Software Engineering and Applications 软件工程与应用, 2017, 6(1), 8-19
Published Online February 2017 in Hans. http://www.hanspub.org/journal/sea
https://doi.org/10.12677/sea.2017.61002

Hans汉斯

Single-Channel Speech Enhancement Based on Sparse Regressive Deep Neural Network

Haixia Sun, Sikun Li

National University of Defense Technology (NUDT), Changsha Hunan
Email: shx_one@163.com, sikunli@126.com

Received: Feb. 7th, 2017; accepted: Feb. 25th, 2017; published: Feb. 28th, 2017

Abstract

Speech enhancement is a mean to improve the quality and intelligibility by noise suppression and enhancing the SNR at the same time, which has been widely applied in voice communication equipments. In recent years, Deep Neural Network (DNN) has become a research hot point due to its powerful ability to avoid local optimum, which is superior to the traditional neural network. However, the existed DNN costs storage and has a bad generalization. Now, this document puts forward a sparse regression DNN model to solve the above problems. First, we will take two regularization skills called Dropout and sparsity constraint to strengthen the generalization ability of the model. Obviously, in this way, the model can reach the consistency between the pre-training model and the training model. Then network compression by weights sharing and quantization is taken to reduce storage cost. Next, spectral subtraction is used in post-processing to overcome stationary noise. The result proofs that the improved framework gets a good effect and meets the requirement of the speech processing.

Keywords

Speech Enhancement, DNN, Regularization Technique, Network Compression, Spectral Subtraction

基于稀疏回归深度神经网络的单通道语音增强

孙海霞，李思昆

国防科学技术大学，湖南 长沙
Email: shx_one@163.com, sikunli@126.com

收稿日期：2017年2月7日；录用日期：2017年2月25日；发布日期：2017年2月28日

文章引用: 孙海霞, 李思昆. 基于稀疏回归深度神经网络的单通道语音增强[J]. 软件工程与应用, 2017, 6(1): 8-19.
https://doi.org/10.12677/sea.2017.61002

摘　要

语音增强可以改进语音质量，抑制、降低噪声干扰，提高信噪比，在手机等语音通信设备中广泛应用。近年来，由于深度神经网络学习的语音增强技术，可有效克服传统神经网络语音消噪算法易陷于局部最优的不足，取得更好的语音消噪效果，成为语音增强技术领域的研究热点。本文针对已有深度神经网络模型泛化能力较弱、存储开销较大等问题，研究提出一种基于稀疏回归深度神经网络的语音增强算法。该算法通过在预训练阶段引入丢弃法(Dropout)和稀疏约束正则化技术改进训练模型保持预训练和调优阶段模型结构一致性，提升模型泛化能力。通过权值共享和权值量化进行网络压缩，降低存储开销。用谱减法进行后处理，有效去除稳态噪声，提高语音质量。仿真实验结果表明，改进算法可达到较高的语音性能评价指标，取得较好的语音增强效果，可满足语音增强处理要求。

关键词

语音增强，深度神经网络，正则化技术，网络压缩，谱减法

1. 引言

语音增强技术广泛应用于语音信号处理领域中，用于提高语音自然可懂度和说话人的可辨度。现实中噪声特性多样，传统的语音增强技术具有针对性，这就要求有一种可以应用于多种噪声环境的语音增强方案。

传统的单通道无监督语音增强算法如非负矩阵分解(NMF) [1]、人工神经网络(ANN) [2]等灵活性高，并行性好，但是 NMF 依赖于矩阵的构建和稀疏参数的选择，ANN 模型结构简单，构建非线性关系能力有限，模型的泛化能力比较弱，此外数据处理是基于干净语音和噪声相互独立的假设，这显然是不合理的。自 2006 年 Hinton 提出深度神经网络(DNN) [3] [4]的概念后，深度神经网络在语音识别 [5]、图像识别 [6]等领域获得成功应用。在语音增强领域，由于深度神经网络通过大数据集的离线训练，能够充分地学习带噪语音和纯净语音之间的非线性关系，提取出语音信号中的结构化信息和高层特征信息，可有效克服传统神经网络语音消噪算法易陷于局部最优的不足，取得更好的语音消噪效果，近年来成为研究热点。2015 年 Xu Y.等 [7]人提出一种基 DNN 的语音增强处理框架。在该框架中用对数功率谱作为训练 DNN 模型的特征，DNN 则作为映射函数，可以从带噪语音中预测出纯净语音。采用丢弃法和全局方差均衡法解决训练中过拟合和过平滑问题，采用动态噪声告知训练法提高对噪声环境的预见能力，取得较好的仿真结果；2016 年 Vu [8]等人提出将稀疏非负矩阵分解(SNMF)和 DNN 相结合的语音增强方法，首先用 SNMF 求取语音和噪声的激活系数，再将激活系数输入 DNN 进行学习，最后重构出语音信号。DNN 训练分两步进行，即预训练和有监督调优训练。已有算法将重点放在了调优训练阶段，实际上，建立合理的预训练模型，使 DNN 在预训练阶段获得较好的网络系数，可有效降低调优阶段的资源消耗，增强 DNN 的噪声适应性和泛化能力。另外，在实际应用中，DNN 的存储开销大、泛化能力提高等问题需要深入研究解决。2016 年 Song Han [9]等人提出基于剪枝、权值量化和霍夫曼编码的深度神经网络模型，用以解决深度神经网络存储开销大等应用问题。系统泛化能力提升主要从数据集和网络模型两方面入手 [3]。经

