文章编号:1004-9037(2014)02-0171-09

深度语音信号与信息处理:研究进展与展望

戴礼荣 张仕良

(中国科学技术大学语音与语言信息处理国家工程实验室,合肥,230027)

摘要:首先对深度学习进行简要的介绍,然后就其在语音信号与信息处理研究领域的主要研究方向,包括语音识别、语音合成、语音增强的研究进展进行了详细的介绍。语音识别方向主要介绍了基于深度神经网络的语音声学建模、大数据下的模型训练和说话人自适应技术;语音合成方向主要介绍了基于深度学习模型的若干语音合成方法;语音增强方向主要介绍了基于深度神经网络的若干典型语音增强方案。最后对深度学习在语音信与信息处理领域的未来可能的研究热点进行展望。

关键词:深度学习;深度神经网络;语音识别;语音合成;语音增强

中图分类号:TP391.4 **文献标志码**:A

Deep Speech Signal and Information Processing: Research Progress and Prospect

Dai Lirong, *Zhang Shiliang*

(National Engineering Laboratory of Speech and Language Information Processing, University of Science and Technology of China, Hefei, 230027, China)

Abstract: Deep learning is briefly introduced at first. Then, a review on the research progress of deep speech signal and information processing is provided along the main research branches including speech recognition, speech synthesis and speech enhancement. For speech recognition, the acoustic modeling methods based on deep neural network(DNN), DNN model training technologies for big speech data and DNN speaker adaptation methods are introduced. For speech synthesis, several speech synthesis methods based on models in deep learning are summarized. For speech enhancement, a couple of typical DNN based speech enhancement frameworks are presented. Finally, the possible future research points of deep speech signal and information processing are discussed.

Key words: deep learning; deep neural network; speech recognition; speech synthesis; speech enhancement

引 言

近年来深度学习[1-2]逐渐成为机器学习领域的研究热点,与此同时,深度学习也在语音信号和信息处理研究领域受到重视并取得较成功的应用,有可能给语音信号与信息处理研究领域带来新的发展变化。

深度学习是指利用多层的非线性信号与信息处理技术进行有监督或者无监督以特征提取、信号转换和模式分类等为目的的机器学习方法[3]。这里的深度是指采用深层的结构[4]模型对信号和信息进行处理。传统的机器学习模型,很多属于浅层结构模型,例如:隐马尔科夫模型(Hidden Markov model, HMM)、线性或者非线性动态系统、条件随机场(Conditional random fields,CRFs)、支持向量机(Support vector machines, SVMs)、单隐层的神经网络(Neural network, NN)等。这些浅层结

基金项目:国家重点基础研究发展计划("九七三"计划)(2012CB326405)资助项目;国家自然科学基金(61273264)资助项目。

收稿日期:2013-12-28;**修订日期**:2014-01-08

构模型的共同特点是对于原始的输入信号只经过较少层次(如一层)的线性或者非线性处理以达到信号与信息处理的目的。其优点在于结构简单、易于学习,而且在数学上有比较完善的算法。但是对于一些复杂的信号,采用浅层的结构模型其表达能力具有一定的局限性,如浅层模型输入和输出中间是有限的线性或者非线性变换组合,所以并不能充分地学习到信号中复杂的结构信息。而深层结构的模型,由于其多层非线性变换的复杂性[5],具有更强的表达与建模能力,更适合于处理复杂类型的信号,如语音信号。

人类语音信号的产生和感知过程就是一个复杂的过程,而且在生物学上是具有明显的多层次或深层次处理结构[6]。所以,对于语音这种复杂信号,采用浅层结构模型对其处理显然有很大的局限性,而采用深层的结构,利用多层的非线性变换提取语音信号中的结构化信息和高层信息,是更为合理的选择。因此在近年深度学习研究领域取得一定进展的同时,很多的语音信号与信息处理研究领域的专家和学者对深度学习给予了极大的关注并开展了积极的研究,并在语音信号与信息处理一些主要研究方向取得了可喜的进展,包括:语音识别、语音合成、语音增强、语音转换、语种识别等。

1 深度学习及深度神经网络

1.1 深度学习

深度学习的概念最早起源于人工神经网络(Artificial neural network, ANN)的研究[7]。人工神经网络是机器学习与人工智能领域的一种模型[8],它的提出是为了模拟人类神经系统对事物的认知和学习能力。最早的神经网络是经典的感知器[9]。由于感知器是单层网络结构,处理能力有限,因此,多层感知器(Multilayer perceptron, MLP)被提出来。它是多个单层感知器的叠加,并采用连续非线性激活函数。由于多层感知器输入到输出之间是多层的非线性变换的组合,所以具有较强的表达能力。在深度学习的研究中,通常将具有两个以上隐层的多层感知器称为深度神经网络(Deep neural network, DNN)。DNN 模型参数可以通过误差后向传播算法(Back propagation, BP)[10]进行训练。由于 DNN 的各层激励函数均为非线性函数,模型训练中的损失函数是模型参数的非凸复杂函数,这导致当采用随机初始化模型参数时,BP 算法很容易陷入局部最优解。DNN 包含的隐层数越多这种现象越严重,从而导致 DNN 难以表现出其强大的表达和建模能力。直到 2006 年,Hinton 等[1]提出一种采用无监督的生成型模型——深度置信网络(Deep belief network, DBN)来初始化深层神经网络,一定程度上解决了上述的问题,使得其强大的学习和表达能力在机器学习中得以发挥。

