文章编号：2095-2163(2019)06-0097-04 中图分类号：TN912.35 文献标志码：A

基于混合深度神经网络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：针对基于混合深度神经网络的语音增强方法展开研究，阐述了该方法提出的背景、模型原理和实施过程。搭建了基于混合深度神经网络的深度学习语音增强模型，并与仅基于 DNN 的语音增强模型进行了对比实验，验证了基于混合深度神经网络的语音增强方法，进一步提高了增强语音的质量。

关键词：混合深度神经网络；语音增强；深度学习；语音质量

Research on speech enhancement method based on hybrid deep neural networks

LIU Peng

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

【Abstract】 The background, model principle and implementation process of speech enhancement based on hybrid deep neural networks are systematically expounded. A hybrid DNN-based deep learning speech enhancement model was built and compared with the stand alone DNN-based model. Experimental results show that the speech enhancement method based on hybrid DNN further improves the quality of enhanced speech.

【Key words】 hybrid deep neural networks; speech enhancement; deep learning; speech quality

0 引 言

近几十年来，语音增强(speech enhancement)由于其在移动电话、语音识别、助听器设计等实时应用方面的重要性而受到研究者的关注。语音增强方法的主要目的是在不失真的情况下提高退化语音(deteriorated speech)信号的语音质量。为此，各国学者设计了许多算法。比如，谱减法是带噪语音减去短期噪声频谱的估计值，从而产生纯净语音的估计值频谱[1]。信号子空间法是将带噪语音信号通过矩阵分解的方法分解为信号子空间和噪声子空间，进而获得纯净语音信号的频谱估值[2]。但是，在这些传统方法中经常遇到的问题是：由此产生的增强语音经常受到一种人为因素的影响，即“音乐噪声”[3]。而且，由于传统的语音增强方法往往假设噪声信号是平稳的并且噪声信号与语音信号不存在相关关系，这使得传统语音增强算法无法适用于非平稳噪声的现实情况。

上世纪 90 年代，考虑到噪声对语音干扰的复杂过程，部分学者开始采用神经网络等非线性模型来建立带噪语音与纯净语音信号之间的映射关系。文献[4]和文献[5]利用浅层神经网络(shallow neural networks)作为非线性滤波器来预测时域或频域内的纯净信号。然而，浅层神经网络的网络规模小，不能充分学习带噪语音特征与目标信噪比之间的关系。不仅如此，浅层神经网络的随机初始化常常会出现明显的局部极小值或停滞，对于包含更多隐藏层的体系结构，问题会更为明显[6]。2006 年 Hinton 等学者在其论文“A fast learning algorithm for deep belief nets”和“Reducing the dimensionality of data with neural networks”中提出了一种贪婪的分层学习算法，为训练深度架构带来了突破，同时也迎来深度学习技术的大繁荣[7-8]。深度学习模型的每一层都进行预训练，以学习其输入(或前一层的输出)的高级表示。对于回归任务，深度学习已被应用于多个语音合成任务中[9-10]。在文献[11]和[12]中，堆叠降噪自编码器(stacked denoising autoencoders)作为一种深度模型来建立带噪语音和纯净语音信号特性之间的关系。为了捕捉语音信号的时间特性，部分学者还引入了循环神经网络(recurrent neural networks)，从而消除了多层感知器(multilayer perceptrons)中对上下文窗口的显式选择，文献[13]和[14]采用深度循环神经网络(deep recurrent neural networks)为鲁棒语音识别(robust speech recognition)进行特征增强。但在有限噪声类型下训练的深度循环神经网络泛化能力较弱。此外，近年

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：2019-08-27

来基于对带噪语音频谱图(spectrograms)处理的语音增强算法也不断被提出。Fu 等学者使用卷积神经网络(convolutional neural networks)直接从带噪语音的频谱图中估计出了纯净语音的频谱图,该方法较基于深度神经网络(deep neural networks)的幅度处理方法相比性能有了很大提高[15]。