过近几年的研究，基于 DNN 的语音增强虽有较大的技术进展，但至今未见实际应用的报道。

本文针对已有基于 DNN 的语音增强方法存在的深度神经网络存储开销大、容易过拟合等问题，研究提出一种基于稀疏回归深度神经网络的语音增强算法。该算法通过在预训练阶段和调优训练阶段都引入丢弃法和稀疏约束正则化技术改进预训练模型，既可提升模型泛化能力，又可保持预训练和调优阶段模型结构一致性；通过权值共享和权值量化提高网络泛化能力，降低存储开销。用谱减法进行后处理，有效去除稳态噪声，提高语音质量。仿真实验结果表明，该算法可达到较高的语音性能评价指标，取得较好的语音增强效果，可满足语音增强处理要求。

2. 基于深度神经网络学习的语音增强原理

2.1. DBN-DNN 网络

由 Hinton 提出的深层神经网络框架是一种前馈神经网络。通过训练受限玻尔兹曼机(RBM) [10]来初始化网络模型，堆叠的 RBM 形成一个深度置信网络(DBN)，再在最后一层添加一个输出层，通过随机梯度下降算法逐步调优形成一个深度神经网络(DNN)，将这样的网络称为 DBN-DNN，这种模型一定程度上解决网络模型陷入局部最优的情况。整个算法流程框图如下 图 1 所示。

模型的训练分两步：第一步称为预训练(Pre-training)，即使用大量没有标注的数据通过无监督学习算法进行模型初始化；第二步称为精细调优(Fine-tuning)，通过带有标注的数据，利用传统的 BP 算法来学习模型的参数。预训练的过程是通过逐层训练玻尔兹曼机得到一个深度置信网络。

2.1.1. 预训练与精细调优

1) 预训练

预训练使用带噪语音训练受限玻尔兹曼机(RBM)，RBM 是一种基于能量的模型，其网络是一个二部图，第一层是可视层(v)，第二层是隐含层(h)，中间通过 sigmoid 激活函数连接，其可见层和隐含层的联合概率定义为：

$$p(v,h)=\frac{\exp\{-E(v,h)\}}{Z} \tag{1}$$

其中 E 是能量函数；Z 是归一化常量，定义为 $\sum_{x,y} e^{-E(x,y)}$。由于语音信号是实值分布，第一个 RBM 通常是 Gaussian-Bernoulli RBM (GRBM)，之后叠加的是 Bernoulli-Bernoulli RBM (BBRBM)。

对于 GRBM，其能量函数定义为：

$$E(v,h)=\sum_{i\epsilon\text{visible}}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{j\epsilon\text{hidden}}c_j h_j-\sum_{i,j}\frac{v_i}{\sigma_i}h_j w_{ij} \tag{2}$$

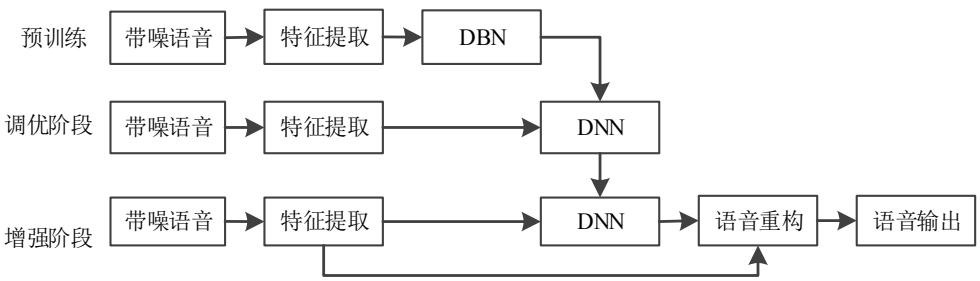


Figure 1. The theory of speech enhancement based on regressive DNN
图 1. 基于回归深度神经网络语音增强方法原理

GRBM 的可视层和隐含层的条件概率如下：

$$\begin{cases} P\left(h_j=1\middle|v\right)=\text{sigmoid}\left(\sum_i w_{ij}\frac{v_i}{\sigma_i}+b_j\right) \\ P\left(v_i=v\middle|h\right)=N\left(\sigma_i\sum_j w_{ij}h_j+a_i,\sigma_i^2\right) \end{cases} \tag{3}$$

对于 BBRBM，其能量函数定义为：

$$E\left(v,h\right)=-\sum_{i\epsilon\text{visible}}a_i v_i-\sum_{j\epsilon\text{hidden}}b_j h_j-\sum_{i,j}v_i h_j w_{ij} \tag{4}$$

BBRBM 的可视层和隐含层的条件概率如下：

$$\begin{cases} P\left(h_j=1\middle|v\right)=\text{sigmoid}\left(\sum_i w_{ij}v_i+b_j\right) \\ P\left(v_i=1\middle|h\right)=\text{sigmoid}\left(\sum_j w_{ij}h_j+a_i\right) \end{cases} \tag{5}$$

采用 Gibbs 采样进行逐层训练，Gibbs 采样的思想是：给一个训练样本 $v1$，根据公式 $P\left(h_j=1\middle|v\right)$ 求 $h1$ 中每个节点的条件概率，再根据公式 $P\left(v_i=1\middle|h\right)$ 求 $v2$ 中每个节点的条件概率，然后依次迭代，执行 k 步，此时 $P\left(v\middle|h\right)$ 的概率收敛于 $P\left(v\right)$ 的概率，过程如下 图 2 所示。

