视点聚焦

语音识别技术的研究进展与展望

王海坤，潘嘉，刘聪
（科大讯飞股份有限公司人工智能研究院，安徽 合肥 230088）

摘　要：自动语音识别（ASR）技术的目的是让机器能够“听懂”人类的语音，将人类语音信息转化为可读的文字信息，是实现人机交互的关键技术，也是长期以来的研究热点。最近几年，随着深度神经网络的应用，加上海量大数据的使用和云计算的普及，语音识别取得了突飞猛进的进展，在多个行业突破了实用化的门槛，越来越多的语音技术产品进入了人们的日常生活，包括苹果的 Siri、亚马逊的 Alexa、讯飞语音输入法、叮咚智能音箱等都是其中的典型代表。对语音识别技术的发展情况、最近几年的关键突破性技术进行了介绍，并对语音识别技术的发展趋势做了展望。

关键词：自动语音识别；深度神经网络；声学模型；语言模型
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000-0801.2018095

Research development and forecast of automatic speech recognition technologies

WANG Haikun, PAN Jia, LIU Cong
AI Research Institute of IFLYTEK Co., Ltd., Hefei 230088, China

Abstract: The purpose of automatic speech recognition (ASR) is to make the machine to be able to “understand” the human speech and transform it to readable text information. ASR is one of the key technologies of human machine interaction and also a hot research domain for a long time. In recent years, due to the application of deep neural networks, the use of big data and the popularity of cloud computing, ASR has made great progress and break through the threshold of application in many industries. More and more products with ASR have entered people’s daily life, such as Apple’s Siri, Amazon’s Alexa, IFLYTEK speech input method and Dingdong intelligent speaker and so on. The development status and key breakthrough technologies in recent years were introduced. Also, a forecast of ASR technologies’ trend of development was given.

Key words: automatic speech recognition, deep neural network, acoustic model, language model

1　引言

语音是人类最自然的交互方式。计算机发明之后，让机器能够“听懂”人类的语言，理解语言中的内在含义，并能做出正确的回答就成为了人们追求的目标。这个过程中主要涉及 3 种技术，

收稿日期：2018-01-15；修回日期：2018-01-31

即自动语音识别（automatic speech recognition，ASR）；自然语言处理（natural language processing，NLP），目的是让机器能理解人的意图；语音合成（speech synthesis，SS），目的是让机器能说话。

语音识别技术的目的是让机器能听懂人类的语音，是一个典型的交叉学科任务，涉及模式识别、信号处理、物理声学、生理学、心理学、计算机科学和语言学等多个学科。

语音识别技术的研究最早开始于20世纪50年代，1952 年贝尔实验室研发出了 10 个孤立数字的识别系统[1]。从 20 世纪 60 年代开始，美国卡耐基梅隆大学的 Reddy 等开展了连续语音识别的研究，但是这段时间发展很缓慢。1969 年贝尔实验室的 Pierce J 甚至在一封公开信中将语音识别比作近几年不可能实现的事情，例如“将水转化为汽油，从海里提取金子，治疗癌症”等。20 世纪 80 年代开始，以隐马尔可夫模型（hidden Markov model，HMM）方法[2,3]为代表的基于统计模型方法逐渐在语音识别研究中占据了主导地位。HMM 模型能够很好地描述语音信号的短时平稳特性，并且将声学、语言学、句法等知识集成到统一框架中。此后，HMM 的研究和应用逐渐成为了主流。例如，第一个“非特定人连续语音识别系统”是当时还在卡耐基梅隆大学读书的李开复研发的 SPHINX[4]系统，其核心框架就是 GMM-HMM 框架，其中 GMM（Gaussian mixture model，高斯混合模型）用来对语音的观察概率进行建模，HMM 则对语音的时序进行建模。20 世纪 80 年代后期，深度神经网络（deep neural network，DNN）的前身——人工神经网络（artificial neural network，ANN）也成为了语音识别研究的一个方向[5]。但这种浅层神经网络在语音识别任务上的效果一般，表现并不如 GMM-HMM 模型。20 世纪 90 年代开始，语音识别掀起了第一次研究和产业应用的小高潮，主要得益于基于 GMM-HMM 声学模型的区分性训练准则和模型自适应方法的提出。这时期剑桥发布的 HTK 开源工具包[6]大幅度降低了语音识别研究的门槛。此后将近 10 年的时间里，语音识别的研究进展一直比较有限，基于 GMM-HMM 框架的语音识别系统整体效果还远远达不到实用化水平，语音识别的研究和应用陷入了瓶颈。

2006 年 Hinton[7]提出使用受限波尔兹曼机（restricted Boltzmann machine，RBM）对神经网络的节点做初始化，即深度置信网络（deep belief network，DBN）。DBN 解决了深度神经网络训练过程中容易陷入局部最优的问题，自此深度学习的大潮正式拉开。2009 年，Hinton 和他的学生 Mohamed D[8]将 DBN 应用在语音识别声学建模中，并且在 TIMIT 这样的小词汇量连续语音识别数据库上获得成功。2011 年 DNN 在大词汇量连续语音识别上获得成功[9]，语音识别效果取得了近 10 年来最大的突破。从此，基于深度神经网络的建模方式正式取代 GMM-HMM，成为主流的语音识别建模方式。

