ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 32,No. 2,Mar. 2017,pp. 221－231
DOI:10.16337/j.1004-9037.2017.02.002

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于深度学习的语音识别技术现状与展望

戴礼荣　张仕良　黄智颖

(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: 首先对深度学习的发展历史以及概念进行简要的介绍。然后回顾最近几年基于深度学习的语音识别的研究进展。这一部分内容主要分成以下5点进行介绍:声学模型训练准则,基于深度学习的声学模型结构,基于深度学习的声学模型训练效率优化,基于深度学习的声学模型说话人自适应和基于深度学习的端到端语音识别。最后就基于深度学习的语音识别未来可能的研究方向进行展望。

关键词: 深度学习;深度神经网络;语音识别;说话人自适应

中图分类号: TN912.3　　**文献标志码:** A

Deep Learning for Speech Recognition: Review of State-of-the-Arts Technologies and Prospects

Dai Lirong, Zhang Shiliang, Huang Zhiying

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

Abstract: In this paper, deep learning is briefly introduced. Then, a review of the research progress of deep learning based speech recognition is presented from the following five points: Training criterions for deep learning based acoustic models, different model architectures for deep learning based speech recognition acoustic modeling, scalable and distributed optimization methods for deep learning based acoustic model training, speaker adaptation for deep learning based acoustic model, and deep leaning based end-to-end speech recognition. At the end of this paper, the future possible research points of deep learning based speech recognition are also proposed.

Key words: deep learning; deep neural network; speech recognition; speaker adaptation

引　　言

2006年,由于深度学习[1-2]理论在机器学习中初步的成功应用,开始引起人们的关注。在接下来的几年里,机器学习领域的研究热点开始逐步地转向深度学习。深度学习使用多层的非线性结构将低层特征变换成更加抽象的高层特征,以有监督或者无监督的方法对输入特征进行变换,从而提升分类或者预测的准确性[3]。深度学习模型一般是指更深层的结构[4]模型,它比传统的浅层模型拥有更多层的非线性变换,在表达和建模能力上更加强大,在复杂信号的处理上会更具优势。

基金项目: 安徽省科技重大专项(15czz02007)资助项目;国家重点研发计划(2016YFB1001300)资助项目。
收稿日期: 2017-01-05;**修订日期:** 2017-02-08

语音信号是一种非平稳的随机信号,其形成和感知的过程就是一个复杂信号的处理过程。同时,人类大脑是一种多层或者深层处理结构,对语音信号的处理是一种层次化的处理过程。浅层模型在语音信号的处理过程中会相对受限,而深层模型在一定程度上模拟人类语音信息的结构化提取过程。由此可见,深层模型比浅层模型更适合于语音信号处理。深度学习的优势吸引了很多语音信号处理领域的研究人员的关注,人们开始对其展开了积极的研究。在后来的几年里,经过研究人员的不懈努力,取得了很多突破性的进展。2009 年,深度学习首次被应用到语音识别任务[5],相比于传统的高斯混合模型-隐马尔科夫模型(Gaussian mixture model-hidden Markov model, GMM-HMM)语音识别系统获得了超过 20%的相对性能提升。此后,基于深度神经网络(Deep neural networks, DNN)的声学模型逐渐替代了 GMM 成为语音识别声学建模的主流模型,并极大地促进了语音识别技术的发展,突破了某些实际应用场景下对语音识别性能要求的瓶颈,使语音识别技术走向真正实用化。本文首先对深度学习做一个简要概述,然后重点就基于深度学习的语音识别技术现状进行较为详细的讨论,最后就基于深度学习的语音识别未来可能的研究方向进行展望。

1 深度学习简介

深度学习的概念来源于人工神经网络(Artificial neural network, ANN)。人工神经网络是机器学习与人工智能领域的一种模型[6]。它从信息处理角度对人类大脑的神经元网络进行抽象,从而达到模拟人脑认知和学习能力的目的。第 1 个人工神经元叫做阈值逻辑单元(Threshold logic unit, TLU),是由 McCulloch 和 Pitts 在 1943 年提出的,由此,开创了人工神经网络研究时代。1958 年,Rosenblatt 提出了感知器模型[7],该模型是使用线性阈值函数的一个非常重要的人工神经网络,在人工神经网络的发展过程中具有开拓性意义。另外,Rosenblatt 还通过模拟人类学习的过程,在 Hebb 学习法则的基础上,提出了一种迭代、试错的学习算法——感知器学习算法。感知器是第 1 个可学习的神经网络模型。这时感知器一般是指单层非线性变换的网络结构,它仅对线性可分问题具有分类能力,对于线性不可分问题只能做近似分类。多层感知器(Multilayer perceptron, MLP)解决了单层感知器的局限性问题,利用多个单层感知器堆叠而成,并且采用的激活函数不是线性阈值函数而是连续非线性函数。由于 MLP 是一种多层的非线性变换模型,其具有强大的表达和建模能力。同时,MLP 可以通过误差后向传播算法(Back propagation, BP)[8]进行训练。但由于 MLP 的各层激活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。并且,随着层数的增多,非凸目标函数越来越复杂,局部最小值点成倍增长,很难进行优化,使用 BP 进行算法进行网络训练时很难获得全局最优解。因此,目标函数难以优化的问题导致了 MLP 难以展现其强大的表达和建模能力。