利用对比散度(CD)算法更新 RBM 参数，前一层的输出是后一层的输入，最后构成堆叠的 RBM 网络。GRBM 模型参数梯度公式为：

$$\begin{aligned} \Delta W_{ij}&=\mu\left(\left\langle\frac{1}{\sigma_i}v_i h_j\right\rangle_{\text{data}}-\left\langle\frac{1}{\sigma_i}v_i h_j\right\rangle_{\text{recon}}\right) \\ \Delta b_i&=\mu\left(\left\langle\frac{1}{\sigma_i^2}v_i\right\rangle_{\text{data}}-\left\langle\frac{1}{\sigma_i^2}v_i\right\rangle_{\text{recon}}\right) \\ \Delta c_j&=\mu\left(\left\langle h_j\right\rangle_{\text{data}}-\left\langle h_j\right\rangle_{\text{recon}}\right) \end{aligned} \tag{6}$$

BBRBM 模型参数梯度公式为：

$$\begin{aligned} \Delta W_{ij}&=\mu\left(\left\langle v_i h_j\right\rangle_{\text{data}}-\left\langle v_i h_j\right\rangle_{\text{recon}}\right) \\ \Delta b_i&=\mu\left(\left\langle v_i\right\rangle_{\text{data}}-\left\langle v_i\right\rangle_{\text{recon}}\right) \\ \Delta c_j&=\mu\left(\left\langle h_j\right\rangle_{\text{data}}-\left\langle h_j\right\rangle_{\text{recon}}\right) \end{aligned} \tag{7}$$

μ 是学习率， data 和 recon 分别表示训练数据的概率分布和重构后的概率分布，通过参数修改使模型能量减小。实际操作中，K 为 1 即可满足大部分采样需求。

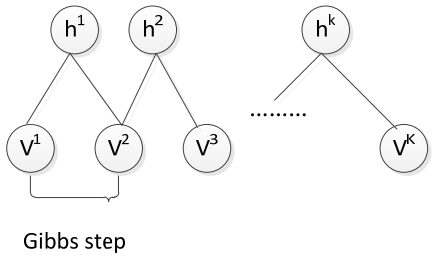


Figure 2. The process of Gibbs sample
图 2. Gibbs 采样过程

2) 精细调优：

精细调优是有监督学习过程，对DBN添加与DNN模型目标函数相关的线性回归输出层，节点个数和输入层节点个数相同。精细调优主要分三个阶段：① 前向传递；② 反馈传导；③ 修改权值。

① 前向传递：将最小批语音特征输入神经网络，将每层的激活值前向传递至输出层，获得基于最小均方误差准则的代价函数：

$$\text{Loss}=\frac{1}{N}\sum_{n=1}^{N}\sum_{d=1}^{D}\left(\hat{X}_n^d\left(W^l,b^l\right)-X_n^d\right)^2 \tag{8}$$

N 为最小批大小，D 为语音输入特征向量总维度，$\hat{X}_n^d\left(W^l,b^l\right)$、$X_n^d$ 分别为第 n 个样本第 d 维的增强语音特征和期望语音特征。

② 反馈传导：先计算输出层每个节点的残差，再向前传递获得其他隐含层的残差。然后依据残差求取每层权值偏导数。

③ 修改权值：随机梯度下降算法被用来调整网络权值：

$$W^{(l)}=W^{(l)}-\alpha\left[\left(\frac{1}{N}\Delta W^{(l)}\right)+\lambda W^{(l)}\right] \tag{9}$$

$$b^{(l)}=b^{(l)}-\alpha\left[\left(\frac{1}{N}\Delta b^{(l)}\right)\right] \tag{10}$$

α 是学习率；λ 是权值衰减参数(weight—decay)，用于控制权值幅值，防止过拟合。

如此反复执行上述步骤，直至训练完成。

2.1.2. 增强阶段

带噪语音经深度神经网络前馈，增强阶段采用均值网络获得隐含层的输出，即对所有权值乘以 $(1-p)$。获得增强语音特征后要对语音波形重构，波形重构的过程是预处理的逆过程。实际上，我们不能获取纯净语音的相位角，大量实验结果表明，人耳对语音相位信息并不敏感，因此我们用带噪语音信号的相位代替增强的语音相位。假设获取的纯净语音估计 $Y'(d)$，原始带噪语音 $X(d)$，则：

$$Y''(d)=\exp\{Y'(d)/2\}\exp\{j\angle X(d)\} \tag{11}$$

时域信号由逆傅里叶变换得到：

$$y(l)=\frac{1}{L}\sum_{d=0}^{L-1}Y''(d)\mathrm{e}^{j2\pi dl/L} \tag{12}$$

整个句子波形通过重叠相加算法(overlap and add)得到。

2.2. 提升模型泛化能力的正则化技术

为了提升模型的泛化能力，防止网络过拟合，丢弃法(Dropout) (Srivastava *et al.*, 2014) [11]可以解决训练集不匹配测试问题。Dropout是指模型训练时按一定比例 p 随机让层间某些权值不工作，不工作的节点暂时认为不是网络的一部分，但是权值要保留下来，因为下次训练时它可能又要工作，这样权值的更新不依赖于固有的层间节点的作用，是一种稀疏化思想。