近年来,深度学习理论研究及其在信号与信息处理领域应用研究均是非常活跃的研究领域。近期有很多关于深度学习理论及其在信号和信息处理领域的应用的专题研讨会,如:ICASSP2013 年关于“面向语音识别及其它应用的新型深度神经网络”的专题研讨会[11];2010,2011,2012 年 NIPS 关于“深度学习和无监督特征提取”的研讨会[12];ICML2011 年关于“语音和视觉信息处理的学习构架、表达和优化方法”[13];2012 年关于“表达学习”[14]以及 2013 年关于“深度学习应用于音频,语音及语言信息处理”[15]的研讨会。也有一些该领域的专刊,如:英文期刊 IEEE Transactions on Audio, Speech, and Language Processing, 2012 年 1 月关于“深度学习应用于语音和语言信息处理”的专刊等。

1.2 深度神经网络训练

在 Hinton 等[16]提出的深层神经网络的学习框架中,使用的是前馈型神经网络。模型的训练分为两步:首先使用大量的没有标注的数据通过无监督学习的算法来进行模型参数的初始化,这一步称为预训练(Pre-training);然后使用较少量的标注数据,利用传统的神经网络的学习算法(如 BP 算法)来学习模型的参数,这一步称为模型精细调整(Fine-tuning)。其中 Pre-training 主要是通过逐层训练受限波尔兹曼机(Restricted Boltzmann machine, RBM)得到一个生成模型 DBN;Fine-tuning 过程是对 DBN 添加一个与 DNN 模型训练目标相关的 Softmax 输出层或线性回归层,然后采用传统的 BP 算法对模型参数进行精细的调整。

1.2.1 受限波尔兹曼机

RBM 是一种包含可见层和隐含层的双层图模型,如图 1 所示。

在给定模型参数 $\theta=\{w_{ij},b_i,a_j,i=1,\cdots,M;j=1,\cdots,N\}$,可见层节点的状态 $\boldsymbol{v}$ 和隐含层节点的状态 $\boldsymbol{h}$ 时,RBM 模型定义了一个能量分布函数 $E(\boldsymbol{v},\boldsymbol{h};\theta)$。当所有节点变量服从伯努利分布时,称为伯努利 RBM,定义的能量函数

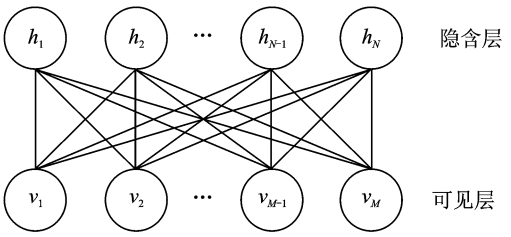

图 1　受限制玻尔兹曼机结构图

Fig. 1　Block diagram of restricted Boltzmann machine

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\sum_{i=1}^{M}\sum_{j=1}^{N}v_i w_{ij}h_j-\sum_{i=1}^{M}v_i b_i-\sum_{j=1}^{N}h_j a_j \tag{1}$$

对于可见层节点变量服从高斯分布，隐含层节点服从伯努利分布时，称为高斯-伯努利 RBM，定义的能量函数

$$E(\boldsymbol{v},\boldsymbol{h};\theta)=-\sum_{i=1}^{M}\sum_{j=1}^{N}v_i w_{ij}h_j+\frac{1}{2}\sum_{i=1}^{M}(v_i-b_i)^2-\sum_{j=1}^{N}h_j a_j \tag{2}$$

RBM 模型定义的联合分布

$$P(\boldsymbol{v},\boldsymbol{h};\theta)=\frac{\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))}{Z} \tag{3}$$

式中 $Z=\sum_{\boldsymbol{v}}\sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))$ 称为规整因子或者配分函数（Partition function）。模型关于可见层节点的状态的边缘概率为

$$P(\boldsymbol{v};\theta)=\frac{\sum_{\boldsymbol{h}}\exp(-E(\boldsymbol{v},\boldsymbol{h};\theta))}{Z} \tag{4}$$

RBM 的模型参数可以通过最大似然准则进行无监督学习得到。训练的目标函数为

$$\tilde{\theta}=\arg\max_{\theta}\log P(\boldsymbol{v};\theta) \tag{5}$$

对于目标函数求偏导，可以得到参数的更新公式为

$$\Delta w_{ij}=E_{\text{data}}(v_i h_j)-E_{\text{model}}(v_i h_j) \tag{6}$$

式中 $E_{\text{data}}(v_i h_j)$ 是关于训练集数据的可见层和隐含层状态的期望值。关于 $E_{\text{model}}(v_i h_j)$ 的直接计算很困难。在文献[17]中，Hinton 给出了通过对比散度算法（Contrastive divergence，CD）近似计算的方法。