深度置信网络(Deep belief networks, DBN)[1]是 Hinton 等学者在 2006 年提出的一种无监督的概率生成模型,用 DBN 来初始化 MLP 各层的网络参数能够解决其目标函数难以优化的问题。一般称使用 DBN 来初始化的 MLP 为 DNN。DNN 模型的训练阶段大致分为两个步骤:(1)预训练(Pre-training),利用无监督学习的算法来训练受限波尔兹曼机(Restricted Boltzmann machine, RBM),RBM 通过逐层训练并堆叠成 DBN;(2)模型精细调整(Fine-tuning),在 DBN 的最后一层上面增加一层 Softmax 层,将其用于初始化 DNN 的模型参数,然后使用带标注的数据,利用传统神经网络的学习算法(如 BP 算法)来学习 DNN 的模型参数。如此,具有很多隐层的(即深层的,一般指隐层数大于 2 乃至几百上千)的大规模模型参数(一般参数数量百万级左右或以上)的学习或训练问题在训练数据充分的条件下一定程度上得到了解决,使得其强大的学习和表达能力在机器学习中得以发挥,也直接导致机器学习领域掀起了深度学习的热潮,同时,有别于 MLP 的各种新的深层神经网络结构模型也被提出。

早期的 DNN 主要是特指前馈全连接深层神经网络(Feedforward fully-connected deep neural networks, FNN)。此后随着深度学习的发展,卷积神经网络(Convolutional neural networks, CNN)和递归神经网络(Recurrent neural networks, RNN)等网络结构在机器学习不同任务中得到应用,并且相比于 DNN 展现出各自的优势,受到越来越广泛的关注。

2 基于深度学习的语音识别技术

2.1 基于深度学习的声学模型训练准则

本节以基于 DNN-HMM 的语音识别声学模型框架讨论基于深度学习的声学模型训练准则,该框架如图 1 所示。相比于传统的基于 GMM-HMM 的语音识别框架,其最大的改变是采用 DNN 替换 GMM 模型来对语音的观察概率进行建模。DNN 相比于 GMM 的优势在于:(1)使用 DNN 估计 HMM 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 可以利用相邻语音帧所包含的结构信息。

最初主流的深层神经网络是最简单的 FNN。对于 1 个包含 L 个隐层的 FNN,其整个模型可以表示为

$$h^0 = X \tag{1}$$

$$h^l = f(W^l h^{l-1} + b^l) \qquad 1 \leqslant l \leqslant L \tag{2}$$

$$y = \mathrm{Softmax}(w^{L+1} h^L + b^{L+1}) \tag{3}$$

式中:X 表示输入的语音声学特征;$\{W^l, b^l\}$ 分别表示 l 层的连接权重和偏量;$f(\cdot)$ 表示隐层的非线性激活函数。输出层采用 Softmax 函数得到每个建模单元的后验概率输出。通过网络的输出和对应的标注可以设计相应的优化目标函数进行模型的优化。交叉熵(Cross-entropy, CE)准则经常被用作优化目标函数。CE 用来衡量目标输出概率分布和实际输出概率分布之间的相似程度,其值熵越小相似程度越高,从而模型的性能也就越好。基于 CE 准则的优化目标函数为

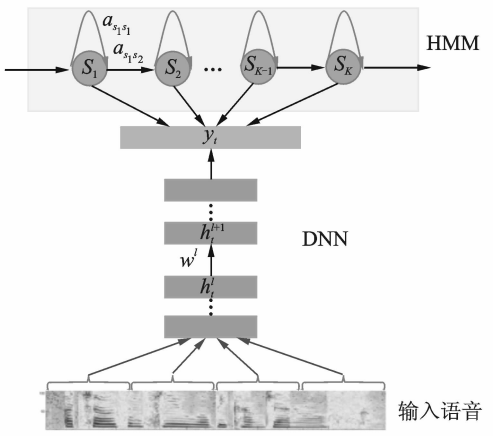

图 1 基于 DNN-HMM 的语音识别系统

Fig. 1 Illustration of hybrid DNN-HMM based speech recognition system

$$F_{\mathrm{CE}}(W) = -\sum_{r=1}^{N}\sum_{t=1}^{T_r} \log y_{rt}(s_{rt}) \tag{4}$$

式中:$y_{rt}(s)$ 表示在 t 时刻第 r 句话在状态 s 下对应 Softmax 层的输出,s_{rt} 表示 X_{rt} 对应的标注。

CE 是定义在帧级别上的优化准则,由于语音信号是一个时序信号,所以更为合适的优化准则应该是定义在整个序列上的优化准则。文献[9]中对比了不同句子级的区分性准则,包括最大互信息量(Maximum mutual information, MMI)、最小音素错误率(Minimum phone error, MPE)、状态级最小贝叶斯风险(State-level minimum Bayes risk, sMBR)和增强型最大互信息量(Boosted MMI, BMMI),用来训练 DNN-HMM 声学模型。结果表明,不同句子级区分性准则可以获得相近的性能,同时相比于 CE 准则可以获得大概 10%的相对性能提升。句子级区分性准则通过引入句子级的来自声学模型、词典和语言模型的约束来调整网络参数,这些约束以网格的形式存储,往往需要占用很大的存储空间。针对于此,文献[10]提出了一种与网格无关的 MMI 准则用于 DNN 的句子级区分性训练。