稀疏技术在字典学习、自动降噪编码机上都取得了实质性效果，神经科学发现神经元具有稀疏激活性，可以将稀疏正则化技术引入深度神经网络模型 [12]，其实现原理是：引入稀疏度目标 p，惩罚项 q 鼓励实际被激活的概率。

$$q_{\text{new}}=\alpha q_{\text{old}}+(1-\alpha)q_{\text{current}} \tag{13}$$

q_{current} 是隐含层的激活概率，α 是稀疏因子。实际分布和期望分布交叉熵函数作为稀疏惩罚：

$$\text{Sparsity penalty} \propto -p\log q-(1-p)\log(1-q) \tag{14}$$

深度神经网络反向调节时对神经元输入有一个 $q-p$ 的梯度，即对(10)对 q 求偏导，这个梯度衰减用于调节权值和偏值。

3. 提升深度神经网络语音增强训练的泛化能力

深度神经网络训练的一个重要问题就是对匹配测试集效果较好，而对非匹配测试集泛化能力较弱。提高模型泛化能力一种方式是通过提高特征层面的噪声适应性，或者说是提高训练数据集规模，另外的途径则是改进模型。已有的基于 DNN 的语音增强算法注重在调优训练阶段提高模型泛化能力，主要采用文献 [11]中的丢弃法。实际上，一个理想的预训练模型使 DNN 在调优训练阶段获得较好的初始系数，从而降低调优阶段的训练时长，增强 DNN 的噪声适应性和泛化能力。

本文对基于 DNN 的语音增强训练模型进行了改进，在预训练和调优训练阶段将丢弃法和稀疏约束[12]相结合方法来提高 DNN 语音增强训练模型的泛化能力。通过丢弃法控制隐层节点激活率，然后采用稀疏正则约束弥补丢弃法的盲目性，对层间权值大小以及稀疏度进行约束，使每一层权值之和尽可能小来减少系数参数，使大部分网络系数为 0，从而最大化利用数据。既保证了预训练和调优训练阶段模型结构的一致性，又使得训练模型获得比已有方法更好的泛化能力。改进的训练模型如 图 3。

图 3(a) RBM 过程模型图，在 CD-*K* 算法中，使用丢弃法计算隐层输出，反复迭代 *K* 次后，获取隐层稀疏惩罚项后得到稀疏约束梯度，和经过 CD-*K* 算法得到的权值梯度累加来修改权值和偏值。

图 3(b)调优训练过程模型图，前向传递过程中对隐层输出做丢弃处理，计算该层的稀疏值，重复该过程直至输出层；反向传导过程中根据最小均方误差获得隐层权值梯度，并根据稀疏值计算每层稀疏惩罚项，将二者累加作为最终的权值梯度。

DNN 训练过程模型的泛化能力通常是通过对训练得到的深度神经网络语音增强算法模型，在非匹配测试集上做性能评价，评价结果见第 6 节，可以看出改进模型获得了更好的语音质量评价指标。

已有和改进的 DNN 训练过程模型调优误差如 图 4 所示。图 4(a)给出了已有的 DNN 训练调优误差随迭代次数的变化曲线，经 32 次迭代以后误差曲线逐渐收敛，趋于稳定值；图 4(b)给出了改进后的 DNN 训练调优误差随迭代次数的变化曲线，其调优误差曲线收敛特性和收敛值与原有 DNN 训练过程模型保持一致，可以得出改进模型并未降低模型精度，与下文网络压缩形成对比。

4. 降低深度神经网络存储开销

深度神经网络计算量高、存储开销大，对基于深度神经网络的语音增强嵌入式平台有很大的存储需求。为了减少存储开销，Song [9]等人提出通过剪枝、权值量化、霍夫曼编码的方法对网络压缩，通过在 AlexNet 上的分类实验获得了不错的效果。本文将权值量化的思想用于回归类问题，采用权值共享和权值量化网络压缩技术，在保证网络精度的同时降低存储空间。以 4*4 的权值矩阵为例，权值共享和权值量化的主要思想是：用 K-means 对每层权值做聚类，属于同一聚类的就用相同的权值大小，这个相同的权值就是聚类中心，其中初始的聚类中心采用线性初始化，如 图 5 上半部分。权值更新时，不用对权值调优，只需对聚类中心调优，依据权值聚类索引对其权值梯度做分组，同一组的做累加操作乘上学习率获得聚类中心梯度，再用原来的聚类中心减去这个梯度，获得新的聚类中心，如 图 5 的下半部分。离线测试时，只需要存储权值索引和聚类中心码本，大大减少了存储开销。

在回归深度神经网络前向传递阶段，权值数组分别被聚类到 2048，4196，4196，2048 个聚类中心(码字)，原来每个权值需要 32 bits 存储，现在只需要 11 bits，12 bits，12 bits，11 bits，压缩率 2.6~2.9 倍。

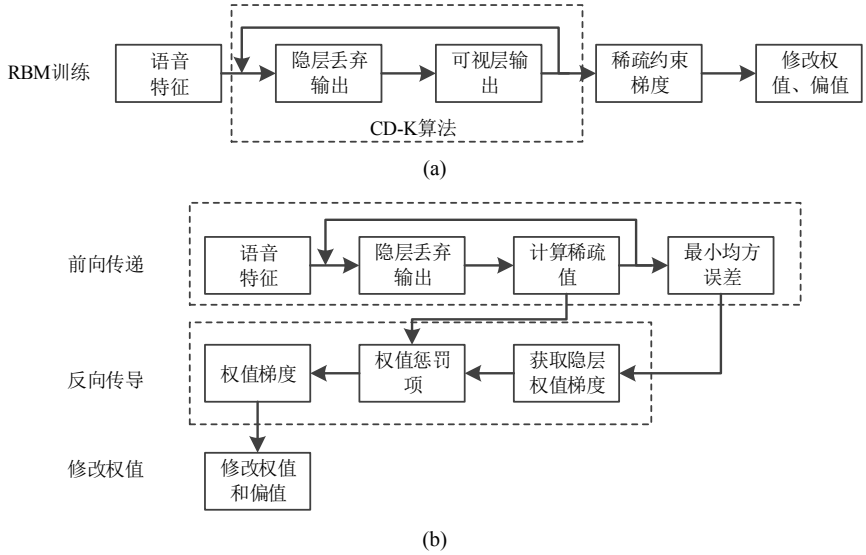


Figure 3. (a) RBM training model; (b) Fine-tuning training model
图 3. (a) RBM 训练模型；(b) 调优训练模型