1.2.2　DBN-DNN

通过自下而上逐层训练 RBM 的方式可以堆积得到一个生成模型，即为 DBN，如图 2 左边所示。

通常对于处理连续的信号如语音信号，最底层的 RBM 采用的是高斯-伯努利 RBM（Gaussian RBM，GRBM），对于二值的信号如二值化的图像，最低层采用的是伯努利 RBM。当训练完第 1 个 RBM 后，其隐层的输出可以用于训练第 2 个 RBM；当第 2 个 RBM 训练完成后，其隐层的输出可以用于训练第 3 个 RBM，等。通过这种逐层训练的方式最终可以得到深度置信网络 DBN，如图 2 所示。

采用无监督的预训练得到的 DBN 模型是一个概率生成模型。当把 DBN 应用于分类等任务时需要在 DBN 的顶层添加一个 Softmax 输出层，如图 2 右边图所示，形成具有初始化网络参数的 DNN（连接 Softmax 输出层网络参数除外，该层参数通常可随机初始化）。Softmax 输出层对应 DNN 输出目标值，例如在语音识别任务中可是音节、音素、音素状态等类别多选一编码值。经 DBN 初始化的 DNN 进一步通过传统的 BP 算法对网络参数进行精细的调整。通常该训练过程需要利用语音信号的标注信息，训练过程所采用的目标函数一般是最大化每个类别的后验概率，所以该过程又称为有监督的区分性训练过程（Discriminative training，DT），简称为 Fine-tuning。

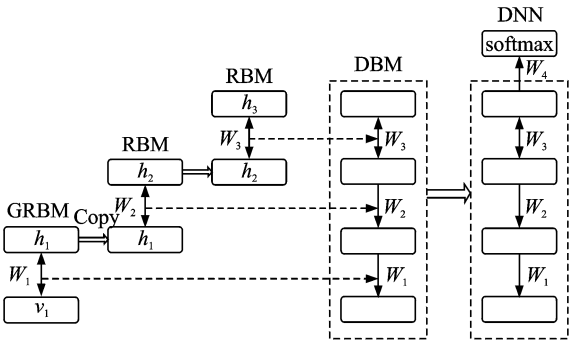

图 2　DBN-DNN 训练流程图

Fig. 2　Flow chart of DBN-DNN training

2　语音识别

本节介绍深度学习在语音识别方向的研究进展，包括 HMM-DNN 声学模型，大语音数据下 HMM-DNN 声学模型训练及 HMM-DNN 声学模型的说话人自适应。

2.1　HMM-DNN 声学模型

传统的语音识别技术普遍采用的是 HMM-GMM（Gaussian mixture model）声学模型。如引言所述，HMM-GMM 声学模型是一种浅层模型。最近，一种基于深度神经网络的称为 HMM-DNN

声学模型[16,18-21]被提出并成功应用于语音识别，并且在多种语音识别任务上一致性地取得相比于传统 HMM-GMM 声学模型较大幅度的性能提升。HMM-DNN 模型可简单看作是用 DNN 模型代替 HMM-GMM 模型中的 GMM 模型。DNN 相比于 GMM 的优势在于：(1)使用 DNN 估计 HMM 的状态的后验概率分布不需要对语音数据分布进行假设；(2)DNN 的输入特征可以是多种特征的融合，包括离散或者连续的；(3)DNN 可以利用相邻的语音帧所包含的结构信息。在文献[22]中的研究表明，DNN 的性能提升主要是归功于第 3 点。基于此，在文献[23-25]中采用 HMM-GMM-BN 框架，即把 DNN 作为一种特征提取网络，利用 DNN 提取一种称为瓶颈特征(Bottle neck feature, BN)的参数替代传统的语音特征参数，用于训练传统的 HMM-GMM。实验结果表明基于 HMM-GMM-BN 框架的语音识别系统可以取得和 HMM-DNN 可比的性能。

HMM-DNN 声学模型中 DNN 网络的激活函数通常都为 sigmoid 函数。而在文献[26-27]中提出采用一种称为 ReLUs(Rectified linear units)激活函数代替 sigmoid 激活函数。两种激活函数为

$$\text{sigmoid}:\sigma(x)=\frac{1}{1+\mathrm{e}^{-x}}$$

$$\text{ReLUs}:f(x)=\max(0,x) \tag{7}$$

采用 ReLUs 激活函数的 DNN，由于其复杂度降低，可以对其采用随机初始化。实验结果表明，使用 ReLUs 激活函数的 DNN，不但可随机初始化，而且可以取得更好的语音识别性能。