2 语音识别声学模型中深度神经网络的应用

2.1 深度学习比浅层模型更适合语音处理

深度学习（deep learning，DL）是指利用多层的非线性信号和信息处理技术，通过有监督或者无监督的方法，进行信号转换、特征提取以及模式分类等任务的机器学习类方法[10]的总称。因为采用深层结构（deep architecture）[11] 模型对信号和信息进行处理，所以这里称为“深度”学习。传统的机器学习模型很多属于浅层结构（shallow structure）模型，例如支持向量机（support vector machine，SVM）、GMM、HMM、条件随机场（conditional random field，CRF）、线性或者非线性动态系统、单隐层的神经网络（neural network，NN）等。原始的输入信号只经过比较少的层次（通常是一层）的线性或者非线性处理以达到信号与信息处理，是这些结构模型的共同特点。浅层模

型的优点在于在数学上有比较完善的算法，并且结构简单、易于学习。但是浅层模型使用的线性或者非线性变换组合比较少，对于信号中复杂的结构信息并不能有效地学习，对于复杂信号的表达能力有局限性。而深层结构的模型则更适合于处理复杂类型的信号，原因在于深层结构具备多层非线性变换[12]，具有更强的表达与建模能力。

人类语音信号产生和感知就是这样一个极其复杂的过程，并且在生物学上被证明具有明显的多层次甚至深层次的处理结构[13]。所以，对于语音识别任务，采用浅层结构模型明显有很大的局限性。利用深层次结构中的多层非线性变换进行语音信号中的结构化信息和更高层信息的提取，是更加合理的选择。

2.2　DNN 在语音识别系统中的应用和局限性

从 2011 年之后，基于 DNN-HMM 声学模型[14-18]在多种语言、多种任务的语音识别上取得了比传统 GMM-HMM 声学模型大幅度且一致性的效果提升。基于 DNN-HMM 语音识别系统的基本框架如图 1 所示，采用 DNN 替换 GMM 模型来建模语音观察概率，是其和传统的 GMM-HMM 语音识别系统最大的不同。前馈型深度神经网络（feed-forward deep neural network，FDNN）由于比较简单，是最初主流的深层神经网络。

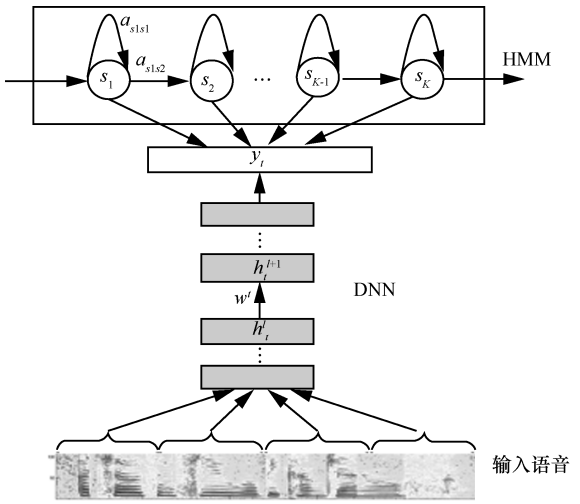


图 1　基于 DNN-HMM 的语音识别系统框架

使用 DNN 取代 GMM 主要有以下几个原因：DNN 可以将相邻的语音帧拼接起来作为输入特征，使得更长时的结构信息得以描述；DNN 的输入特征可以是多种特征的融合，也可以是离散或者连续的特征；不需要对语音数据分布进行假设，也是使用 DNN 估计 HMM 状态的后验概率分布的一个特点。

语音识别的特征提取需要首先对波形进行加窗和分帧，然后再提取特征。训练 GMM 模型的输入是单帧特征，DNN 则一般采用多个相邻帧拼接在一起作为输入，这种方法使得语音信号更长的结构信息得以描述，研究表明，特征拼接输入是 DNN 相比于 GMM 可以获得大幅度性能提升的关键因素。由于说话时的协同发音的影响，语音是一种各帧之间相关性很强的复杂时变信号，正要说的字的发音和前后好几个字都有影响，并且影响的长度随着说话内容的不同而时变。虽然采用拼接帧的方式可以学到一定程度的上下文信息，但是由于 DNN 输入的窗长（即拼接的帧数）是事先固定的，因此 DNN 的结构只能学习到固定的输入到输入的映射关系，导致其对时序信息的更长时相关性的建模灵活性不足。

2.3　递归神经网络在声学模型中的应用

语音信号具有明显的协同发音现象，因此必须考虑长时相关性。由于循环神经网络（recurrent neural network，RNN）具有更强的长时建模能力，使得 RNN 也逐渐替代 DNN 成为语音识别主流的建模方案。DNN 和 RNN 的网络结构如图 2 所示，RNN 在隐层上增加了一个反馈连接，是其和 DNN 最大的不同。这意味着 RNN 的隐层当前时刻的输入不但包括了来自上一层的输出，还包括前一时刻的隐层输出，这种循环反馈连接使得 RNN 原则上可以看到前面所有时刻的信息，这相当于 RNN 具备了历史记忆功能。对于语音这种时序信号来说，使用 RNN 建模显得更加适合。