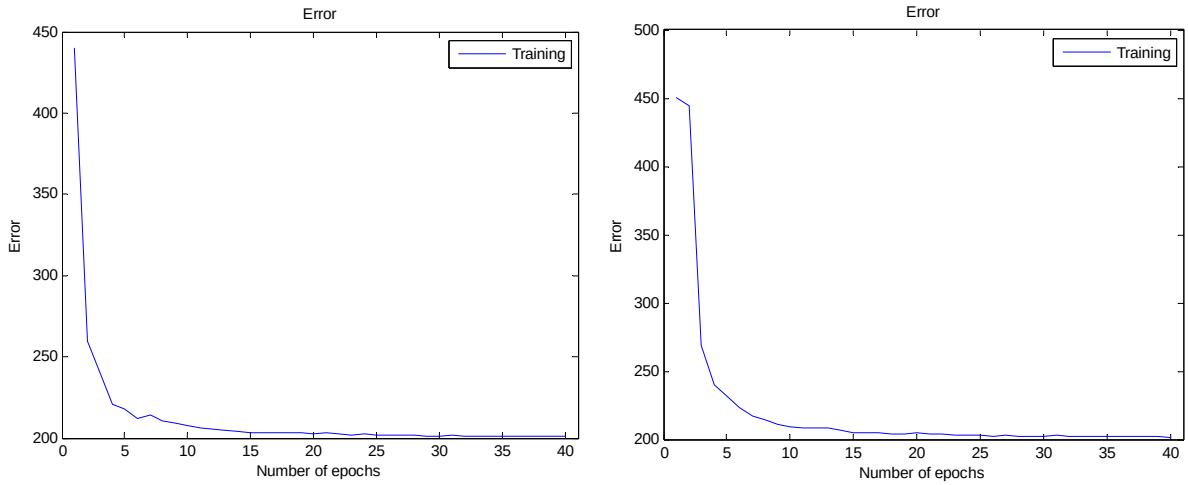


Figure 4. (a) Reconstruction error of original DNN training process; (b) Reconstruction error of improved DNN training process
图 4. (a) 已有 DNN 训练过程模型训练重构误差；(b) 改进的 DNN 训练过程模型训练重构误差

另外层间权值聚类算法互不干扰，基于多核的加速训练节约了训练时间 [13]。图 6 给出了采用权值共享和权值量化网络压缩技术后的调优训练误差曲线图，由图中可以看出精度有一定的损失，但经 32 次迭代以后逐渐收敛，误差趋于稳定。具体实验结果见第 6 节。

5. 后处理去除残留稳态噪声

Wei [14]将谱减法作为深度神经网络语音增强算法的前置环节，通过训练深度神经网络消除谱减法残留的“音乐噪声”，使得在小数据量训练情况下仍然能获得较好的语音增强效果。该算法注重消除谱减法残留的“音乐噪声”，只是构建了浅层神经网络，未充分利用深度神经网络构建带噪语音和纯净语音之间的非线性关系能力，训练模型泛化能力和噪声鲁棒性较差。另外，由原始回归深度神经网络语音增强算法框架重构的增强语音，残留了一部分的稳态噪声，该算法无法消除此类残留稳态噪声。