除 HMM-DNN 声学模型外，深层卷积网络(Convolutional neural network，CNN)和递归神经网络(Recurrent neural network，RNN)近期也被应用于语音识别的声学建模。文献[28-30]研究了 CNN 用于语音识别的声学建模。CNN 采用局部滤波和最大池技术，以期提取与语音谱峰频率位移无关的语音特征参数，从而提高对不同说话人的稳健性。文献[31]把 CNN 和 ReLUs 相结合用于语音的声学建模，相比于 CNN 获得了进一步的性能提升。文献[32]尝试了将 RNN 用于语音识别的声学建模，在 TIMIT 语料库上取得了目前最好的识别性能。但是 RNN 由于训练过程过于复杂，对初始化要求比较高，网络训练非常耗时，所以到目前为止还没有在大词汇量连续语音识别(Large vocabulary continuous speech recognition, LVCSR)系统中得到成功运用。

2.2 大语音数据下 HMM-DNN 声学模型训练

尽管基于 HMM-DNN 的语音识别系统在大词汇量连续语音识别任务中取得了相比于传统的 HMM-GMM 系统显著的性能提升[33-34]。但是 DNN 的训练是一个相当耗时的工作。例如，即使通过 GPU(Graphics processing unit)加速，训练一个在 1 000 h 语音数据集上的普通的 6 隐层、隐层节点数为 2 048 的 DNN，通常仍需要数周的时间。造成这种情况的潜在原因是 DNN 训练中的基本算法，随机梯度下降算法(Stochastic gradient descent, SGD)，收敛相对较慢，而且由于它本质上是一个串行的算法，使得很难对 SGD 进行并行化。因此，如何提高在大语音数据下 DNN 的训练效率，是迫切需要解决的问题。

解决 DNN 训练效率的第一个可行的方法是通过利用 DNN 模型参数的稀疏性简化模型结构来提高训练效率。文献[35]中，通过将 DNN 模型参数中 80% 的较小参数强制为 0 来减小模型大小，同时几乎没有性能损失。这个方法在减小模型大小方面很出色，但是由参数稀疏性带来的高度随机内存访问使得训练时间并没有明显减小。沿着这条路线，文献[36]中，提出将 DNN 中权重矩阵分解为两个低秩矩阵的乘积，从而达到 30%～50%的效率提升。

解决 DNN 训练效率的另外一种方法是试图通过使用多个 CPU 或者 GPU 来并行训练 DNN。文献[37-38]通过将训练数据分成许多小块到不同的机器来计算更新矩阵，实现并行训练。类似地，文献[39]在每遍迭代中，训练数据被分成 N 个不相交的子集，每个子集用来训练一个 sub-MLP，最后这些 sub-MLP 通过另一个在其他子集上训练的合并网络结合。文献[40]把这种方式扩展到了上千个 CPU 核的计算集群，通过一种异步梯度下降(Asynchronous SGD)算法训练深层神经网络。文献[41]将异步梯度下降算法应用到了多个 GPU 中。文献[42]提出了一种管道式的 BP 算法，通过将 DNN 中不同层的计算分配到不同的 GPU 单元来取得并行训练的效果，在使用 4 块 GPU 的情况下，该方法相对使用单个 GPU 训练取得了大约 3.1倍的效率提升。然而，以上这些并行训练的方法都面临着并行计算单元之间的通信开销问题，即需要收集梯度数据，重新分配更新后的模型参数以及在不同计算单元之间传递模型输出值等。不同

计算单元之间过于频繁的数据传递,成为该类方法提升训练效率的主要瓶颈,尤其是当模型较大而且并行计算单元较多时,这种现象更加明显。为此,一种新的基于状态聚类的多深层神经网络建模方法[43]被提出以实现 DNN 并行训练的目的。该方法通过将训练数据在状态层面进行聚类,得到彼此状态集不相交的子集。这种在状态层面的训练数据划分,避免了不同计算单元神经网络之间的数据传递,使得每个 DNN 可以完全独立并行训练。SWB(Switchboard)数据集上实验表明在使用 4 块 GPU,聚类数为 4 类的情况下,这种状态聚类的多 DNN 方法取得了约 4 倍的训练效率提升。

2.3　HMM-DNN 声学模型的说话人自适应

一般来说,说话人无关声学模型的识别性能比说话人相关声学模型的识别性能要低。因此,随着 HMM-DNN 模型在语音识别领域的成功应用,基于 HMM-DNN 声学模型的说话人自适应技术正受到越来越多的关注。但可惜的是,传统的 HMM-GMM 自适应技术并不能直接应用于 HMM-DNN 声学模型的自适应,需要寻找新的适合 HMM-DNN 声学模型自适应方法。

