithm）来指出计算机应该如何改变其内部参数来发现大数据集中复杂的结构，而这些参数用于从上一层的表示中来计算网络层次中的每一层表示[12]。

2　基于深度学习的语音增强方法

基于深度学习的语音增强框架是将 DNN 作为回归模型，从带噪语音的对数功率谱（LPS）中预测出纯净语音的对数功率谱（LPS）特征。DNN 也可以作为映射函数来学习纯净语音和带噪语音特征之间的关系，而不需要做出任何假设（传统的语音增强方法通常需要假设噪声信号与纯净语音信号不相关）。在基于 DNN 的语音增强中，目标特征与预测特征之间通常选取最小均方误差（*MMSE*）作为成本函数（cost function），运算时可参照式（1）和式（2），具体如下：

$$Cost(x,\tilde{x}') = \| x - \tilde{x}' \|^2 , \tag{1}$$

$$Cost(x,\tilde{x}') = \left\| x - \sigma'(\boldsymbol{W}'(\sigma(\boldsymbol{W}\tilde{x}' + \boldsymbol{b})) + \boldsymbol{b}') \right\|^2 . \tag{2}$$

其中，x 是所参考的纯净语音输入数据；$\tilde{x}'$ 是输出数据（由带噪语音所估计的数据）；$\boldsymbol{W}$ 是编码权重矩阵；$\boldsymbol{W}'$ 是解码权重矩阵；$\boldsymbol{b}$ 是编码偏置向量；$\boldsymbol{b}'$ 是解码偏置向量；σ 和 σ' 分别是编码激活函数和解码激活函数（这里均选取为 sigmoid 函数）。

通常，设计一个较 *MMSE* 更好的成本函数来直接优化 DNN 模型是困难的，特别是对于当语音特征和噪声特征存在相关关系时。Liu 等人[13]在其论文“Experiments on Deep Learning for Speech Denoising”中说明了其它成本函数，如 Kullback Leibler 散度或 Itakura-Saito 散度，在表现上要逊色于 *MMSE*。

基于 DNN 的语音增强框架如图 1 所示。模型采用 DNN 作为从带噪语音到清晰语音特征的映射函数，系统分 2 个阶段进行构建。对此可做阐释分述如下。

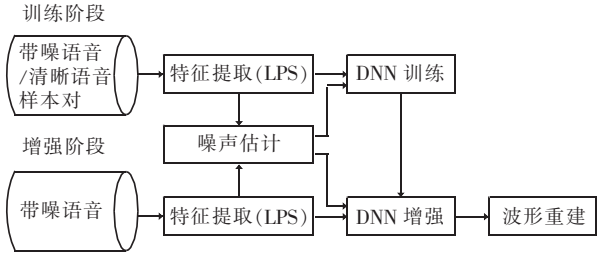

图 1　基于 DNN 的语音增强框架

Fig. 1　Speech enhancement framework based on DNN

（1）在训练阶段，利用从训练集中的带噪语音和清晰语音样本对中提取到的对数功率谱特征来训练基于 DNN 的回归模型。这里，采用对数功率谱特征（LPS）是为了提供与听觉感知相关的参数。DNN 训练时，先通过无监督预训练（unsupervised pre-training）的方式，初始化前面数层隐藏层参数。然后，通过反向传播（back-propagation）算法来训练 DNN，利用所估计的归一化对数功率谱特征与参考的纯净语音来建立 *MMSE* 成本函数，这是因为对数功率谱域的 *MMSE* 准则与人类听觉感知系统具有一致性。最后，采用微调（fine-tuning）来对整个 DNN 网络中的所有参数进行有监督训练，采用小批量（mini-batch）随机梯度下降（stochastic gradient

descent)算法进行不断优化[12]。

(2)在增强阶段,先利用训练好的 DNN 模型对带噪语音特征进行提取,预测出清晰的语音特征。然后,将所得到的对纯净语音 LPS 的估计 $\hat{X}^{l}(d)$ 通过波形重建获得重建后的增强语音谱 $\hat{X}^{f}(d)$,其数学公式可表示为:

$$\hat{X}^{f}(d)=\exp\{\hat{X}^{l}(d)/2\}\exp\{j\angle X^{f}(d)\} \quad (3)$$

其中,$\angle X^{f}(d)$ 表示带噪语音信号的第 d 维相位。

3 基于深度学习的语音增强建模实验

本节基于深度神经网络(DNN)建立语音增强模型,并与传统的语音增强算法(子空间法)在语音增强的可懂度效果上进行了实验对比。对此部分可详述如下。

3.1 实验步骤

3.1.1 实验环境搭建及数据准备

在 TensorFlow 深度学习框架中搭建了基于 DNN 的语音增强模型。噪声信号选取为 NOISEX-92 标准库中的 4 种噪声,分别为 babble、car、street 和 train,纯净语音句子来源于 IEEE 句子库,信噪比分别为-15 dB、-10 dB 和-5 dB。信号的量化精度为 16 bit,采样频率设置为 8 kHz。