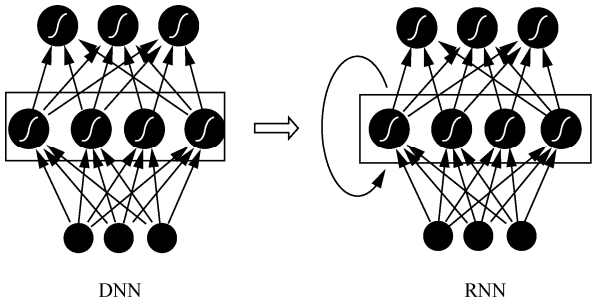


图 2　DNN 和 RNN 的结构区别示意

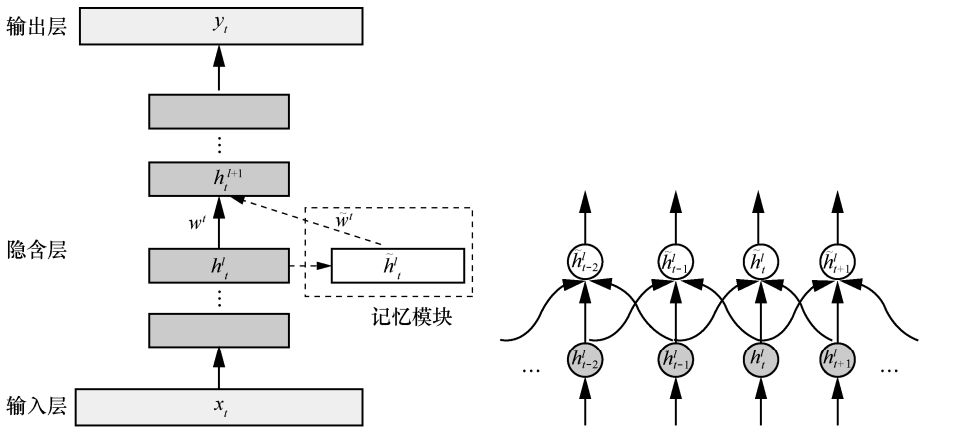
但是，传统的 RNN 在训练过程中存在梯度消失的问题，导致该模型难以训练。为了克服梯度消失问题，有研究人员提出了长短时记忆（long-short term memory，LSTM）RNN[19]。LSTM-RNN 使用输入门、输出门和遗忘门来控制信息流，使得梯度能在相对更长的时间跨度内稳定地传播。双向 LSTM-RNN （BLSTM-RNN）对当前帧进行处理时，可以利用历史的语音信息和未来的语音信息，从而容易进行更加准确的决策，因此也能取得比单向 LSTM 更好的性能提升。

尽管双向 LSTM-RNN 的性能更好，但它并不适合实时系统，由于要利用较长时刻的未来信息，会使得该系统具有很大时延，主要用于一些离线语音识别任务。基于此，研究人员提出了延迟受控 BLSTM（latency control-BLSTM）[20] 和行卷积 BLSTM 等模型结构，这些模型试图构建单向 LSTM 和 BLSTM 之间的折中：即前向 LSTM 保持不变，针对用来看未来信息的反向 LSTM 做了优化。在 LC-BLSTM 结构中，标准的反向 LSTM 被带有最多 N 帧前瞻量的反向 LSTM 替代，而在行卷积模型中被集成了 N 帧前瞻量的行卷积替代。

2.4　基于 FSMN 的语音识别系统

目前国际上已经有不少学术或工业机构在进行 RNN 架构下的研究。目前效果最好的基于 BLSTM-RNN 的语音识别系统存在时延过大的问题，这对于实时的语音交互系统（如语音输入法），并不合适。尽管可以通过 LC-BLSTM 和行卷积 BLSTM 将 BLSTM 做到实时语音交互系统，由于 RNN 具有比 DNN 更加复杂的结构，海量数据下的 RNN 模型训练需要耗费大量的时间。最后，由于 RNN 对上下文相关性的拟合较强，相对于 DNN 更容易陷入过拟合的问题，容易因为训练数据的局部问题而带来额外的异常识别错误。

为了解决以上问题，科大讯飞结合传统的 DNN 框架和 RNN 的特点，研发出了一种名为前馈型序列记忆网络（feed-forward sequential memory network，FSMN）的新框架[21]，具体如图 3 所示。FSMN 的结构采用非循环的前馈结构，只需要 180 ms 的时延，就达到了和 BLSTM-RNN 相当的效果。