随着学者对深度学习模型研究的不断深入,人们开始尝试将深度学习模型与原有机器学习模型(如SVM或GMM)或者不同深度学习模型之间进行联合,构建出混合的深度学习模型结构,比如:DNN—HMM 结构、DNN—GMM 结构、CNN—RNN 结构、CNN—HMM 结构以及 RNN—HMM 结构等。研究发现,使用这些混合网络相较于单一网络结构能够获得更好的性能和实验效果[16]。

1 基本方法概述

1.1 语音增强的概念

语音增强是指通过抑制噪声来改善听众对带噪语音某方面的感知体验。在实际应用中,语音增强对带噪语音感知体验的改善主要有质量(quality)和可懂度(intelligibility)两个方面。针对带噪语音质量的改善是非常必要的,特别是在其长时间暴露于诸如工厂生产车间或航空飞机场等高分贝噪音环境下,语音质量的改善可以减少听众的听觉疲劳。使用语音增强算法可以在一定程度上降低或抑制背景噪声,因此有时也称其为噪声抑制算法(noise suppression algorithms)[3]。

1.2 深度学习模型

深度学习指的是广泛的机器学习技术以及基于多层非线性信息处理的体系结构,这些信息处理本质上被认为是分层的。深度学习的模型结构可以分为单一独立(Standalone)结构(通常包括 DNNs、CNNs 和 RNNs 等)和混合(hybrid)结构(包括 DNN—HMM、DNN—GMM、CNN—RNN、CNN—HMM 和 RNN—HMM 等)[16]。

卷积神经网络(CNNs)被认为是一个由多个特征提取阶段所构成的深层体系结构,其中每个阶段都包含一个卷积层和一个池化层以及非线性激活函数(ReLU),通过这样的组合方式力求接近复杂的非线性模型函数。卷积层共享了权值,而池化层对来自卷积层的输出进行采样,降低了数据维度。CNNs 假设特征具有不同层次结构并可以通过卷积内核提取。在监督训练过程中,通过学习层次特征来完成既定的任务。

循环神经网络(RNNs)是一类允许通过网络的不同层共享参数的深度神经网络。RNNs 是基于类似树的结构上循环地使用相同的权值集来开发的,该树按拓扑顺序遍历。RNNs 主要用于利用已有的数据样本预测未来的数据序列。当涉及到语音或文本等序列数据的建模时,RNNs 是非常流行的。

将卷积神经网络(CNNs)与循环神经网络(RNNs)相结合,用于对音频信号或单词序列等序列数据进行建模,这种混合模型称为卷积循环神经网络(CRNNs)。通过用 RNNs 替换最后一层卷积,可以将 CRNNs 描述为一个经过修改的 CNNs。在 CRNNs 中,CNNs 和 RNNs 分别扮演着特征提取器和时间归纳器的角色。采用 RNNs 对特征进行聚类,使得网络能够考虑全局结构,而局部特征由卷积层提取。这种结构最初是在文献[17]中提出用于文档分类,文献[18]采用该结构进行了音乐标注。

2 基于混合深度神经网络的语音增强方法

2.1 模型概述

基于混合深度神经网络的语音增强模型由三个部分组成:首先,将带噪语音频谱图与若干个卷积核(kernel)进行卷积,形成特征图(feature maps),并将所有特征图拼接成一个二维特征图;然后,利用双向 RNNs 在时间维度对二维特征图进行进一步的变换,建立连续帧之间的动态关联;最后,建立预测频谱图和纯净语音频谱图之间的成本函数(cost function),利用全连接层(Fully Connected Layer)对纯净语音频谱图逐帧进行预测。与已有的 DNNs 和 RNNs 模型相比,由于卷积内核的稀疏性,该混合网络具有更高的数据效率和处理效率。此外,双向循环网络使得模型能够自适应地对连续帧之间的动态关联进行建模。