HMM-DNN 声学模型自适应最直接的方法是利用目标说话人的数据直接训练更新已收敛的说话人无关 DNN 模型参数[44-45],但由于目标说话人数据偏少及神经网络的"灾难性遗忘"特性[46],该方法非常容易出现过训练。为解决这一问题,研究人员从不改变或少量改变原有 DNN 模型参数的角度出发提出了很多行之有效的方法。文献[47-48]分别提出了线性输入网络(Linear input network,LIN)和线性隐层网络(Linear hidden network,LHN)方法。LIN 方法在输入特征和第一个隐层间增加了一个线性变换层,对于不同说话人分别估计不同的变换参数,以减少不同说话人语音差异对识别性能的影响。LHN 方法的思想与 LIN 方法类似,不同之处在于线性变换层被加在了最后一个隐层和输出层之间。文献[49]提出一种重训部分隐层单元(Retrained sub-set hidden units)的方法,在自适应时该方法首先选择隐层中的部分活跃节点,然后重新训练与这部分活跃节点相连接的权重参数,由于需要训练的参数只是全部参数的子集,因此可以防止过训练的出现。文献[50-51]则从另一个角度引入了一种基于正交厄米特多项式的隐层激活函数自适应方法,该方法改变了隐层节点的激活函数,通过对不同说话人估计不同的多项式系数来提升识别性能。以上方法主要针对隐层数较少的 HMM-DNN 模型进行了相关实验,而对于隐层数较多的 HMM-DNN 声学模型,以上自适应方法所带来的识别性能提升非常有限,甚至没有提升。因此如何在隐层数目较多的 DNN 上进行自适应是目前研究的难点。文献[52]引入 Kullback-Leibler 距离来规整权重使得模型参数的调整在一个比较小的范围内进行,不至于偏离原说话人无关 DNN 网络参数过多而引起过训练。文献[53-54]提出一种基于说话人编码(Speaker code,SC)的自适应技术;该方法通过引入所有说话人共享的自适应变换网络和每个说话人独特的编码表示来将说话人相关的声学特征变换成为说话人无关的特征。训练时采用随机梯度下降算法 SGD,并使用所有训练数据训练所有说话人共享的自适应变换网络,而当对目标说话人进行自适应时,只需要利用目标说话人部分数据根据反向传播误差生成该说话人的编码,而后与测试数据一同输入到网络中得到输出层音素状态后验概率。该方法的优点在于自适应时所需估计的参数较少,可以避免过训练,其缺点是增加了较多额外的训练时间,在 TIMIT 数据集上的实验显示该方法可以取得 10%的 PER(Phone error rate)相对错误率下降。

3　语音合成

基于 HMM 参数语音合成方法已成为当前一种主流的语音合成方法。该方法的优点是合成语音质量稳定性高,需要的存储和计算资源较小,可以方便地进行音色等方面的调整;其缺点是相对于原始语音,音质下降明显。导致音质下降的主要原因包括声码器性能的限制,声学建模不够精确,生成参数过平滑[55]。为改善基于 HMM 参数语音合成方法的合成语音质量,近年有研究人员尝试将深度学习引入语音合成技术。

Ling 等[56-57]提出 HMM-RBM 和 HMM-DBN 语音合成方法。该方法根据谱参数进行决策树状态聚类,每个状态对应的谱包络数据分别训练对应的 RBM 或 DBN;合成阶段采用 RBM 或 DBN 显层概率密度函数的模式替代高斯均值。该方法的主要优势有:对相关性很强的高维谱包络直接建模,更好地保留了频谱细节;通过 RBM/DBN 模型强大的建模能力,可以更好地拟合谱包络的分布特性,减弱了合成语音的过平滑。主客观实验表明该方法合成语音的质量优于传统的基于 HMM 参数合成方法。Zen 等[58]提出一种基于 DNN 的语音

合成方法，该方法在训练阶段，利用 DNN 取代传统基于 HMM 参数合成方法中的决策树和 GMM 模型，建立语言学特征到声学特征的映射关系；在合成阶段直接用 DNN 预测值替换传统方法的高斯均值，对应的训练数据方差替换传统方法中高斯模型的方差，进行参数生成。Kang 等[56]提出了基于 DBN 的语音合成方法，该方法针对语音合成的特点提出 MD-DBN(Multi-distribution deep belief network)。借助 MD-DBN 中不同类型的 RBM 可以同时对频谱/基频特征以及清浊信息建模，并估计音节和声学特征的联合概率分布。

4 语音增强

语音增强作为语音信号处理的一个重要分支，从 20 世纪 60～70 年代就得到了广泛的关注。语音增强的一个主要目标是从带噪语音信号中提取尽可能纯净的原始语音信号，提高语音信号的质量、清晰度和可懂度。目前非平稳噪声语音增强仍是没有很好解决的研究问题，可能的原因之一是目前语音增强方法或算法难以对语音谱在时频域上的结构化信息进行有效建模和利用。

由于深度学习中的 RBM，DNN 等模型擅长对数据中的结构化信息进行建模，而且具有从数据的低层结构化信息提取更高层的结构化信息的能力。因此，将深度学习中的 RBM，DNN 等模型应用于语音增强也是近年语音增强研究热点之一。文献[59]提出了一种基于理想二元时频掩蔽估计的语音增强方法，该方法把语音增强问题转化成用 DNN 估计理想二元时频掩蔽估计的分类问题，如图 3 所示。该方法对于低信噪比非平稳语音增强可得到高可懂度的增强语音，但语音音质损失严重。