DNN 模型的训练集由 IEEE 句子库中的前 600 个句子,依据 4 种类型噪声×3 种信噪比、共计 12 种加噪条件产生的带噪语音和其所参考的清晰语音构成。因此,实验中由 7 200 个语音样本对组成 DNN 模型的训练数据集。

DNN 模型的测试集由 IEEE 句子库中的后 120 个句子,依据 4 种类型噪声×3 种信噪比、共计 12 种加噪条件产生的带噪语音组成。因此,由 1 440 个语音样本组成实验中 DNN 模型的测试数据集。

3.1.2 特征提取

在模型训练阶段,首先对训练数据集中的带噪语音和纯净语音信号样本对进行短时傅里叶分析,分别计算每个重叠窗口帧的离散傅里叶变换(DFT),然后分别计算其对数功率谱(LPS)来作为 DNN 模型训练的特征数据。在语音增强阶段,将测试数据集中的带噪语音进行短时傅里叶分析后计算每个重叠窗口帧的离散傅里叶变换(DFT),再将其对数功率谱(LPS)作为模型的输入数据。

3.1.3 DNN 模型建立及参数配置

实验中 DNN 模型由 1 个输入层,3 个隐藏层(每层 500 个神经元)和 1 个输出层构成。每层的预训练轮数(*epoch*)设置为 20,预训练的学习速率设置为 0.000 5。在参数微调时,前 10 轮(*epoch*)的学习速率设置为 0.1,此后每轮学习速率都下降 10%,总共进行 50 轮训练。采用小批量(mini-batch)随机梯度下降(stochastic gradient descent)算法进行调优处理,小批量(mini-batch)数据集大小设置为 N=128。

3.2 实验结果及分析

本文的语音可懂度测试采用归一化协方差法(NCM)。研究表明,子空间法是传统的语音增强算法中语音可懂度增强效果较好的一种增强算法[14]。故而实验选用了子空间法和加噪未增强两种处理方式与本文的增强算法进行对比。实验中语音可懂度的 NCM 评价结果见表 1 ~ 表 3。

表 1 *SNR*= -15 dB 语音可懂度的 NCM 评价结果

Tab. 1 NCM evaluation results of speech intelligibility under *SNR*= -15 dB

噪声类型	加噪未增强	子空间法增强	基于 DNN 增强
Babble	0.26	0.22	0.31
Car	0.29	0.27	0.45
Street	0.32	0.25	0.42
Train	0.29	0.24	0.38

表 2 *SNR*= -10 dB 语音可懂度的 NCM 评价结果

Tab. 2 NCM evaluation results of speech intelligibility under *SNR*= -10 dB

噪声类型	加噪未增强	子空间法增强	基于 DNN 增强
Babble	0.41	0.36	0.47
Car	0.47	0.52	0.66
Street	0.52	0.48	0.61
Train	0.42	0.40	0.53

表 3 *SNR*= -5 dB 语音可懂度的 NCM 评价结果

Tab. 3 NCM evaluation results of speech intelligibility under *SNR*= -5 dB

噪声类型	加噪未增强	子空间法增强	基于 DNN 增强
Babble	0.60	0.61	0.67
Car	0.68	0.69	0.76
Street	0.70	0.70	0.73
Train	0.61	0.58	0.69

实验结果中的 NCM 数值越大,表示其可懂度越高,从表 1 ~ 表 3 语音 NCM 测试值可以看出:对比其它 2 种对带噪语音的处理(加噪未增强,子空间法增强),基于 DNN 的语音增强方法提高了增强后带噪语音的可懂度。

由于噪声或信噪比估计误差会导致语音增强处理频谱中出现伪峰,几乎所有传统的语音增强方法都出现了音乐噪声。与之不同的是,基于深度学习

的语音增强中没有发现音乐噪声。此外，深度学习模型可以恢复被噪声掩盖了的语音高频频谱[15]。因此，基于深度学习的语音增强方法较传统的语音增强能够表现出更好的语音可懂度增强效果。

4　结束语

本文针对基于深度学习的语音增强方法展开研究，系统阐述了基于深度学习的语音增强方法提出的背景、模型原理和实施过程。在 TensorFlow 平台上搭建了基于 DNN 的深度学习语音增强模型，并进行了实验，验证后可知基于 DNN 的语音增强方法提高了增强语音的可懂度。

值得注意的是，基于深度学习的语音增强方法需要用到规模较大的语音训练集样本对，特别是当所构建的模型规模较大而训练集的样本数量又极少时，模型极易出现过拟合现象，这将最终使得模型在语音增强阶段失效。

参考文献

[1] LOIZOU P C. Speech enhancement: Theory and practice[M]. 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.

[2] 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.