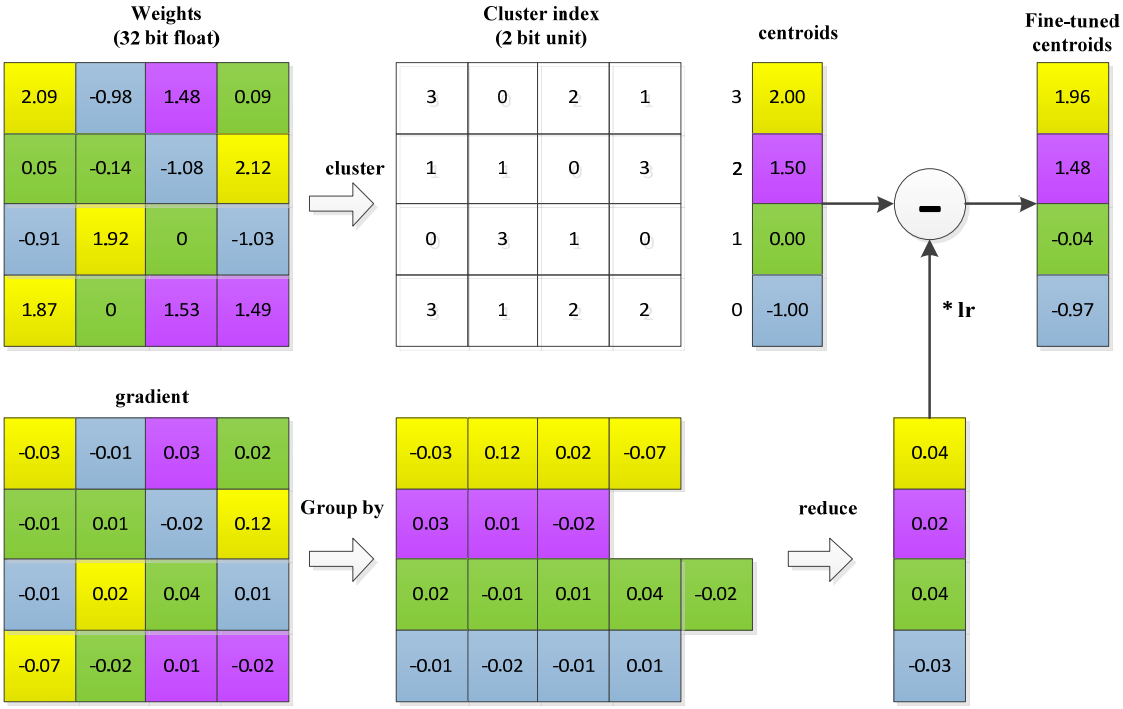


Figure 5. Weight sharing and scalar quantization and centroids fine-tuning [9]
图 5. 权值共享和量化以及聚类中心调优 [9]

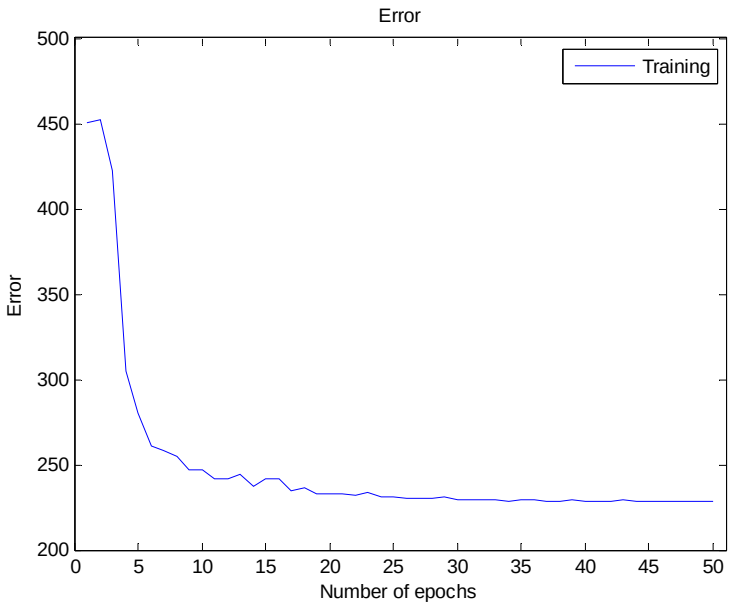


Figure 6. Reconstruct error based on compressed DNN
图 6. 引入网络压缩的深度神经网络训练重构误差

谱减法具有消除稳态噪声的良好性能。语音的前 6 帧一般代表语音所在的噪声环境，并且实验中是稳态噪声，所以只需估计前 6 帧的语音作为噪声，本文将谱减法作为深度神经网络语音增强阶段的后置环节，以后处理方式消除深度神经网络语音增强算法重构的增强语音中残留的稳态噪声。语音增强模型如下 图 7。

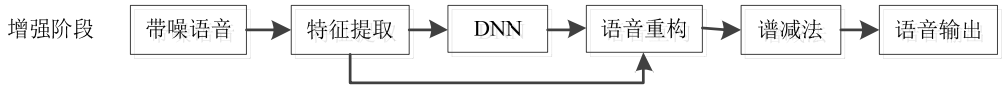


Figure 7. Speech enhancement model based on regressive DNN
图 7. 回归深度神经网络语音增强模型

带噪语音经过训练得到的深度神经网络语音增强算法模型后获得增强语音特征，经语音重构获得增强语音的时域波形，通过后处理谱减方法去除稳态噪声，最后得到增强后的语音。实验结果详见第 6 节，可以看出，后处理方式极大的提高了语音质量，取得了较好的消噪效果。

6. 实验

6.1. 实验设置

6.1.1. 样本数据处理

实验选取 NOISEX-92 噪声数据集 f16，machinegun，m109，white 和 IEEE 中 30 条纯净语音合成信噪比分别为−5~20 dB 的带噪语音 540 条作为训练数据集，180 条作为非匹配测试集。使用基于大规模数据的特征增强方法，将带噪语音、纯净语音和噪声并行输入回归深度神经网络，预测出更好的增强语音，提高了噪声鲁棒性。我们知道，语音信号的特征有多种，如响度、音频、振幅、短时能量、短时过零率、梅尔倒谱系数(MFCC)等。这里选取对数功率谱作为语音特征，主要是基于以下几点：1) 实验的研究表明，人耳对声音的强弱感觉与能量的对数成正比，大量基于频域对数功率特征的实验获得很好的效果；2) 对数可以压缩语音信号特征动态范围；3) 对数使得声学耦合的变化在特征提取中可有可无。4) 可以移除相位信息。固本文提取语音数据对数功率谱特征，对数据进行加窗—分帧处理，帧长 16 ms，帧移为 8 ms。将输入输出数据做全局 0 均值 1 方差的归一化处理，上下文帧数为 11。