2.2 基于深度学习的语音识别声学模型的结构或类型

2009 年,DNN 首次被应用到语音识别中[5],其采用的模型为如图 1 所示的 DNN-HMM。当时的实

验是在 3 h 的 TIMIT 数据库上进行的音素识别实验。网络的输入是拼接帧的语音声学特征，利用 DNN 进行特征提取和变换，预测目标则是 61 个音素对应的 183 个 HMM 状态。实验验证了通过 Pre-training 技术可以训练包含多个隐层的神经网络，而且随着隐层数目的增加，效果也在提升。而在文献[11]中，使用绑定的音素状态作为建模单元，使得基于 DNN 的语音识别首次在大词汇量连续语音识别任务上取得突破。

在早期的 DNN-HMM 声学模型中，DNN 通常采用基于 sigmoid 的非线性激活函数，而最近的一些研究[12-14]则提出了一种更为有效的非线性激活函数，称之为修正的线性单元(Rectified linear units, ReLUs)。这两种激活函数的公式可表达为

$$\text{Sigmoid}: \sigma(x) = \frac{x}{1 + e^{-x}} \tag{5}$$

$$\text{ReLUs}: f(x) = \max(x, 0) \tag{6}$$

相关的研究结果表明，采用基于 ReLUs 激活函数的 DNN 相比基于 Sigmoid 的 DNN 不仅可以获得更好的性能，而且不需要进行 Pre-training，直接采用随机初始化即可。文献[14] 发现通过合理的参数设置，基于 ReLUs 的 DNN(RL-DNN)可以采用很大批量的随机梯度下降(Stochastic gradient descent, SGD)算法进行优化，从而可以很容易地利用多图形处理器(Graphics processing unit, GPU)进行并行化训练，而且还进一步地提出了一种绑定标量的规整技术，使得基于大批量优化的 RL-DNN 训练更加稳定。

CNN 是另一种著名的深度学习模型，在图像领域获得了广泛的应用。相比于 DNN，CNN 通过采用局部滤波和最大池化技术可以获得更加鲁棒性的特征。而语音信号的频谱特征也可以看做一幅图像，每个人的发音存在很大的差异性，例如共振峰的频带在语谱图上就存在不同。所以通过 CNN，有效地去除这种差异性将有利于语音的声学建模。最近的一些工作[15-17]也表明，基于 CNN 的语音声学模型相比于 DNN 可以获得更好的性能。文献[17]中通过采用 2 层 CNN，再添加 4 层 DNN 的结构，相比于 6 层 DNN，在大词汇量连续语音识别任务上可以获得相对 3%～5%的性能提升。文献[18]提出将 CNN 和 RL-DNN 相结合，可以获得进一步的性能提升。虽然 CNN 被应用到语音识别中已有很长一段时间，但是都只是把 CNN 当作一种鲁棒性特征提取的工具，所以一般只是在底层使用 1～2 层的 CNN 层，然后高层再采用其他神经网络结构进行建模。而在最近的一些研究中，CNN 在语音识别得到了新的应用，相比于之前的工作，最大的不同是使用了非常深层的 CNN 结构(Deep convolutional neural networks, DCNN)[19-22]，包含 10 层甚至更多的卷积层。研究结果也表明深层的 CNN 往往可以获得更好的性能。

语音信号是一种非平稳时序信号，如何有效地对长时时序动态相关性进行建模至关重要。由于 DNN 和 CNN 对输入信号的感受视野相对固定，所以对于长时时序动态相关性的建模存在一定的缺陷。RNN 通过在隐层添加一些反馈连接，使得模型具有一定的动态记忆能力，对长时时序动态相关性具有较好的建模能力。文献[23]最早尝试将 RNN 用于语音识别的声学建模，在 TIMIT 语料库上取得了当时最好的识别性能。由于简单的 RNN 会存在梯度消失问题，一个改进的模型是基于长短时记忆单元(Long-short term memory, LSTM)[24]的递归结构。文献[25]使用 LSTM-HMM 在大数据库上获得了成功。此后大量的研究人员转移到基于 LSTM 的语音声学建模的研究中。基于双向 LSTM 的语音声学模型系统可以获得相比基于 DNN 系统超过 20%的相对性能提升。文献[26]结合 CNN，DNN 以及 LSTM 各自的优点，提出了 CLDNN 结构用于语音的声学建模。

虽然 LSTM 相比于 DNN 在模型性能上有极大的优势，但是训练 LSTM 需要使用沿时间展开的反向传播算法(Back propagation through time, BPTT)算法，会导致训练不稳定，而且训练相比于 DNN 会