2.2 模型建立

假定 $\boldsymbol{y}$ 和 $\boldsymbol{x}$ 分别为带噪语音和其所对应的纯净语音频谱图,其维度均为 $\boldsymbol{d} \times \boldsymbol{t}$。其中,$\boldsymbol{d}$ 表示频谱图的频带数目,$\boldsymbol{t}$ 表示频谱图的长度。假定 $\boldsymbol{z}$ 为卷积核,其维度为 $\boldsymbol{b} \times \boldsymbol{w}$。将带噪语音频谱图 y 与内核 z 进行卷积,所形成的特征图如公式(1)所示。

$$\boldsymbol{h}_z(\boldsymbol{y}) = \sigma(\boldsymbol{y} * \boldsymbol{z}), \tag{1}$$

其中,$\sigma(\cdot)$ 为 ReLU 激活函数。每个这样的卷积 $\boldsymbol{z}$ 核都会产生一个二维特征图。将 k 个单独的卷积核应用到输入频谱图上,就得到一个二维特征图集合 $\{\boldsymbol{h}_{z_i}(\boldsymbol{y})\}_{i=1}^{k}$。然后,将所得到的二维特征图集合按照公式(2)所示,沿着特征维垂直拼接形成

一个堆叠的二维特征图 $\boldsymbol{H}(\boldsymbol{y})$，其中包含了之前卷积特征图的所有信息。

$$\boldsymbol{H}(\boldsymbol{y}) = [\boldsymbol{h}_{z_1}(\boldsymbol{y}); \ldots; \boldsymbol{h}_{z_k}(\boldsymbol{y})], \quad (2)$$

通过双向长短期记忆网络（Long Short-Term Memory）建立双向循环神经网络。在每一时间步 t 上对二维特征图 $\boldsymbol{H}(\boldsymbol{y})$ 进行变换，每个单向 LSTM 神经元通过门控制器按照公式（3）~（7）更新其隐藏层表示 $\overrightarrow{\boldsymbol{H}_t}$。

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{yi}\boldsymbol{H}_t + \boldsymbol{W}_{hi}\overrightarrow{\boldsymbol{H}}_{t-1} + \boldsymbol{W}_{ci}\boldsymbol{c}_{t-1}), \quad (3)$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{yf}\boldsymbol{H}_t + \boldsymbol{W}_{hf}\overrightarrow{\boldsymbol{H}}_{t-1} + \boldsymbol{W}_{cf}\boldsymbol{c}_{t-1}), \quad (4)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + i_t \odot \tanh(\boldsymbol{W}_{yc}\boldsymbol{H}_t + \boldsymbol{W}_{hc}\overrightarrow{\boldsymbol{H}}_{t-1}), \quad (5)$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{yo}\boldsymbol{H}_t + \boldsymbol{W}_{ho}\overrightarrow{\boldsymbol{H}}_{t-1} + \boldsymbol{W}_{co}\boldsymbol{c}_t), \quad (6)$$

$$\overrightarrow{\boldsymbol{H}_t} = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t), \quad (7)$$

其中，$\sigma(\cdot)$ 为 sigmoid 激活函数；$\odot$ 表示点乘运算；$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$ 依次代表了输入门、遗忘门和输出门。

双向 LSTM 的隐藏层表示 $\widetilde{\boldsymbol{H}_t}$ 由两个不同方向的单向隐藏层表示拼接而成，即 $\widetilde{\boldsymbol{H}_t} = [\overrightarrow{\boldsymbol{H}_t}; \overleftarrow{\boldsymbol{H}_t}]$。

利用全连接层对纯净语音频谱图逐帧进行预测，为保证预测结果非负，采用公式（8）进行线性回归预测。

$$\hat{\boldsymbol{x}}_t = \max\{0, \boldsymbol{W}\tilde{\boldsymbol{H}}_t + \boldsymbol{b}\}, \quad (8)$$

其中，$\boldsymbol{W}$ 为权重矩阵，$\boldsymbol{b}$ 为偏置向量。最后，通过纯净语音预测值 $\hat{\boldsymbol{x}}$ 和对应的纯净语音 x 之间的均方误差（MSE）建立模型的成本函数，如公式（9）。

$$\mathrm{Cost}(\boldsymbol{x}, \hat{\boldsymbol{x}}) = \| \boldsymbol{x} - \hat{\boldsymbol{x}} \|^2. \quad (9)$$