（a）FSMN结构示意　　（b）第l个隐层记忆模块的时序展开示意（左右各看一帧）

图 3　FSMN 结构示意

FSMN 的结构示意如图 3(a)所示，其主要是基于传统 DNN 结构的改进，在 DNN 的隐层旁增加了一个“记忆模块”，这个记忆模块用来存储对判断当前语音帧有用的语音信号的历史信息和未来信息。图 3(b)画出了记忆模块左右各记忆 N 帧语音信息的时序展开结构。需记忆的历史和未来信息长度 N 可根据实际任务的需要来调整。FSMN 记忆块的记忆功能是使用前馈结构实现的，这点有别于传统的基于循环反馈的 RNN 模型。采用这种前馈结构存储信息有两大好处：首先，传统双向 RNN 必须等待语音输入结束才能对当前语音帧进行判断，双向 FSMN 对未来信息进行记忆时只需要等待有限长度的未来语音帧即可，这个优点使得 FSMN 的时延是可控的。实验证明，使用双向 FSMN 结构，时延控制在 180 ms 时就能取得和传统双向 RNN 相当的效果；其次，传统简单的 RNN 实际并不能记住无穷长的历史信息，而是只能记住有限长的历史信息，原因是其训练过程中存在梯度消失的问题。然而 FSMN 的记忆网络完全基于前馈展开，在模型训练过程中，梯度则沿着记忆块与隐层的连接权重（如图 3 所示）往回传给各个时刻，对判断当前语音帧的影响的信息通过这些连接权重来决定，而且这种梯度传播是可训练的，并且在任何时刻都是常数衰减，以上的实现方式使得 FSMN 也具有了类似 LSTM 的长时记忆能力，这相当于使用了一种更为简单的方式解决了传统 RNN 中的梯度消失问题。另外，由于 FSMN 完全基于前馈神经网络结构，也使得它的并行度更高，GPU 计算能力可利用得更加充分，从而获得效率更高的模型训练过程，并且 FSMN 结构在稳定性方面也表现得更加出色。

2.5　基于卷积神经网络的语音识别系统

卷积神经网络（convolutional neural network，CNN）的核心是卷积运算（或卷积层），是另一种可以有效利用长时上下文语境信息的模型[22]。继 DNN 在大词汇量连续语音识别上的成功应用之后，CNN 又在 DNN-HMM 混合模型架构下被重新引入。重新引入 CNN 最初只是为了解决频率轴的多变性[23-26]来提升模型的稳定性，因为该混合模型中的 HMM 已经有很强的处理语音识别中可变长度话语问题的能力。早期 CNN-HMM 模型仅使用了 1~2 个卷积层，然后和全连接 DNN 层堆叠在一起。后来，LSTM 等其他 RNN 层也被集成到了该模型中，从而形成了所谓的 CNN-LSTM-DNN（CLDNN）[27]架构。

基于 CNN-HMM 框架的语音识别吸引了大量的研究者，但是始终鲜有重大突破，最基本的原因有两个：首先是他们仍然采用固定长度的语音帧拼接作为输入的传统前馈神经网络的思路，导致模型不能看到足够的上下文信息；其次是他们采用的卷积层数很少，一般只有 1~2 层，把 CNN 视作一种特征提取器来使用，这样的卷积网络结构表达能力十分有限。针对这些问题，科大讯飞在 2016 年提出了一种全新的语音识别框架，称为全序列卷积神经网络（deep fully convolutional neural network，DFCNN）。实验证明，DFCNN 比 BLSTM 语音识别系统这个学术界和工业界最好的系统识别率提升了 15%以上。基于 DFCNN 语音识别框架示意如图 4 所示。

如图 4 所示，DFCNN 先对时域的语音信号进行傅里叶变换得到语音的语谱图，DFCNN 直接将一句语音转化成一张图像作为输入，输出单元则直接与最终的识别结果（比如音节或者汉字）相对应。DFCNN 的结构中把时间和频率作为图像的两个维度，通过较多的卷积层和池化（pooling）层的组合，实现对整句语音的建模。DFCNN 的原理是把语谱图看作带有特定模式的图像，而有经验的语音学专家能够从中看出里面说的内容。

为了理解 DFCNN 的优势所在，下面从输入端、模型结构和输出端 3 个角度更具体地分析。首先，在输入端，传统语音识别系统的提取特征

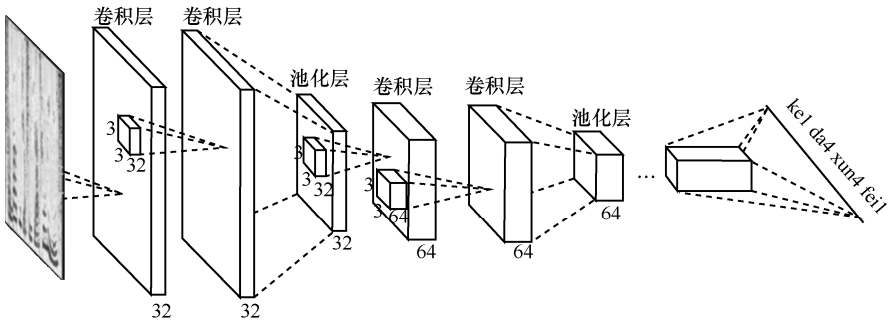


图 4　基于 DFCNN 语音识别框架示意

方式是在傅里叶变换后用各种类型的人工设计的滤波器，比如 Log Mel-Filter Bank，造成在语音信号频域，尤其是高频区域的信息损失比较明显。另外，传统语音特征采用非常大的帧移来降低运算量，导致时域上的信息会有损失，当说话人语速较快的时候，这个问题表现得更为突出。而 DFCNN 将语谱图作为输入，避免了频域和时域两个维度的信息损失，具有天然的优势。其次，从模型结构上来看，为了增强 CNN 的表达能力，DFCNN 借鉴了在图像识别中表现最好的网络配置，与此同时，为了保证 DFCNN 可以表达语音的长时相关性，通过卷积池化层的累积，DFCNN 能看到足够长的历史和未来信息，有了这两点，和 BLSTM 的网络结构相比，DFCNN 在顽健性上表现更加出色。最后，从输出端来看，DFCNN 比较灵活，可以方便地和其他建模方式融合，比如和连接时序分类模型（connectionist temporal classification，CTC）方案结合，以实现整个模型的端到端声学模型训练。DFCNN 语音识别框架可以方便地和其他多个技术点结合，实验证明，在数万小时的中文语音识别任务上，和目前业界最好的语音识别框架 BLSTM-CTC 系统相比，DFCNN 系统获得了额外 15%的性能提升。