更加耗时。因此如何让前馈型的神经网络也能像 LSTM 一样具有长时时序动态相关性的建模能力是一个研究点。文献[27]中提出将 RNN 沿着时间展开,可以在训练速度和 DNN 可比的情况下获得更好的性能。但是进一步的把 LSTM 结构沿时间展开就比较困难。文献[28]中提出的时间延时神经网络(Time delay neural networks, TDNN)是另外一种可以对长时时序动态相关性进行建模的前馈型神经网络。在最近的工作[29-30]中,TDNN 被应用到 LVCSR 任务上,性能上略差于 LSTM。另外,文献[31~32]中提出的一种前馈序列记忆网络(Feedforward sequential memory networks, FSMN),可以用更小的模型和更快的速度,取得比 LSTM 更好的性能。

2.3 基于深度学习的语音识别声学模型训练效率优化

目前 DNN-HMM 取代 GMM-HMM 成为语音识别的主流声学模型,获得了显著的性能提升[33-34],但是基于 DNN 的语音识别声学模型的训练是一个相当耗时的过程。随着大数据时代的到来,可以获得的语音数据也越来越多,因此基于大数据的深度学习语音识别声学模型的训练效率是一个迫切需要解决的问题。这方面的工作大致可以分成两类:(1)如何利用神经网络的特性,设计结构更加简洁的网络,从而加速网络的训练;(2)如何利用多 GPU 进行并行化训练。

DNN 通过增加隐层以及隐层节点的数目,可以获得很强的模型表达能力。例如在语音识别声学建模任务中一个常用的 DNN 网络结构包含 6 个隐层,每个隐层包含 2 048 个节点。这样的网络具有很强的冗余性,这可以通过训练收敛后网络权重的稀疏性得到验证。文献[35]的研究表明,DNN 中有大量的权重阈值小于 0.1,这些很小的权重可以强制置为 0,不会对网络性能产生很大的影响。相关实验结果表明,可以将网络中 80%的权重置 0,从而几乎不损失性能。该做法可以有效地减小模型的参数,但是并不能加快训练速度。文献[36]则进一步分析了 DNN 的稀疏特性,发现越往高层,权重的稀疏性越强,因而提出了一种隐层节点递减的结构,可有效地减少网络的参数,同时带来接近一倍的训练效率提升。根据稀疏矩阵不满秩的特性,文献[37]引入了矩阵低秩分解的方法,将原本的 DNN 权重矩阵分解成两个小矩阵相乘的形式,从而可以将网络的参数减少 30%~50%。文献[25]进一步地将矩阵低秩分解和 LSTM 相结合,称为所谓的 LSTMP 结构,也可以大幅度地加快 LSTM 的训练。文献[38]提出一些节点剪枝的方法,可以大致将 50%的节点从网络中去除,而基本不对性能造成损失。

由于单 GPU 的计算能力有限,很难处理海量数据,所以探究如何进行多 CPU 或者 GPU 并行计算是一个热门研究点。关于这方面的研究首先是分数据的策略,在文献[39~40]中提出将训练数据分成很多小份,然后每份在一个单独的 GPU 上进行运算,将得到的梯度求平均去更新模型。这种方法受限于 SGD 训练时必须采用小批量(mini-batch)不同机器间的频繁交互会导致通信代价很高,从而没法带来很大的训练速度提升。文献[41]提出将原始数据平均分成 N 份,然后每份数据利用一台机器单独训练一个子网络,每次迭代后将这些子网络求平均得到一个总模型,再分到各个机器上进行训练。这种方式可以有效避免机器之间的通讯代价,但是会导致较大的性能损失。由于机器之间的通讯代价是并行计算的一个瓶颈,文献[42]提出异步随机梯度下降(Asynchronous stochastic gradient descent, ASGD),可以有效地掩蔽通讯代价,利用包含数千个 CPU 的集群来进行 DNN 的并行训练。而文献[43]将这种方法扩展到了 GPU 上,利用多 GPU 进行并行化训练,节约了设备成本。虽然 ASGD 可以有效地掩蔽不同计算单元之间的通讯代价,但其扩展性却比较差,当想进一步扩展到更多 GPU 时,往往会导致明显的性能损失。针对该问题,文献[44]提出了块模型更新过滤(Blockwise model-update filtering, BMUF)算法,通过引入梯度动量的方式,不仅可以减少多 GPU 之间的交互次数,而且基本实现了训练随着 GPU 数目的增加而线性加速。

2.4 基于深度学习的语音识别声学模型的说话人自适应

一般来说,说话人无关模型在语音识别性能上要劣于说话人相关模型,但说话人相关模型需要每一特定说话人的大量语音用于训练,实际应用不具可行性。语音识别声学模型的说话人自适应一般可使识别性能优于说话人无关模型,并且所需的特定说话人的数据量远低于说话人相关模型的数据量要求。

传统的基于 GMM-HMM 的语音识别声学模型的说话人自适应技术已经有了很多较成熟的技术。其大致可以分成两类,一种是模型域的说话人自适应,另一种是特征域的说话人自适应。模型域的说话人自适应通过对训练好的通用模型进行自适应,从而得到一个说话人相关的模型;特征域自适应是通过自适应得到说话人无关的特征。