模型训练采用 Adadelta 算法进行模型参数的优化。

3　基于混合深度神经网络的语音增强实验

将基于混合深度神经网络所建立的语音增强模型与仅基于 DNN 的语音增强模型在语音增强的质量效果进行了实验对比。

3.1　实验过程

3.1.1　数据准备

分别搭建基于混合深度神经网络和仅基于 DNN 的语音增强模型。纯净语音选自 TIMIT 数据库，噪声信号选取 NOISEX-92 中的 babble、ca、street 和 train 四种噪声，按照-5 dB、0 dB 和 5 dB 分别加噪。

两种模型的训练数据集均由 TIMIT 数据库中的全部训练集4 620 个句子，按照不同噪声类型（4 种）结合不同信噪比（3 种）所产生的不同加噪条件（12 种）的带噪语音和与之对应的纯净语音组成。所以，采用了55 440 个语音对来构成两种模型的训练数据集。

两种模型的测试数据集均由 TIMIT 数据库中的全部测试集1 680 个句子，按照不同噪声类型（4 种）结合不同信噪比（3 种）所产生的不同加噪条件（12 种）的带噪语音和与之对应的纯净语音组成。所以，采用了20 160 个语音对来构成两种模型的测试数据集。

3.1.2　模型参数配置

基于混合深度神经网络的语音增强模型实验中，作为预处理步骤，首先使用短时傅里叶变换（STFT）从每个话语中提取频谱图。每个频谱图中有 256 个频带（$\boldsymbol{d}$ = 256）和 500 帧（$\boldsymbol{t}$ = 500）。模型卷积层中有 256 个维度为 32 × 11 的卷积核，滑动步长（stride）频率维度为 16，时间维度为 1，边缘外自动补 0。在卷积层之后使用了两层双向 LSTMs，每层都有 1 024 个隐藏单元。

仅基于 DNN 的语音增强模型实验中，DNN 模型包含 3 个隐藏层，每个层都有 2 048 个隐藏单元。

3.2　实验结果及分析

实验中语音质量的评价选用 PESQ 方法，语音质量的 PESQ 评价结果见表 1~表 3 所示。

表 1　SNR= -5 dB 语音质量的 PESQ 值

Tab. 1　PESQ value of speech quality under SNR=-5 dB

噪声类型	仅基于 DNN 增强	基于混合网络增强
Babble	2.03	2.40
Car	1.89	2.24
Street	1.75	2.07
Train	1.80	2.13

表 2　SNR= 0 dB 语音质量的 PESQ 值

Tab. 2　PESQ value of speech quality under SNR= 0 dB

噪声类型	仅基于 DNN 增强	基于混合网络增强
Babble	2.49	2.90
Car	2.32	2.70
Street	2.18	2.54
Train	2.25	2.63

表 3　SNR= 5 dB 语音质量的 PESQ 值

Tab. 3　PESQ value of speech quality under SNR= 5dB

噪声类型	仅基于 DNN 增强	基于混合网络增强
Babble	2.87	3.30
Car	2.68	3.08
Street	2.55	2.94
Train	2.62	3.01

语音质量的 PESQ 值越高说明对应的语音主观听觉质量越好,从表 1~表 3 语音 PESQ 测试值可以看出:相较于仅基于 DNN 的语音增强模型,基于混合深度神经网络的语音增强模型进一步提高了增强语音的质量。

由于在所构建的混合深度神经网络中,CNNs 和 RNNs 分别扮演了特征提取器和时间归纳器的角色。采用双向 LSTMs 对特征进行聚类,使得网络能够考虑语音的全局结构,而局部特征可以由卷积层提取。因此,基于混合深度神经网络的语音增强方法较仅基于 DNN 的语音增强方法能够学习到语音中更多的上下文全局信息,表现出更好的语音质量增强效果。

4 结束语

本文针对基于混合深度神经网络的语音增强方法展开了研究,阐述了该方法提出的背景、模型原理和实施过程,搭建了基于混合深度神经网络的语音增强模型和仅基于 DNN 的语音增强模型,进行了对比实验,验证了基于混合深度神经网络的语音增强方法,进一步提高了增强语音的质量。

参考文献

[1] BOLL, S F. Suppression of acoustic noise in speech using spectral subtraction [J]. IEEE Trans. Acoust. Speech Signal Process., 1979, 27(2):113-120.