6.1.2. 参数设置

训练一个节点个数分别为 1408-2048-2048-2048-1408 的三隐含层回归 DNN，隐含层激活函数为 sigmoid 函数，输出层采用线性回归模型。预训练的模型初始化参数采随机初始化权值和阈值，保证每次运行时数字的随机性。GRBM 的学习率为 0.001，学习率过大会导致迭代不收敛，太小则会收敛速度太慢，GRBM 训练迭代次数为 30 次。BBRBM 的学习率通常是 GRBM 学习率的 1~2 个数量级 [10]，这里为 0.01，迭代次数为 20 次。Dropout 比例为 0.5，稀疏因子选为 0.1。调优阶段迭代次数为 30，学习率为 0.0005，学习动量为 0.5，经过 15 次迭代以后动量增至 0.9，权值衰减系数 0.0001，学习率衰减因子 0.9。

6.2. 实验结果

6.2.1. 语音质量评价

对重构后的语音进行语音质量评价，普遍采用文献 [15]中的语音质量评价指标。主要是主观语音质量评估(PESQ)，客观方法使用分段信噪比(SegSNR)，以及频域对数谱距离(LSD)。原始语音(ORG)、基于已有深度神经网络框架(DNN)和改进的提升模型泛化能力(GENE)、权值共享和量化(WSQ)、后处理谱减(PSS)对比实验结果如下 表 1、表 2、表 3：

以上实验结果基于非匹配测试集，可以发现预训练引入正则化技术的回归深度神经网络具有更好的泛化能力，采用网络压缩方法减少了存储空间但是以牺牲网络精度为代价的，最后谱减法后处理方式极大的提高了语音质量。

6.2.2. 时域波形图

时域波形图显示了语音信号幅度随时间的变化规律，可以形象直观地看出语音增强效果。基于改进

的回归深度神经网络在低信噪比−5 dB、0 dB、5 dB 情况下原始带噪语音、增强后的语音以及纯净语音数据，用 Matlab 产生的时域波形图如下 图 8。

Table 1. The SegSNR of the original noisy and the SegSNR of enhanced speech based on improved model
表 1. 原始带噪语音和改进模型增强后的 SegSNR 值

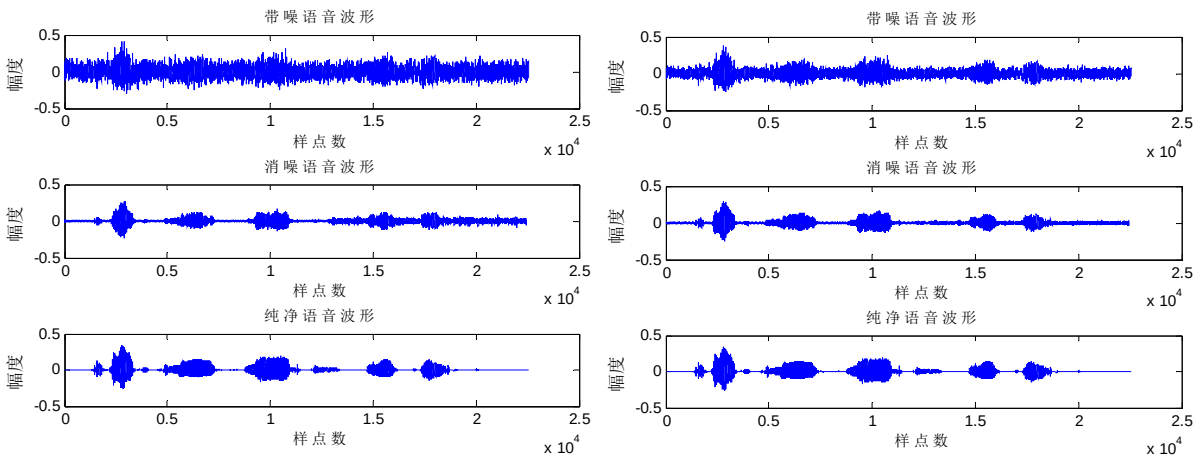
输入信噪比	SegSNR				
	ORG	DNN	GENE	WSQ	PSS
−5 dB	**−6.880263**	**−4.408206**	**−3.033468**	**−3.497555**	**−0.914351**
0 dB	**−4.588185**	**−3.455436**	**−2.401388**	**−2.366636**	**0.993789**
5 dB	**−3.588916**	**−1.015831**	**−1.869330**	**−1.513949**	**3.503478**
平均	**−5.01912**	**−2.95982**	**−2.43473**	**−2.45938**	**1.194305**

Table 2. The LSD of the original noisy and the LSD of enhanced speech based on improved model
表 2. 原始带噪语音和采用提升模型泛化能力增强后的 LSD 值

输入信噪比	LSD				
	ORG	DNN	GNENE	WSQ	PSS
−5 dB	**18.119964**	**14.975621**	**13.371705**	**13.431508**	**8.919710**
0 dB	**14.944845**	**11.957203**	**11.156676**	**11.607875**	**6.933970**
5 dB	**11.143723**	**9.374866**	**8.499927**	**8.419374**	**4.998014**
平均	**14.73618**	**12.10256**	**11.00944**	**11.15292**	**6.950565**

Table 3. The PESQ of the original noisy and the PESQ of enhanced speech based on improved model
表 3. 原始带噪语音和采用提升模型泛化能力增强后 PESQ 值

输入信噪比	PESQ				
	ORG	DNN	GENE	WSQ	PSS
−5 dB	**1.901588**	**2.093061**	**2.518918**	**2.165188**	**2.877767**
0 dB	**2.109744**	**2.116952**	**2.738987**	**2.201471**	**3.198172**
5 dB	**2.011579**	**2.180421**	**3.030972**	**2.238185**	**3.221166**
平均	**2.007637**	**2.130145**	**2.762959**	**2.201615**	**3.099035**