基于深度学习的语音识别声学模型的说话人自适应是近年来语音识别声学模型的说话人自适应的研究热点。文献[45]将传统的 GMM-HMM 声学模型的约束最大似然线性回归(Constrained maximum likelihood linear regression,CMLLR)自适应算法应用到了基于深度学习的声学模型自适应,提出了特征空间区分性线性回归(Feature-space discriminative linear regression, FDLR)的特征域说话人自适应方法。基于深度学习的语音识别声学模型的自适应研究主要集中在模型域自适应,主要可以归纳为如下几种:

(1)基于说话人特征的自适应方法。其主要思路是通过一种包含说话人信息并且能够区分不同说话人的特征矢量,实现对基于深度学习的语音识别声学模型的自适应。鉴别性矢量(identity vector, i-vector)是一种包含说话人信息和信道信息的矢量,基于 i-vector 的说话人自适应方法利用每个说话人的语料提取对应的 i-vector,然后将 i-vector 同声学特征相融合[46-47],从而实现模型域上的说话人自适应。文献[48]通过提取瓶颈(Bottleneck, BN)特征,将该 BN 特征以类似的方式融入到基于深度学习的语音识别声学模型中。另外,基于说话人编码(Speaker code)的说话人自适应[49-51]通过生成说话人特定的 Speaker code,然后将 Speaker code 输入至基于深度学习的语音识别声学模型的所有层进行模型域的说话人自适应。

(2)基于模型正则化的说话人自适应方法。该方法直接用特定说话人的少量数据调整一个说话人无关模型,并通过模型正则化避免易产生的模型过拟合问题。文献[52]提出了一种基于 KL 散度(Kullback-Leibler divergence)的说话人自适应方法,该方法通过 KL 散度约束自适应后模型的后验概率分布不至于偏离说话人无关模型的分布太远来实现模型的规整。另外,文献[53~54]通过增加对上、下文无关(Context independent, CI)的 HMM 状态类别的自适应学习,一定程度上实现了对上下文相关(Context dependent, CD)声学模型的规整。

(3)基于线性变换的说话人自适应方法。该方法在原始的说话人无关的基于深度学习的语音识别声学模型中插入一个或若干线性变换层,该变换层通过自适应训练后起到将说话人无关模型转换为特定说话人模型的作用。文献[55~57]采用直接插入线性变换层的方式分别在输入层、隐层和输出层进行线性变换。但是,对于这种直接插入线性变换层的方式来说,每个说话人所需要训练的参数量太大,易导致模型过拟合。文献[58~60]将原始的说话人无关模型进行 SVD 去冗余处理,然后再在 SVD 层中插入线性变换层,该线性变换层的参数量比直接插入的方式要少很多,这样大大减少了自适应阶段所需要训练的参数量,在一定程度上缓解过拟合问题。另外,根据不同说话人在隐层单元的激活程度大小不一样,文献[61]提出了基于线性隐层单元分布(Linear hidden unit contribution, LHUC)的说话人自适应方法,该方法针对每个说话人学习其对应的隐层单元分布,取得了比较好的说话人自适应效果。

(4)基于多基融合的说话人自适应方法。该方法在声学模型空间建立一组基,这组基可以是基于深度学习的语音识别声学模型[62],也可以是对应的深层声学模型网络的联结权重[63-64]。再利用每个说话

人的语音数据通过训练来获得对应的插值矢量,通过该插值矢量来对基进行插值,从而获得特定说话人的声学模型。

(5)基于激活函数的说话人自适应方法。该方法认为每个说话人在深层声学模型网络节点上的激活程度不一样,因而可以对每个说话人构造一组特定的激活函数实现说话人自适应,该激活函数可以利用赫尔米特正交函数[65],或者是参数化的 Sigmoid 和参数化的 ReLUs 函数[66]来构建。

说话人无关模型只需要训练和测试两个阶段,而说话人自适应模型一般需要训练、自适应和测试 3 个阶段。因此,在实际应用中,自适应阶段会影响语音识别模型的实时性。基于 i-vector 的说话人特征自适应虽然不需要自适应阶段、在实时性上能够满足实际要求,但是,从较短、带噪的句子提取的 i-vector 往往不能够非常好地表达说话人信息,因而会出现自适应后性能提升不明显甚至性能变差的情况。另外,实验表明,现有的说话人自适应方法大多会出现少部分人经过自适应后性能变差的情况,也是值得注意需要解决的问题。

2.5 基于深度学习的端到端语音识别

以上所讨论的基于深度学习的语音识别声学模型建模技术,在模型训练上仍依赖于传统的基于 GMM-HMM 语音识别技术,声学模型框架上仍采用类似 GMM-HMM 的语音识别模型框架。如:声学模型和语言模型的训练是独立的,通过后端的解码将两者进行融合。声学模型的训练过程中首先需要利用 HMM 进行对齐得到训练数据帧级别的标注,所以整个模型的训练分成很多个阶段。针对此问题,基于深度学习的语音识别技术近期的一个研究热点是如何进行端到端的语音识别。

文献[67~69]提出采用连续时序分类 (Connectionist temporal classification, CTC)[70]和 LSTM 结合的声学模型,该模型直接对一句语音的音素序列或者绑定的音素(Context-dependent phone, CD-Phone)序列与对应的语音特征序列进行序列层面建模,不需要利用 HMM 进行强制对齐得到帧级别的标注,可以取得相比于传统 LSTM-HMM 声学模型更好的性能。