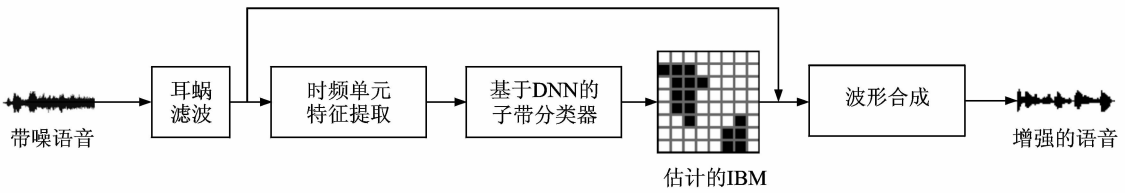

图 3 基于分类深度神经网络的语音增强

Fig. 3 Block diagram of DNN classification based speech separation

文献[60]提出了一种基于 DNN 的最小均方误差回归拟合语音增强方案，如图 4 所示。

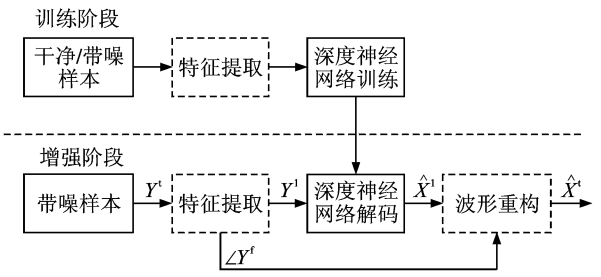

图 4 基于回归深度神经网络的语音增强

Fig. 4 Block diagram of DNN-based speech enhancement system

该方法基于对数功率谱最小均方误差准则，通过 DNN 对带噪语音和干净语音间的复杂关系进行回归拟合建模。实验表明，多帧扩展对提升语音增强质量和连续性有很大帮助，这也在某种程度上说明语音谱的结构化信息对语音增强具有重要的作用。该语音增强方案还说明大语音数据训练能保证 DNN 较充分学习到噪声语音谱和干净语音谱之间复杂的非线性关系。类似的工作包括文献[61]采用的一种堆叠式去噪自编码器(Stacked denoise autoencoder)来进行语音增强的方法。

5 总结与展望

本文就深度学习在语音信号与信息处理领域的研究进展情况进行了较为详细的介绍。首先介绍了深度学习的历史以及 DNN 训练的基本原理和算法，然后重点介绍和讨论了深度学习在语音信号与信息处理领域的语音识别、语音增强和语音合成研究方向的研究进展。相关研究进展表明，深度学习在语音信号与信息处理领域的主要课题方向均取得了较传统方法一定的优势，已成为语音信号与信息处理领域新的研究热点。

相比于传统的 HMM-GMM 语音识别声学模型，基于深度学习的 HMM-DNN 语音识别声学模型在大词汇量连续语音识别任务上已取得 20%～30%的相对性能提升。深度学习在语音识别研究

方向的进一步研究热点可能包括:首先,由于 DNN 训练过程采用的是基于梯度下降的 BP 算法,阻碍了训练的并行化。当在大语音数据上训练 DNN 模型时,所需时间在实际中有时难以忍受。目前关于如何加快模型的训练已经取得了一定的进展,但是这些技术并没有从本质上解决网络的训练耗时问题。所以在未来的研究中探索更有效的训练方法和算法将是有待进一步关注的研究问题。其次,探索如何设计训练算法使 DNN 模型参数收敛到识别性能更好的局部最优点或甚而全局最优点也是一个极其具有挑战性的研究点。再者,在模型结构上,基于 RNN 声学模型的语音识别技术仍是值得进一步深入研究的方向,由于 RNN 能直接对语音信号时序性进行建模,所以,RNN 可以完全替代 HMM-GMM 声学模型,是一种对于语音信号与信息处理非常具有潜力的模型。最后,DNN-HMM 的自适应技术仍将是活跃的研究点,基于 DNN-HMM 自适应技术的研究尚处于起步阶段,目前最有效的基于说话人编码的自适应技术仍存在诸多的有待完善之处,如说语人编码并不具有真实地表达说话人声纹信息的物理意义等。

深度学习在语音增强方面的进一步研究点可能包括:进一步提升对不包含在训练集噪声环境下的语音增强性能;语音增强 DNN 模型对噪声环境的自适应问题;及进一步将深度学习应用到多声道语音增强等。目前深度学习在语音合成的应用研究也只能算是一些初步的尝试,进一步完善基于深度学习的语音合成技术还需要进一步深入的研究。这方面可能的研究点包括:寻找更适合语音合成的深层网络结构与参数生成方法;如何更好地基于深度学习进行基频建模以及韵律建模,并将深度学习应用到统计拼接语音合成中去;以及在实际应用中如何解决采用深度神经网络完全取代传统方法所带来的运算量问题。关于深度学习在语音信号与信息处理领域的其他研究方向还包括:语种识别、说话人识别以及语音转换等。深度学习在语种识别和语音转换的研究目前已有了初步的进展,但是在说话人识别方向还未有成功运用的研究报道,因此,这方面的研究也值得关注。

致谢

感谢中国科学技术大学语音与语言信息处理国家工程实验室的博士研究生周盼、杨辰雨、薛少飞、徐勇、蒋兵,硕士研究生刘利娟,他们对深度语音信号与信息处理相关的大量研究资料进行了调研和整理。

参考文献:

[1] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets[J]. Neural computation, 2006,18(7):1527-1554.