[2] HU Y, LOIZOU P C. A generalized subspace approach for enhancing speech corrupted by colored noise [J]. Speech and Audio Processing, IEEE Transactions on, 2003, 11(4): 334-341.

[3] LOIZOU P C. Speech Enhancement: Theory and Practice (Second Edition) [M]. Boca Raton, FL, USA: CRC Press, 2013:1-2, 225-227.

[4] TAMURA S I. An analysis of a noise reduction neural network [C]//ICASSP, 1989: 2001-2004.

[5] XIE F, COMPERNOLLE D V. A family of MLP based nonlinear spectral estimators for noise reduction[C]// ICASSP, 1994:53-56.

[6] BENGIO Y. Learning deep architectures for AI [J] Foundat. Trends Mach. Learn., 2(1), 2009: 1-127.

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

[8] HINTON G E, SALAKHUTDINOV R R. Reducing the dimensionality of data with neural networks [J]. Science, 313 (5768), 2006:504-507.

[9] CHEN L H, LING Z H, LIU L J, et al. Voice conversion using deep neural networks with layer-wise generative training [J]. IEEE/ACM Trans. Audio, Speech, Lang. Process., 22(12), 2014:1859-1872.

[10] LING Z H, DENG L, YU D. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis [J]. IEEE Trans. Audio, Speech, Lang.Process.,21(10), 2013:2129-2139.

[11] XIA B Y, BAO C C. Speech enhancement with weighted denoising Auto-Encoder [J]. Interspeech, 2013:3444-3448.

[12] LU X G, TSAO Y, MATSUDA S, et al. Speech enhancement based on deep denoising autoencoder [J]. Interspeech, 2013, 25-29:436-440.

[13] MAAS A L, LE Q V, O'NEIL T M, et al. Recurrent neural networks for noise reduction in robust ASR [J]. Interspeech, 2012: 22-25.

[14] WOLLMER M, ZHANG Z, WENINGER F, et al. Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise[C] // ICASSP, 2013: 6822-6826.

[15] FU S, HU T, TSAO Y, et al. Complex spectrogram enhancement by convolutional neural network with multi-metrics learning[C]// IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP), 2017:1-6.

[16] NASSIF A B, SHAHIN I, ATTILI I, et al. Speech recognition using deep neural networks: a systematic review [J]. IEEE Access, 2019: 19143-19165.

[17] TANG D, QIN B, LIU T. Document modeling with gated recurrent neural network for sentiment classification [C]// Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1422-1432.

[18] CHOI K, FAZEKAS G, SANDLER M, et al. Convolutional recurrent neural networks for music classification [C]//ICASSP, 2017.

(上接第 96 页)

[3] 蔡立新,李嘉欢. 大数据时代企业财务风险预警机制与路径探究[J]. 财会月刊,2018(15):38-43.

[4] 张润驰,杜亚斌. 基于粒子群优化聚类算法的多预测器信用评估模型[J]. 系统工程,2017,35(10):154-158.

[5] 满春涛,刘博,曹永成. 粒子群与遗传算法优化支持向量机的应用[J]. 哈尔滨理工大学学报,2019,24(3):87-92.

[6] 汤深伟,贾瑞玉. 基于改进粒子群算法的 K 均值聚类算法[J]. 计算机工程与应用,2019,55(18):140-145.

[7] 王佳信,周宗红,付斌,等. 因子分析-概率神经网络模型在边坡稳定性评价中的应用[J]. 水文地质工程地质,2018,45(2):123-130.

[8] 蒋玉秀,赵晓欢,邓元望. 基于概率神经网络的电子油门踏板故障诊断[J]. 中南大学学报(自然科学版),2019,50(6):1370-1376.

[9] 刘嘉蔚,李奇,陈维荣,等. 基于概率神经网络和线性判别分析的 PEMFC 水管理故障诊断方法研究[J]. 中国电机工程学报,2019,39(12):3614-3622.