语音识别声学模型的输入往往是根据信号处理和人耳的听觉特性设计的声学特征,例如 MFCC, PLP 和 FBK 等。但是这些特征的提取和声学模型的训练相互独立,而且声学特征的提取准则和声学模型优化准则之间存在一定的不匹配性,所以让神经网络直接从原始的语音波形中去学习特征更为合理。文献[71]提出一种直接输入对数功率谱,利用网络学习得到 FBK 特征的方法,这样 FBK 特征的提取和后端的声学模型就可以联合优化,可以获得相比于 FBK 特征更好的性能。文献[72]直接使用原始的语音波形进行输入,利用 CNN 学习特征,后端采用 LSTM 和 DNN 进行声学建模,可以取得和 FBK 特征可比的性能。文献[73~75]将基于时域波形的单通道语音声学建模进一步推广到多通道上,实验结果表明采用时域卷积层可以利用多通道的语音波形进行波束形成和空域滤波,并且相比传统的基于阵列信号处理的多通道语音增强技术,例如 Delay-and-sum 和 Filter-and-sum,可以获得更好的性能。文献[76]对上述方法进行进一步的改进,将时域卷积的滤波器系数变成和输入相关的一组系数,从而使得模型的滤波具有一定的自适应效果。

端到端的语音识别的另外一个方法是基于编码和解码(encoder-decoder)模型以及注意(attention)模型[77],直接实现从语音声学特征序列到最终句子级的音素序列、字符序列或词序列的输出。该方法同样不需要进行分帧以及得到帧级别的标注。文献[78~79]使用基于 Attention 的 Encoder-decoder 模型在 TIMIT 数据库上取得了和主流混合神经网络以及 HMM 模型相当的性能。但是在大词汇量连续语音识别任务上,该方法的性能[80-81]目前和最好的语音识别系统的性能还有一定的差距。

3 结束语

当前基于深度学习的语音识别技术相比于传统 GMM-HMM 技术已经取得了很大的进展。在安静

环境下目前基于深度学习的语音识别技术已经达到了实用化水平。但是在一些特殊环境下，比如噪声干扰比较强或者是在远场情况下，语音识别系统的性能依然没有达到实用化要求。目前远场识别的错误率是近场的2倍左右，所以解决远场以及强噪声干扰情况下的语音识别是目前有待进一步研究的问题。这方面目前的主要做法是将语音识别和麦克风阵列相结合。通过阵列信号处理技术，将多通道语音进行增强，然后后端再利用深度学习的方法进行声学建模。显然这种方案有待进一步优化，如：如何将阵列信号处理技术和深度学习方法相结合，利用阵列信号处理的知识指导深度神经网络的结构设计，从而直接从多通道语音信号中学习多通道语音增强方法然后和后端声学模型联合优化。目前较为成熟的基于深度学习的语音识别技术在语音识别系统训练流程上还是比较复杂，需要很多中间步骤，例如需要进行强制对齐得到标注，需要根据词典训练语言模型及需要将声学模型和语言模型进行联合解码，所以探究更为简单的、高识别性能的端到端语音识别技术是未来一个值得关注的研究方向。

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 [Research Frontier][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] Mohamed A, Dahl G, Hinton G. Deep belief networks for phone recognition[C]// Nips Workshop on Deep Learning for Speech Recognition and Related Applications. Whistler, BC, Canada: MIT Press, 2009: 39.

[6] 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.

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

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

[9] Vesely K, Ghoshal A, Burget L, et al. Sequence-discriminative training of deep neural networks[C]// Interspeech. Lyon, France: IEEE, 2013: 2345-2349.

[10] Povey D, Peddinti V, Galvez D, et al. Purely sequence-trained neural networks for ASR based on lattice-free MMI[C]// Interspeech. San Francisco, California: IEEE, 2016: 2751-2755.

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

[12] Dahl G E, Sainath T N, Hinton G E. Improving deep neural networks for LVCSR using rectified linear units and dropout[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 8609-8613.

[13] Zeiler M D, Ranzato M, Monga R, et al. On rectified linear units for speech processing[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 3517-3521.

[14] Zhang S, Jiang H, Wei S, et al. Rectified linear neural networks with tied-scalar regularization for LVCSR[C]// Interspeech. Dresden, Germany: IEEE, 2015: 2635-2639.

[15] Abdel-Hamid O, Mohamed A, Jiang H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012: 4277-4280.

[16] Abdel-Hamid O, Deng L, Yu D. Exploring convolutional neural network structures and optimization techniques for speech recognition[C]// Interspeech. Lyon, France: IEEE, 2013: 3366-3370.

[17] Sainath T N, Mohamed A, Kingsbury B, et al. Deep convolutional neural networks for LVCSR[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada: IEEE, 2013: 8614-8618.