[2] Arel I, Rose D C, Karnowski T P. Deep machine learning-A new frontier in artificial intelligence research [J]. Computational Intelligence Magazine, IEEE, 2010,5(4):13-18.

[3] Deng L. An overview of deep-structured learning for information processing[C]//Proc Asian- Pacific Signal and Information Processing-Annual Summit and Conference (APSIPA-ASC). Xi′an, China:[s. n.], 2011.

[4] Bengio Y. Learning deep architectures for AI[J]. Foundations and Trends in Machine Learning, 2009, 2(1):1-127.

[5] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002,14(8):1771-1800.

[6] Baker J, Deng L, Glass J, et al. Developments and directions in speech recognition and understanding, Part 1[J]. Signal Processing Magazine, IEEE, 2009, 26(3):75-80.

[7] Yu D, Deng L. Deep learning and its applications to signal and information processing[J]. Signal Processing Magazine, IEEE, 2011,28(1):145-154.

[8] Hopfield J J. Neural networks and physical systems with emergent collective computational abilities[J]. Proceedings of the National Academy of Sciences, 1982,79(8):2554-2558.

[9] Orbach J. Principles of neurodynamics perceptrons and the theory of brain mechanisms[J]. Archives of General Psychiatry, 1962,7(3):218.

[10] Rumelhart D E, Hinton G E, Williams R J. Learning representations by back-propagating errors[J]. Cognitive Modeling, 2002,1:213.

[11] ICASSP. New types of deep neural network learning for speech recognition and its applications[EB/OL]. http://www. icassp2013. com/SpecialSessions. asp, 2013.

[12] NIPS. Deep learning and unsupervised feature learning[EB/OL]. http://nips. cc/, 2010-10-12.

[13] ICML. Learning architectures, representations, and optimization for speech and visual information processing[EB/OL]. http://www. icml-2011. org/workshops. php, 2011.

[14] ICML. Representation learning[EB/OL]. http://icml. cc/2012/workshops/, 2012.

[15] ICML. Deep learning for audio, speech, and lan-

guage processing[EB/OL]. http://icml.cc/2013/?page_id=41, 2013.

[16] Mohamed A, Dahl G E, Hinton G. Acoustic modeling using deep belief networks[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012,20(1):14-22.

[17] Hinton G. A practical guide to training restricted Boltzmann machines[J]. Momentum, 2010,9(1):926.

[18] Mohamed A, Dahl G E, Hinton G E. Deep belief networks for phone recognition[C]//NIPS Workshop on Deep Learning for Speech Recognition and Related Applications. Hyatt Regency Vancouver, Canada: [s. n.], 2009:1-9.

[19] Sainath T N, Kingsbury B, Ramabhadran B, et al. Making deep belief networks effective for large vocabulary continuous speech recognition[C]//Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on. USA: IEEE, 2011: 30-35.

[20] Dahl G E, Yu D, Deng L, et al. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2012,20(1):30-42.

[21] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. Signal Processing Magazine, IEEE, 2012,29(6):82-97.

[22] Pan J, Liu C, Wang Z, et al. Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMS in acoustic modeling[C]//Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on. Hong Kong, China: IEEE, 2012: 301-305.

[23] Yu D, Seltzer M L. Improved bottleneck features using pretrained deep neural networks[C] //Interspeech. Florence, Italy: IEEE, 2011:237-240.

[24] Bao Y, Jiang H, Dai L, et al. Incoherent training of deep neural networks to de-correlate bottleneck features for speech recognition[C]//ICASSP. British Columbia: IEEE, 2013:6980-6984.

[25] Sainath T N, Kingsbury B, Ramabhadran B. Auto-encoder bottleneck features using deep belief networks[C]//ICASSP. Kyoto: IEEE, 2012: 4153-4156.

[26] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for lvcsr using rectified linear units and dropout[C]//ICASSP. British Columbia: IEEE, 2013:8609-8613.

[27] Zeiler M D, Ranzato M, Monga R, et al. On rectified linear units for speech processing[C]//ICASSP. British Columbia: IEEE, 2013:3517-3521.

[28] Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//ICASSP. Kyoto: IEEE, 2012:4277-4280.

[29] Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition[C]//Interspeech. Lyon: IEEE, 2013.

[30] Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]//ICASSP. British Columbia: IEEE, 2013:8614-8618.

[31] Tóth L. Convolutional deep rectifier neural nets for phone recognition[C]//Interspeech. Lyon: IEEE, 2013:1722-1726.

[32] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//ICASSP. British Columbia: IEEE, 2013:6645-6649.

[33] Seide F, Li G, Yu D. Conversational speech transcription using context-dependent deep neural networks[C]//Interspeech. Florence, Italy: IEEE, 2011:437-440.

[34] Dahl G E, Yu D, Deng L, et al. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs[C]//ICASSP. Czech: IEEE, 2011: 4688-4691.

[35] Yu D, Seide F, Li G, et al. Exploiting sparseness in deep neural networks for large vocabulary speech recognition[C]//ICASSP. Kyoto: IEEE, 2012: 4409-4412.