2.6　大规模语音数据下神经网络声学模型的训练

相比于传统的 GMM-HMM 系统，基于 DNN-HMM 语音识别系统取得了巨大的性能提升[28,29]。但是 DNN 声学模型的训练却非常耗时。举个例子，在一个配置为 E5-2697 v4 的 CPU 上进行 2 万小时规模的语音数据的声学模型训练，大概需要 116 天左右才能训练完。造成这种情况的潜在原因是将随机梯度下降（stochastic gradient descent，SGD）算法作为神经网络训练中的基本算法，SGD 算法收敛相对较慢，而且是一个串行算法，很难进行并行化训练。而目前工业界主流的语音识别系统涉及的训练数据一般为几千小时甚至几万小时级别，因此，提高在大规模语音数据下深度神经网络的训练速度和训练效率，也成为了研究热点和必须解决的问题。

由于深度神经网络的模型参数非常稀疏，利用这个特点，参考文献[30]将深度神经网络模型中超过 80%的较小参数都设置为 0，几乎没有性能损失，同时模型尺寸大大减少，但是训练时间并没有明显减小，原因是参数稀疏性带来的高度随机内存访问并没有得到太多的优化。进一步地，参考文献[31]提出在深度神经网络中，用两个低秩矩阵的乘积表示权重矩阵，实现了 30%~50%的效率提升。

通过使用多个 CPU 或者 GPU 并行训练来解决神经网络训练效率是另外一种可行的方法。参考文献[32,33]的方式是：把训练数据分成许多小块后并行地送到不同的机器来进行矩阵运算，从而实现并行训练。参考文献[34]的优化方案是：在模型的每遍迭代中，先将训练数据分成 N 个完全不相交的子集，然后在每个子集中训练一个

sub-MLP，最后把这些 sub-MLP 进行合并网络结合。为了进一步提升并行效率，参考文献[35]在上千个 CPU 核的计算集群实现了这种方式，深层网络的训练主要是利用异步梯度下降（asynchronous SGD）算法。参考文献[36]将异步梯度下降算法应用到了多个 GPU 中。在参考文献[37]中，一种管道式的 BP 算法被提了出来，该方法利用不同的 GPU 单元来计算神经网络中不同层，实现并行训练的效果。实验证明，相对使用单个 GPU 训练，该方法通过使用 4 个 GPU 实现了 3.1 倍左右的效率提升。然而，不同计算单元之间极其频繁的数据传递成为该类方法提升训练效率的主要瓶颈。为此，为了更好地实现神经网络并行训练，一种新的基于状态聚类的多深层神经网络建模方法[38]被提出，该方法先将训练数据在状态层面进行聚类，在状态层面进行不相交的子集划分，使得不同计算单元神经网络之间的数据传递规模大幅度减小，从而实现每个神经网络完全独立的并行训练。使用 4 块 GPU，在聚类数为 4 类的情况下，在 SWB（SwitchBoard）数据集上的实验表明，这种状态聚类的多神经网络方法取得了约 4 倍的训练效率提升。

3　语音识别语言模型中深度神经网络的应用

深度学习理论除了在声学模型建模上获得了广泛的应用外，在语音识别系统另外的重要组件——语言模型上也得到了应用。在深度神经网络普及之前，语音识别系统主要采用传统的统计语言模型 *N*-gram 模型[39]进行建模。*N*-gram 模型也具备明显的优点，其结构简单且训练效率很高，但是 *N*-gram 的模型参数会随着阶数和词表的增大而指数级增长，导致无法使用更高的阶数，性能容易碰到瓶颈，在训练语料处于相对稀疏的状态时，可以借助降权（discounting）和回溯（backing-off）等成熟的平滑算法解决低频词或不可见词的概率估计问题，以获得比较可靠的模型估计。

在 20 世纪初，一些浅层前馈神经网络被用于统计语言模型建模[40]。神经网络语言模型是一种连续空间语言模型，平滑的词概率分布函数使得它对于训练语料中的低频词和不可见词的概率估计更为顽健，具有更好的推广性，在语音识别任务上也取得了显著的效果[41]。最近几年，相关研究人员也将深层神经网络用于语言模型建模，并取得了进一步的性能提升[42]。