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

[19] Sercu T, Puhrsch C, Kingsbury B, et al. Very deep multilingual convolutional neural networks for LVCSR[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 4955-4959.
[20] Sainath T N, Kingsbury B, Saon G, et al. Deep convolutional neural networks for large-scale speech tasks[J]. Neural Networks, 2015, 64: 39-48.
[21] Qian Y, Bi M, Tan T, et al. Very deep convolutional neural networks for noise robust speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, 24(12): 2263-2276.
[22] Yu D, Xiong W, Droppo J, et al. Deep convolutional neural networks with layer-wise context expansion and attention[C]// Interspeech. San Francisco, California: IEEE, 2016: 17-21.
[23] Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, B C, Canada: IEEE, 2013: 6645-6649.
[24] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.
[25] Sak H, Senior A W, Beaufays F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]// Interspeech. Singapore: IEEE, 2014: 338-342.
[26] Sainath T N, Vinyals O, Senior A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4580-4584.
[27] Saon G, Soltau H, Emami A, et al. Unfolded recurrent neural networks for speech recognition[C]// Interspeech. Singapore: IEEE, 2014: 343-347.
[28] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1989, 37(3): 328-339.
[29] Peddinti V, Povey D, Khudanpur S. A time delay neural network architecture for efficient modeling of long temporal contexts[C]// Interspeech. Dresden, Germany: IEEE, 2015: 3214-3218.
[30] Peddinti V, Chen G, Povey D, et al. Reverberation robust acoustic modeling using i-vectors with time delay neural networks [C]// Interspeech. Dresden, Germany: IEEE, 2015: 2440-2444.
[31] Zhang S, Liu C, Jiang H, et al. Nonrecurrent neural structure for long-term dependence[J]. IEEE/ACM Transactions on Audio, Speech and Language Processing, 2017,25(4):871-884.
[32] Zhang S, Jiang H, Xiong S, et al. Compact feedforward sequential memory networks for large vocabulary continuous speech recognition[C]// Interspeech. San Francisco, California: IEEE, 2016: 3389-3393.
[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]// International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic: 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]// International Conference on Acoustics, Speech and Signal Processing. Kyoto, Japan: IEEE, 2012: 4409-4412.
[36] Zhang S, Bao Y, Zhou P, et al. Improving deep neural networks for LVCSR using dropout and shrinking structure[C]// International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 6849-6853.
[37] Sainath T N, Kingsbury B, Sindhwani V, et al. Low-rank matrix factorization for deep neural network training with high-dimensional output targets[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 6655-6659.
[38] He T, Fan Y, Qian Y, et al. Reshaping deep neural network for fast decoding by node-pruning[C]// International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 245-249.
[39] Kontár S. Parallel training of neural networks for speech recognition[C]// Proc 12th International Conference on Soft Computing. Brno, Crech Republic: Brno University of Technology, 2006:6-10.
[40] Vesely K, Burget L, Grézl F. Parallel training of neural networks for speech recognition[C]// Text, Speech and Dialogue. Brno, Crech Republic: Springer-Verlag, 2010:439-446.
[41] Park J, Diehl F, Gales M J F, et al. Efficient generation and use of MLP features for Arabic speech recognition[C]// Interspeech. Brighton, United Kingdom: IEEE, 2009: 236-239.
[42] Le Q V. Building high-level features using large scale unsupervised learning[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, BC, Canada: IEEE, 2013: 8595-8598.
[43] Zhang S, Zhang C, You Z, et al. Asynchronous stochastic gradient descent for DNN training[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, B C, Canada: IEEE, 2013: 6660-6663.

[44] Chen K, Huo Q. Scalable training of deep learning machines by incremental block training with intra-block parallel optimization and blockwise model-update filtering[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 5880-5884.

[45] Seide F, Li G, Chen X, et al. Feature engineering in context-dependent deep neural networks for conversational speech transcription[C]// Automatic Speech Recognition and Understanding. Waikoloa, Hawaii: IEEE, 2011: 24-29.

[46] Saon G, Soltau H, Nahamoo D, et al. Speaker adaptation of neural network acoustic models using i-vectors[C]// Automatic Speech Recognition and Understanding. Olomouc, Czech Republic: IEEE, 2013: 55-59.

[47] Miao Y, Zhang H, Metze F. Speaker adaptive training of deep neural network acoustic models using i-vectors[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2015, 23(11): 1938-1949.

[48] Huang H, Sim K C. An investigation of augmenting speaker representations to improve speaker normalisation for DNN-based speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4610-4613.

[49] Abdel-Hamid O, Jiang H. Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, B C, Canada: IEEE, 2013: 7942-7946.

[50] Xue S, Abdel-Hamid O, Jiang H, et al. Fast adaptation of deep neural network based on discriminant codes for speech recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2014, 22(12): 1713-1725.

[51] Huang Z, Tang J, Xue S, et al. Speaker adaptation of RNN-BLSTM for speech recognition based on speaker code[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 5305-5309.

[52] Yu D, Yao K, Su H, et al. KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. Vancouver, B C, Canada: IEEE, 2013: 7893-7897.

[53] Huang Z, Li J, Siniscalchi S M, et al. Rapid adaptation for deep neural networks through multi-task learning[C]// Interspeech. Dresden, Germany: IEEE, 2015: 2329-9290.