[36] Sainath T N, Kingsbury B, Sindhwani V, et al. Low-rank matrix factorization for deep neural network training with high-dimensional output targets[C]//ICASSP. British Columbia: IEEE, 2013:6655-6659.

[37] Kontár S. Parallel training of neural networks for speech recognition[C]//Proc 12th International Conference on Soft Computing. Zakopane: Brno University of Technology, 2006:6-10.

[38] Vesely K, Burget L, Grézl F. Parallel training of neural networks for speech recognition[C]//Text, Speech and Dialogue. Berlin: Springer Berlin Heidelberg, 2010:439-446.

[39] Park J, Diehl F, Gales M J F, et al. Efficient generation and use of MLP features for Arabic speech recognition[C]//Interspeech. Brighton: IEEE, 2009:

236-239.

[40] Le Q V, Ranzato M A, Monga R, et al. Building high-level features using large scale unsupervised learning [C]//ICASSP. British Columbia: IEEE, 2013:8595-8598.

[41] Zhang S, Zhang C, You Z, et al. Asynchronous stochastic gradient descent for DNN training[C]//ICASSP. British Columbia: IEEE, 2013:6660-6663.

[42] Chen X, Eversole A, Li G, et al. Pipelined back-propagation for context- dependent deep neural networks[C]//Interspeech. Protland: IEEE, 2012:429-433.

[43] Zhou P, Liu C, Liu Q, et al. A cluster-based multiple deep neural networks method for large vocabulary continuous speech recognition[C]//ICASSP. British Columbia: IEEE, 2013:6650-6654.

[44] Neto J, Almeida L, Hochberg M, et al. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system[J]. IEEE Trans on Speech and Audio Processing, 1995,9(2):171-185.

[45] Tüskea Z, Schlütera R, Neya H. deep hierarchical bottleneck mrasta feature for LVCSR[C]// ICASSP. British Columbia: IEEE, 2013:6970-6974.

[46] French R M. Catastrophic forgetting in connectionist networks[J]. Trends in cognitive sciences, 1999,3(4):128-135.

[47] Neto J, Almeida L, Hochberg M, et al. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system[C]//Eurospeech. [S.l.]: IEEE, 1995:2171-2174.

[48] Gemello R, Mana F, Scanzio S, et al. Linear hidden transformations for adaptation of hybrid ANN/HMM models[J]. Speech Communication, 2007,49(10):827-835.

[49] Stadermann J, Rigoll G. Two-stage speaker adaptation of hybrid tied-posterior acoustic models[C]//ICASSP. Philadelphia: IEEE, 2005:977-980.

[50] Siniscalchi S M, Li J, Lee C H. Hermitian based hidden activation functions for adaptation of hybrid hmm/ann models [C]//Interspeech. Protland: IEEE, 2012:366-369.

[51] Siniscalchi S M, Li J, Lee C H. Hermitian polynomial for speaker adaptation of connectionist speech recognition systems[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2013,21(10):2152-2161.

[52] Yu D, Yao K, Su H, et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition[C]//ICASSP. British Columbia: IEEE, 2013:7893-7897.

[53] Ossama Abdel-Hamid, Jiang H. Fast speaker adaptation of hybrid NN/ HMMmodel for speech recognition based on discriminative learning of speaker code [C]//ICASSP. British Columbia: IEEE, 2013:1942-1946.

[54] Ossama Abdel-Hamid, Jiang H. Rapid and effective speaker adaptation of convolutional neural network based models for speech recognition [C]//Interspeech. Lyon: IEEE, 2013.

[55] Zen H, Tokuda K, Black A W. Statistical parametric speech synthesis[J]. Speech Communication, 2009, 51(11):1039-1064.

[56] Kang S, Qian X, Meng H. Multi-distribution deep belief network for speech synthesis[C]//ICASSP. Columbia, USA: IEEE, 2013:8012-8016.

[57] Ling Z, Deng L, Yu D. Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis [J]. Audio, Speech, and Language Processing, IEEE Transactions on. 2013,21(10):2129-2139.

[58] Zen H, Senior A, Schuster M. Statistical parametric speech synthesis using deep neural networks[C]//ICASSP. British Columbia: IEEE, 2013:7962-7966.

[59] Wang Y, Wang D. Towards scaling up classification-based speech separation [J]. IEEE Trans Audio, Speech, Lang. Process, 2013,(99):1-23.

[60] Xu Y, Du J, Dai L, et al. An experimental study on speech enhancement based on deep neural networks [J]. IEEE Signal Processing Letters, 2014,21(1):65-68.

[61] Lu X G, Tsao Y, Matsuda S, at el. Speech enhancement based on deep denoising auto-encoder[C]//Proc Interspeech. Lyon: IEEE, 2013:436-440.

作者简介:戴礼荣(1962-),男,教授,博士生导师,研究方向:语音识别、语音合成、说话人识别等,E-mail: lrdai@ustc.edu.cnl; 张仕良(1990-),男,硕士研究生,研究方向:语音识别。