然而，前馈神经网络语言模型只能够处理固定长度的历史信息，其仍然存在 *N* 阶假设，即在预测当前词概率的时候只与之前 *N*-1 个词有关，这在一定程度上影响了模型的准确性。实际上，人类能够记忆和处理的历史信息要长久得多，而标准的 RNN 正好能够通过循环网络结构记忆和处理任意长度的历史信息，因此参考文献[43]将 RNN 引入语言模型建模中。RNN 相比于前馈神经网络取得了更好的性能。然而，由于基于 RNN 的深层网络的复杂特性，模型的训练训练依旧非常耗时，在大文本（100 GB ~ 1 TB）语料上几乎不可实现。参考文献[44]提出在 GPU 上将多个句子拼接为数据组（mini-batch）同时参与训练，大幅度地提升了 RNN 的训练效率。科大讯飞基于参考文献[45]的方法进一步改进，将 RNN 的输出层基于词聚类进行了分解，在中文 LVCSR 任务上获得了 50 倍以上的训练效率提升。在提高训练效率的基础上，RNN 模型相对于传统 *N*-gram 模型也获得了 5%以上的识别效果提升，这也进一步验证了 RNN 的有效性。参考文献[46]提出了基于 LSTM（long short-term memory）的 RNN 语言模型结构，通过对网络结构的调整，有效解决了 RNN 语言模型训练中存在梯度消失（gradient vanishing）的问题[47]，并获得了一定的性能提升。

4　深度学习、大数据和云计算之间的关系

基于深度学习的语音识别技术在21世纪初走向舞台的中央，并不只是由于深度学习类机器学

习算法的进步，而是大数据、云计算和深度学习这 3 个要素相互促进的结果。

不同于之前 GMM-HMM 语音识别框架表达能力有限、效果对于大规模数据易饱和的情况，深度学习框架所具备的多层非线性变换的深层结构，则具有更强的表达与建模能力，使得语音识别模型对复杂数据的挖掘和学习能力得到了空前的提升，使得更大规模的海量数据的作用得以充分的发挥。大数据就像奶粉一样，“哺育”了深度学习算法，让深度学习算法变得越来越强大。

随着移动互联网、物联网技术和产品的普及，更重要的是采用云计算的方式，使得多种类型的海量数据得以在云端汇集。而对大规模的数据的运算的要求则又显著提升了对于云计算方式的依赖，因此云计算成为了本次深度学习革命的关键推手之一。

深度学习框架在云端的部署，则显著增强了云计算的能力。

正是由于深度学习、大数据和云计算三者的相互促进，才成就了本次语音技术的进步，成就了本次人工智能的浪潮。

5 总结和展望

本文对语音识别领域的研究状况和最近几年的关键突破性技术做了比较详细的介绍。首先简要回顾了语音识别技术发展的历史，然后重点介绍了深度神经网络在语音识别声学模型建模中起到的引领作用，也介绍了各种形态（包括 LSTM、FSMN、DFCNN 等）的关键技术突破。相关研究证明，和传统的 GMM-HMM 框架相比，深度学习在大词汇量连续语音识别任务方面取得了 30%~60%的性能提升。也介绍了深度声学模型训练的优化方法以及 RNN 在语言模型建模中的应用，在语言模型领域同样能取得比传统 *N*-gram 语言模型 5%以上的识别效果提升。毫不夸张地说，深度学习技术的确给语音识别的研究和应用带来了革命性的历史突破。

语音识别技术进一步的研究热点方向应该包含以下几个。

首先是端到端的语音识别系统。在目前 DNN-HMM 的混合框架下，声学模型中 DNN、HMM 两个部分以及语言模型都是单独训练的。然而语音识别是一个序列识别的任务，如果能够对声学模型的各个部分以及语言模型进行联合优化，并且去除类似于发音词典等所有需要人工来设计的组件，必定能取得更进一步的效果提升。目前在声学模型建模领域已经出现了端到端的模型应用，即将声学模型中的各个组件做联合优化，且优化目标是输出的词或音素序列，而不是使用交叉熵（cross entropy，CE）准则来优化一帧一帧的标注，比如连接时序分类准则（CTC）被引入[48-50]，并且在多个任务上取得了一定的效果。另外，受到 CTC 的启发，一种被称为无词图最大互信息（lattice free maximum mutual information，LFMMI）的准则被提出，可以实现从头训练的深度神经网络，不需要使用交叉熵做网络的初始化。但是无论是 CTC 还是 LFMMI，都不能称为真正的端到端语音识别模型，它们仍需要发音词典、语言模型等组件，需要大量的专家知识来辅助设计。受到在翻译领域成功应用的 Attention 模型的启发[51,52]，Encoder-Decoder 框架已经不明确区分声学模型和语言模型，并且完全不需要发音词典等人工知识，可以真正地实现端到端的建模。Encoder-Decoder 框架的模型训练难度很大并且收敛比较缓慢，目前 Google（谷歌）和科大讯飞在新一代端到端框架下已经取得了正面的效果提升，后面应该会吸引更多的研究机构和学者进入该领域进行研究。

其次，直接利用时域波形语音建模来代替人工设计的特征（比如 Log Mel-Filter Bank 等）。主要原因是原始的时域波形文件中的信息量是最丰富的，在通过人工设计提取一些特征的同时也

会抛弃一些信息，这些信息对于噪声较大等复杂场景识别十分重要。研究人员也在这个领域进行了相关的工作[53,54]，但是只取得了与人工设计特征相当的效果。科大讯飞最新的研究成果证明，直接利用时域波形来建模在多个任务上都取得了10%以上的识别效果提升，并且认为该方面仍然会有巨大的提升潜力。

最后，利用多个麦克风信号和深度学习来联合建模，用来提升远场环境下的语音识别效果的研究也是近期和长期的热点，但是如何将深度学习对于离线大数据的学习能力和传统的信号处理对于瞬时信号处理能力结合起来，仍需要很多的研究工作要做。

参考文献：

[1] DAVIS K. H, BIDDULPH R, BALASHEK S. Automatic recognition of spoken digits[J]. Journal of the Acoustical Society of America, 1952, 24(6): 637.