[54] Price R, Iso K, Shinoda K. Speaker adaptation of deep neural networks using a hierarchy of output layers[C]// Spoken Language Technology Workshop (SLT). South Lake Tahoe, N V, USA: IEEE, 2014: 153-158.

[55] Neto J, Almeida L, Hochberg M, et al. Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system[C]// European Conference on Speech Communication and Technology(EUROSPEECH). Madrid, Spain: ISCA, 1995:2171-2174.

[56] 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.

[57] Li B, Sim K C. Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems[C]// Interspeech. Makuhari, Chiba, Japan: IEEE, 2010: 526-529.

[58] Xue J, Li J, Yu D, et al. Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network[C]// International Conference on Acoustics, Speech and Signal Processing. Florence, Italy: IEEE, 2014: 6359-6363.

[59] Kumar K, Liu C, Yao K, et al. Intermediate-layer DNN adaptation for offline and session-based iterative speaker adaptation [C]// Interspeech. Dresden, Germany: IEEE, 2015: 1091-1095.

[60] Zhao Y, Li J, Gong Y. Low-rank plus diagonal adaptation for deep neural networks[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 5005-5009.

[61] Swietojanski P, Renals S. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models[C]// Spoken Language Technology Workshop (SLT). South Lake Tahoe, N V, USA: IEEE, 2014: 171-176.

[62] Wu C, Gales M J F. Multi-basis adaptive neural network for rapid adaptation in speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4315-4319.

[63] Delcroix M, Kinoshita K, Hori T, et al. Context adaptive deep neural networks for fast acoustic model adaptation[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4535-4539.

[64] Tan T, Qian Y, Yin M, et al. Cluster adaptive training for deep neural network[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4325-4329.

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

[66] Zhang C, Woodland P C. DNN speaker adaptation using parameterised sigmoid and ReLU hidden activation functions[C]//

International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 5300-5304.
[67] Graves A, Jaitly N. Towards end-to-end speech recognition with recurrent neural networks[C]// International Conference on Machine Learning. Beijing, China: ACM, 2014, 14: 1764-1772.
[68] Sak H, Senior A, Rao K, et al. Learning acoustic frame labeling for speech recognition with recurrent neural networks[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4280-4284.
[69] Graves A, Jaitly N, Mohamed A. Hybrid speech recognition with deep bidirectional LSTM[C]// Automatic Speech Recognition and Understanding (ASRU). Olomouc, Crech Republic: IEEE, 2013: 273-278.
[70] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]// International Conference on Machine Learning. Pittsburgh, Pennsylvania, USA: ACM, 2006: 369-376.
[71] Sainath T N, Kingsbury B, Mohamed A, et al. Learning filter banks within a deep neural network framework[C]// Automatic Speech Recognition and Understanding. Olomouc, Czech Republic: IEEE, 2013: 297-302.
[72] Sainath T N, Weiss R J, Senior A, et al. Learning the speech front-end with raw waveform CLDNNS[C]// Interspeech. Dresden, Germany: IEEE, 2015: 1-5.
[73] Hoshen Y, Weiss R J, Wilson K W. Speech acoustic modeling from raw multichannel waveforms[C]// International Conference on Acoustics, Speech and Signal Processing. South Brisbane, Queensland, Australia: IEEE, 2015: 4624-4628.
[74] Sainath T N, Weiss R J, Wilson K W, et al. Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms[C]// Automatic Speech Recognition and Understanding. Scottsdale, A Z, USA: IEEE, 2015: 30-36.
[75] Sainath T N, Weiss R J, Wilson K W, et al. Factored spatial and spectral multichannel raw waveform CLDNNs[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 5075-5079.
[76] Li B, Sainath T N, Weiss R J, et al. Neural network adaptive beamforming for robust multichannel speech recognition[C]// Interspeech. San Francisco, California: IEEE, 2016: 1976-1980.
[77] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[C]// International Conference on Learning Representations(ICLR). http://arxiv.org/abs/1409.0473, 2015.
[78] Chorowski J, Bahdanau D, Cho K, et al. End-to-end continuous speech recognition using attention-based recurrent NN: First results[J]. arXiv preprint arXiv:1412.1602, 2014.
[79] Chorowski J K, Bahdanau D, Serdyuk D, et al. Attention-based models for speech recognition[C]// Annual Conference on Neural Information Processing Systems. Montreal, Quebec, Canada: MIT Press, 2015: 577-585.
[80] Bahdanau D, Chorowski J, Serdyuk D, et al. End-to-end attention-based large vocabulary speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 4945-4949.
[81] Chan W, Jaitly N, Le Q, et al. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition[C]// International Conference on Acoustics, Speech and Signal Processing. Shanghai, China: IEEE, 2016: 4960-4964.

作者简介:

戴礼荣(1962-),男,教授,博士生导师,研究方向:语音识别、语音合成和说话人识别等,E-mail: lrdai@ustc.edu.cn。

张仕良(1990-),男,博士研究生,研究方向:基于深度学习的语音识别,自然语言理解,E-mail: zsl2008@mail.ustc.edu.cn。

黄智颖(1992-),男,硕士研究生,研究方向:语音声学建模说话人自适应,E-mail:zyhuang@mail.ustc.edu.cn。