[3] RUMELHART D E, HINTON G E, WILLIAMS R J. Learning internal representations by error propagation [M] // Neurocomputing: foundations of research. Cambridge, MA, USA: MIT Press, 1988: 696-699.

[4] HINTON G E, OSINDERO S, TEH Y W. A fast learning algorithm for deep belief nets [J]. Neural Computation, 2006, 18(7):1527-1554.

[5] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks [C] // Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: dblp, 2006: 153-160.

[6] BOTTOU L, CHAPELLE O, DECOSTE D, et al. Large-scale kernel machines[M]. Cambridge, MA, USA: MIT Press, 2007.

[7] KOLBæK M, TAN Zhenghua, JENSEN J. Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems [J]. IEEE/ACM Trans Audio, Speech and Language Processing, 2017, 25(1): 153-167.

[8] TU Y H, DU J, LEE C H. DNN training based on classic gain function for single-channel speech enhancement and recognition [C] // 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Bringhton: IEEE, 2019: 910-914.

[9] ODELOWO B O, ANDERSON D V. A study of training targets for deep neural network-based speech enhancement using noise prediction [C] // 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, AB, Canada: IEEE, 2018: 5409-5413.

[10] LAI Y H, CHEN F, WANG S S, et al. A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation [J]. IEEE Transactions on Biomedical Engineering, 2017, 64(7): 1568-1578.

[11] LAI Y H, TSAO Y, LU X, et al. Deep learning based noise reduction approach to improve speech intelligibility for cochlear implant recipients [J]. Ear Hear, 2018, 39(4): 795-809.

[12] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning [M]. Cambridge, MA, USA: MIT Press, 2016.

[13] LIU Ding, SMARAGDIS P, KIM M. Experiments on deep learning for speech denoising [C] //15th Annual Conference of the International Speech Communication Association (INTERSPEECH-2014). Singapore: ISCA, 2014: 2685-2689.

[14] HU Yi, LOIZOU P C. A comparative intelligibility study of single-microphone noise reduction algorithms[J]. The Journal of the Acoustical Society of America, 2007, 122(3): 1777-1786.

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

(上接第 145 页)

[4] 郝亮，徐涛，崔健，等. 参数化诱导槽设计的吸能盒结构抗撞性多目标优化[J]. 吉林大学学报(工学版)，2013，43(1)：39-44.

[5] 严杰，谭伟，孙伟卿，等. 基于 Ls-Dyna 和 Hyperstudy 的汽车吸能盒优化分析[J]. 汽车科技，2013(6)：63-66.

[6] 雷刚，谭皓文，樊伟，等. 基于汽车正面碰撞的吸能盒设计及优化[J]. 重庆理工大学学报(自然科学)，2013，27(3)：1-5.

[7] 陈有松，孙万朋，安超群，等. 基于低速正碰的吸能盒式防撞梁吸能特性研究[J]. 现代制造工程，2018(3)：53-58.

[8] 朱永梅，戴永健，朱俊臣，等. 蛋形混凝土耐压壳的设计与力学特性研究[J]. 舰船科学技术，2018，40(11)：48-51，120.

[9] 蒋致禹，顾敏童，赵永生. 一种薄壁吸能结构的设计优化[J]. 振动与冲击. 2010，29(2)：111-116.

[10] 于用军，郭永奇，李飞，等. 铝合金吸能盒的结构设计及耐撞性分析[J]. 汽车实用技术，2017(22)：55-57.

[11] 陈宇，纪宝钢，钟金发，等. 基于 FEM 的铝制吸能盒结构优化设计[J]. 工具技术，2015(1)：44-47.

[12] 侯淑娟，龙述尧，李光耀，等. 材料参数对车身吸能元件抗撞性影响的数值分析 [J]. 汽车工程，2008，30(5)：416-419，423.

[13] 万鑫铭，徐小飞，徐中明，等. 汽车用铝合金吸能盒结构优化设计[J]. 汽车工程学报，2013，3(1)：15-21.

[14] 曲明，柳艳杰，夏春艳，等. 吸能盒在低速撞击情况下的仿真与分析[J]. 应用科技，2008，35(8)：59-64.

[15] 柳艳杰. 汽车低速碰撞吸能部件的抗撞性能研究[D]. 哈尔滨：哈尔滨工程大学，2012.

[16] MIRZAEI M, SHAKERI M. SADIGHI M, et al. Crash worthinesss design for cylindrical tube using neural network and genetic algorithm[J]. Procedia Engineering, 2011, 14(4): 3346-3353.

[17] 龚洁. 汽车前保险杠碰撞的有限元仿真分析研究[D]. 沈阳：东北大学，2010.

[18] 雷刚，谭皓文，樊伟，等. 基于汽车正面碰撞的吸能盒设计及优化[J]. 重庆理工大学学报(自然科学)，2013，27(3)：1-5.