[2] FERGUSON J D. Application of hidden Markov models to text and speech[EB]. 1980.

[3] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Readings in Speech Recognition, 1990, 77(2): 267-296.

[4] LEEE K F L M. An overview of the SPHINX speech recognition system[J]. IEEE Transactions on Acoustics Speech & Signal Processing Speech, 1990, 38(1): 35-45.

[5] WAIBEL A, HANAZAWA T, HINTON G. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1990, 1(2): 393-404.

[6] YOUNG S, EVERMANN G, GALES M, et al. The HTK book[EB]. 2005.

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

[8] MOHAMED A R, DAHL G, HINTON G. Deep belief networks for phone recognition[EB]. 2009.

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

[10] DENG L. An overview of deep-structured learning for information processing[C]//Asian-Pacific Signal and Information Processing-Annual Summit and Conference (APSIPA-ASC), October 18, 2011, Xi'an, China. [S.l.:s.n.], 2011.

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

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

[13] BAKER J, DENG L, GLASS J, et al. Developments and directions in speech recognition and understanding[J]. IEEE Signal Processing Magazine, 2009, 26(3): 75-80.

[14] MOHAMED A R, DAHL G, HINTON G. Deep belief networks for phone recognition[EB]. 2009.

[15] SAINATH T N, KINGSBURY B, RAMABHADRAN B, et al. Making deep belief networks effective for large vocabulary continuous speech recognition[EB]. 2011.

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

[17] 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.

[18] 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]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

[19] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[20] ZHANG Y, CHEN G G, YU D, et al. Highway long short-term memory RNNS for distant speech recognition[C]//2016 IEEE International Conference on Acoustics, Speech and Signal Processing, March 20-25, 2016, Shanghai, China. Piscataway: IEEE Press, 2016.

[21] ZHANG S L, LIU C, JIANG H, et al. Feedforward sequential memory networks: a new structure to learn long-term dependency[J]. arXiv:1512.08301, 2015.

[22] LECUN Y, BENGIO Y. Convolutional networks for images, speech and time-series[M]. Cambridge: MIT Press, 1995.

[23] ABDEL-HAMID O, MOHAMED A R, JIANG H, et al. Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition[C]//2012 IEEE International Conference on Acoustics, Speech and Signal Processing, March 20, 2012, Kyoto, Japan. Piscataway: IEEE Press, 2012: 4277-4280.

[24] ABDEL-HAMID O, MOHAMED A R, JIANG H, et al. Convolutional neural networks for speech recognition[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2014, 22(10): 1533-1545.

[25] ABDEL-HAMID O, DENG L, YU D. Exploring convolutional neural network structures and optimization techniques for speech recognition[EB]. 2013.

[26] SAINATH T N, MOHAMED A R, KINGSBURY B, et al. Deep convolutional neural networks for LVCSR[C]//2013 IEEE In-

ternational Conference on Acoustics, Speech and Signal Processing, May 26-30, 2013, Vancouver, BC, Canada. Piscataway: IEEE Press, 2013: 8614-8618.

[27] SAINATH T N, VINYALS O, SENIOR A, et al. Convolutional, long short-term memory, fully connected deep neural networks[C]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing, April 19-24, Brisbane, QLD, Australia. Piscataway: IEEE Press, 2015: 4580-4584.

[28] SEIDE F, LI G, YU D. Conversational speech transcription using context- dependent deep neural networks[C]// International Conference on Machine Learning, June 28-July 2, 2011, Bellevue, Washington, USA. [S.l.:s.n.], 2011: 437-440.

[29] DAHL G E, YU D, DENG L, et al. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs[C]//ICASSP, May 22-27, 2011, Prague, Czech Republic. [S.l.:s.n.], 2011: 4688-4691.

[30] YU D, SEIDE F, LI G, et al. Exploiting sparseness in deep neural networks for large vocabulary speech recognition[C]//ICASSP, March 25-30, 2012, Kyoto, Japan. [S.l.:s.n.], 2012: 4409-4412.

[31] 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, May 26-31, 2013, Vancouver, BC, Canada. [S.l.:s.n.], 2013: 6655-6659.

[32] KONTÁR S. Parallel training of neural networks for speech recognition[C]//13th International Conference on Text, Speech and Dialogue, September 6-10, 2010, Brno, Czech Republic. New York: ACM Press, 2006: 6-10.

[33] VESELÝ K, BURGET L, GRÉZL F. Parallel training of neural networks for speech recognition[C]//13th International Conference on Text, Speech and Dialogue, September 6-10, 2010, Brno, Czech Republic. New York: ACM Press, 2006: 439-446.

[34] PARK J, DIEHL F, GALES M J F, et al. Efficient generation and use of MLP features for Arabic speech recognition[C]//Interspeech, Conference of the International Speech Communication Association, September 6-10, 2009, Brighton, UK. [S.l.:s.n.], 2009: 236-239.

[35] LE Q V, RANZATO M A, MONGA R, et al. Building high-level features using large scale unsupervised learning[J]. arXiv preprint arXiv:1112.6209, 2011.