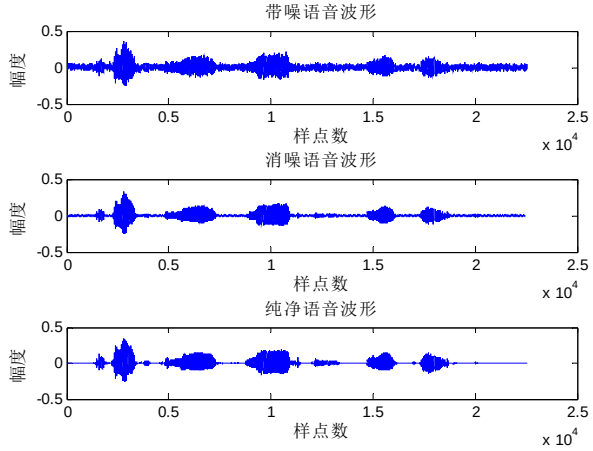


Figure 8. Time waveform of noisy (top left), enhanced speech (top right) and speech (bottom) under −5, 0, 5 SNR
图 8. −5 (上左)、0 (上右)、5 (下)信噪比情况下带噪语音、增强语音和纯净语音时域波形图

7. 结论

逻辑回归(如语音识别)和线性回归(如语音增强)有很大不同，逻辑回归只需在所有可能中找到最近似结果，而线性回归需要充分拟合期望的数据，所以基于回归深度神经网络的语音增强训练更为困难。泛化能力是深度神经网络训练的重要课题，基于现有的框架，本文在预训练阶段引入丢弃法和稀疏约束，使模型结构训练一致，减少调优阶段资源消耗，同时又能提高网络模型的泛化能力，防止过拟合。另外采用网络压缩算法减少深度神经网络存储开销，通过后处理谱减法去除稳态噪声。从实验结果可以看出，改进的回归深度神经网络模型在低信噪比情况下能够取得更好的效果。本文提出的基于稀疏回归的深度神经网络语音增强算法，主要不足之处是对高信噪比的带噪语音信号增强效果不佳，学习训练时间较长。下一步将深入研究语音增强效果更好、训练计算开销小的预训练和调优训练算法。

基金项目

本文得到国家自然科学基金重点项目(编号：61133007)资助。

参考文献 (References)

[1] Le, T.T and Mason, J.S. (1996) Artificial Neural Network for Nonlinear Time-Domain Filtering of Speech. *IEEE Proceedings on Vision, Image and Signal Processing*, **3**, 433-438.

[2] Mohammadina, N., Smaragdis, P. and Leijon, A. (2013) Supervised and Unsupervised Speech Enhancement Using Nonnegative Matrix Factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, **21**, 2140-2151. https://doi.org/10.1109/TASL.2013.2270369

[3] 时文华, 张雄伟, 张瑞昕, 韩伟. 深度学习理论及其应用专题讲座(四)[J]. 军事通信技术, 2016, 37(3): 98-104.

[4] Hinton, G.E., Osindero, S. and The, Y.W. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[5] Dahl, G.E., Yu, D., Deng, L., *et al*. (2012) Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio Speech & Language Processing*, **20**, 30-42. https://doi.org/10.1109/TASL.2011.2134090

[6] Cireşan, D., Meier, U., Gambardella, L., *et al*. (2010) Deep, Big, Simple Neural Nets for Handwritten Digit Recognition. *Neural Computation*, **22**, 3207-3220. https://doi.org/10.1162/NECO_a_00052

[7] Xu, Y., Du, J., Dai, L.R., *et al*. (2014) An Experimental Study on Speech Enhancement Based on Deep Neural Networks. *IEEE Signal Processing Letters*, **21**, 65-68. https://doi.org/10.1109/LSP.2013.2291240

[8] Vu, T.T., Bigot, B. and Chng, E.S. (2016) Combing Non-Negative Matrix Factorization and Deep Neural Network for

Speech Enhancement and Automatic Speech Recognition. In: *IEEE International Conference on Acoustic Speech and Signal Processing*, IEEE Press, Shanghai, 499-503.

[9] Han, S., Mao, H.Z. and Dally, W.J. (2015) Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman coding.

[10] Hinton, G E. (2010) A Practical Guide to Training Restricted Boltzmann Machines. *Momentum*, **9**, 599-619.

[11] Srivastava, N., Hinton, G., Krizhevsky, A., *et al*. (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, **15**, 1929-1958.

[12] Nair, V. and Hinton, G.E. (2009) 3D Object Recognition with Deep Belief Nets. *Advances in Neural Information Processing Systems* 22: *Conference on Neural Information Processing Systems* 2009, Vancouver, British Columbia, Canada, 7-10 December 2009, 1527-1554.

[13] Phan, K.T., Maul, T.H. and Vu, T.T. (2015) A Parallel Circuit Approach for Improving the Speed and Generalization Properties of Neural Networks. *International Conference on Natural Computation*, **45**, 1-7.

[14] 魏泉水. 基于深度神经网络的语音增强算法研究[D]: [硕士学位论文]. 南京: 南京大学，2016.

[15] Hu, Y. and Loizou, P.C. (2006) Evaluation of Objective Quality Measures for Speech Enhancement. *INTERSPEECH* 2006-*ICSLP*, *Ninth International Conference on Spoken Language Processing*, Pittsburgh, PA, USA, September 2006, 229-238. https://doi.org/10.1007/11939993