[36] ZHANG S, ZHANG C, YOU Z, et al. Asynchronous stochastic gradient descent for DNN training[C]//IEEE International Conference on Acoustics, June 27-July 2, 2013, Santa Clara Marriott, CA, USA. Piscataway: IEEE Press, 2013: 6660-6663.

[37] CHEN X, EVERSOLE A, LI G, et al. Pipelined back-propagation for context-dependent deep neural networks[C]//13th Annual Conference of the International Speech Communication Association, September 9-13, 2012, Portland, OR, USA. [S.l:s.n.], 2012: 429-433.

[38] ZHOU P, LIU C, LIU Q, et al. A cluster-based multiple deep neural networks method for large vocabulary continuous speech recognition[C]//ICASSP, May 26-31, 2013, Vancouver, BC, Canada. [S.l.:s.n.], 2013: 6650-6654.

[39] JELINEK F. The development of an experimental discrete dictation recognizer[J]. Readings in Speech Recognition, 1990, 73(11): 1616-1624.

[40] BENGIO Y, DUCHARME R, VINCENT P. A neural probabilistic language model[J]. Journal of Machine Learning Research, 2003(3): 1137-1155.

[41] SCHWENK H, GAUVAIN J L. Training neural network language models on very large corpora[C]//Conference on Human Language Technology & Empirical Methods in Natural Language Processing, October 6-8, 2005, Vancouver, BC, Canada. New York: ACM Press, 2005: 201-208.

[42] ARıSOY E, SAINATH T N, KINGSBURY B, et al. Deep neural network language models[C]//NAACL-HLT 2012 Workshop, June 8, 2012, Montreal, Canada. New York: ACM Press, 2012: 20-28.

[43] MIKOLOV T, KARAFIAT M, BURGET L, et al. Recurrent neural network based language model[C]//11th Annual Conference of the International Speech Communication Association, September 26-30, 2010, Makuhari, Chiba, Japan. [S.l.:s.n.], 2010: 1045-1048.

[44] CHEN X, WANG Y, LIU X, et al. Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch[EB]. 2014.

[45] MIKOLOV T, KOMBRINK S, BURGET L, et al. Extensions of recurrent neural network language model[C]//IEEE International Conference on Acoustics, May 22-27, 2011, Prague, Czech Republic. Piscataway: IEEE Press, 2011: 5528-5531.

[46] SUNDERMEYER M, SCHLUTER R, NEY H. LSTM neural networks for language modeling[EB]. 2012.

[47] BENGIO Y, SIMARD P, FRASCONI P. Learning long term dependencies with gradient descent is difficult[J]. IEEE Transactions on Neural Networks, 1994, 5(2): 157.

[48] SAK H, SENIOR A, RAO K. Learning acoustic frame labeling for speech recognition with recurrent neural networks[C]//2015 ICASSP, April 19-24, 2015, Brisbane, QLD, Australia. [S.l.:s.n.], 2015: 4280-4284.

[49] SAK H, SENIOR A, RAO K, et al. Fast and accurate recurrent neural network acoustic models for speech recognition[J]. arXiv:1507.06947, 2015.

[50] SENIOR A, SAK H, QUITRY F D C, et al. Acoustic modelling with CD-CTC-SMBR LSTM RNNS[C]//2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 13-17, 2015, Scottsdale, AZ, USA. Piscataway: IEEE Press, 2015: 604-609.

[51] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.

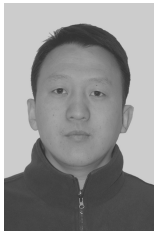
[52] MNIH V, HEESS N, GRAVES A, et al. Recurrent models of visual attention[C]//28th Annual Conference on Neural Infor-

mation Processing Systems, December 8-13, 2014, Montreal, Canada. [S.l.:s.n.], 2014: 2204-2212.

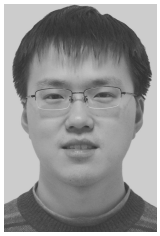
[53] TUSKE Z, GOLIK P, SCHLUTER R, et al. Acoustic modeling with deep neural networks using raw time signal for LVCSR[EB]. 2014.

[54] SAINATH T N, WEISS R J, SENIOR A W, et al. Learning the speech front-end with raw waveform[EB]. 2015.

[作者简介]



王海坤（1984−），男，科大讯飞股份有限公司人工智能研究院副院长，牵头研发科大讯飞嵌入式识别系统和远场识别系统，叮咚音箱技术总负责人，主要研究方向为语音识别、麦克风阵列语音信号处理、回声消除、语音交互等。著有 40 多篇发明专利，多项研究成果获得省级以上表彰。



潘嘉（1985−），男，科大讯飞股份有限公司人工智能研究院语音识别组研究主管，科大讯飞学术委员会委员，主要研究方向为语音识别。在深度神经网络领域有极深的造诣，是科大讯飞语音识别系统研发的主要参与者。

刘聪（1984−），男，博士后，科大讯飞股份有限公司人工智能研究院副院长，长期从事语音识别和人工智能等相关领域的研究工作。从 2014 年底开始，全面负责科大讯飞人脸识别、医学图像识别、视频监控等方向的研究工作，研究成果在多个内部产品中成功应用。2014 年获得北京市科学技术奖一等奖，发表论文 10 余篇，获得专利 10 余项。