文章编号：2095-2163(2019)02-0135-08 中图分类号：TN912.34 文献标志码：A

基于深度学习的语音识别方法研究

邵 娜[1,2]，李晓坤[1,2]，刘 磊[1,2]，陈虹旭[1,2]，郑永亮[1,2]，杨 磊[1,2]
(1 黑龙江恒讯科技有限公司国家博士后科研工作站，哈尔滨 150090；
2 黑龙江省智慧媒体工程技术研究中心(黑龙江恒讯科技有限公司)，哈尔滨 150090)

摘 要：本文主要介绍了深度学习和语音识别技术的发展历史和发展现况，研究意义及目的。通过深度学习技术建立声学模型，从而引入语音识别技术中分析其发展前景。本文希望通过对基于深度学习的语音识别方法研究，将语音识别的效率优化，准确率提高，从而促进语音识别技术的发展。

关键词：语音识别；深度学习；深度神经网络

Research on speech recognition based on depth learning

SHAO Na[1,2], LI Xiaokun[1,2], LIU Lei[1,2], CHEN Hongxu[1,2], ZHENG Yongliang[1,2], YANG Lei[1,2]
(1 Heilongjiang Hengxun Technology Co.,Ltd. Postdoctoral Programme, Harbin 150090, China; 2 Engineering Research Center Of Smart Media,Heilongjiang Province(Heilongjiang Hengxun Technology Co.,Ltd.), Harbin 150090, China)

【Abstract】 The paper mainly introduces the history and development of deep learning and speech recognition technology, and elaborates significance and purpose of the study. The acoustic model is established by deep learning technology to introduce the development prospect of the speech recognition technology. In this paper, based on deep learning, the research could improve the efficiency and the accuracy of speech recognition, so as to promote the development of speech recognition technology.

【Key words】 speech recognition; depth learning; Deep Neural Network

0 引 言

从原始社会开始，语言就是人类之间沟通的桥梁，这是最直接、也是最清晰的表达方式。作为人类交流思想的媒介，语言对文明的进步起着不可磨灭的作用。通过研究可知，语言对人类来说是一项标志性的因素。语言在沟通过程中的一类重要属性就是发出指令，通过指令可以调派某人完成某项任务。进而人们开始思考能否通过语言对人工智能发出命令。从此，语音识别技术开始出现在学界的视野中。语音识别能够将人类的语言与人工智能进行融合，从而实现对计算机下达命令的目的。语音识别的目的是通过计算机接收人类语言，并将人类语言解读为指令，从而实现人类与计算机的交互智能化。近些年来语音识别技术得到了飞跃性的发展，语音识别的研究也日渐受到学界的推崇与重视。语音识别技术已经不再局限于仅是科研人员实验室中的产物，而是融入人类生活中，成为了一种商品。现如今，在互联网、以及市面上均陆续涌现出大量与语音识别相关的软件。凭借着语音识别技术的实用性与准确性，在通讯设备、汽车、智能家居等载体上，语音识别技术的实用性则已在广泛的应用中得到了完美的阐释。相关研究人员将提高语音识别的准确性作为目标，各类研发成果相继问世，这些研究均旨在创造一个准确率方面的新高。本文研究致力于将深度学习与语音识别相互融合，从而优化语音识别的效率。

随着互联网技术的不断发展，人们越来越重视人与计算机之间的交互命令，因此用语音来实现这一目标，主要包括 3 项技术，分别是：语音识别、语音编码和语音合成[1]。突破技术层面的难题，自动识别人类发出的自然信号，对其进行解码转换文本。近些年来，人类从未在语音识别的道路上停止过探索与前行。1952 年，贝尔实验室的 3 名研究人员建立了一个单喇叭数字识别系统。该系统就是通过定位每个声音功率谱中的共振峰来开展工作[2]。20 世纪 60 年代末，苏联研究人员发明了动态时间扭曲算法[3]，虽然已被后来更高效的算法所取代，但是

基金项目：中小企业创新基金(2017FF1GJ023)；专利优势示范企业基金(2017YBQCZ029)；国家自然科学基金(81273649)。

作者简介：邵 娜(1987-)，女，硕士，工程师，主要研究方向：虚拟化、云计算、人工智能等；李晓坤(1979-)，男，研究员级高级工程师，教授，硕士生导师，系统集成高级项目经理，CCF 高级会员，主要研究方向：虚拟化、人工智能、生物特征识别等。

通讯作者：李晓坤 Email：li.xiaokun@163.com

收稿日期：2018-12-18

将信号分割成帧的技术将会继续得以不断的创新及演变。20 世纪 60 年代末,Leonard Baum 在国防分析研究所开发了马尔可夫链的数学模型。Raj Reddy 的学生 James Baker 和 Janetm Bakerk 开始考虑将隐马尔可夫模型(HMM)与语音识别结合,从而研究出一种新型混合模型[4]。最早的语音识别产品是来自 Kurzweil 应用的智能识别器,于 1987 年发布[5-6]。在 21 世纪初,语音识别仍然使用传统的方法,如隐藏的马尔可夫模型和前馈人工神经网络[7]。而回顾整个的语音识别历史发现,人们已经持续多年地始终都在探究研发浅层表现形式和深层的人工神经网络。但这些方法在与高斯混合模型/隐马尔可夫模型(GMM-HMM)技术的较量中从未占据过上风[8]。直至 2009 年,学界才开启深度学习的研究序幕,并逐渐掀起研究热潮。

1 语音识别和深度学习的概述

1.1 语音识别的科学内涵

目前,语音识别已成为学界研究热点,其研究目的就是为了让计算机能够听懂人类发出的指令。选择隐马尔可夫模型(Hidden Markov model,HMM)来建立语音识别系统堪称是当下的首要选择。说话人所发出的语音信号具备短时平稳性。HMM 的状态不能够被研究者直接观察到,故而 HMM 模型是属于马尔科夫链的一类。通过观察某些密度分布产生的概率,从而计算求得其相应的观测向量。在 20 世纪 80 年代,有相关的研究人员尝试将 HMM 与语音识别相结合,得到的结果比较符合预期。HMM 在图像识别、语音识别等领域正迅速成为设计者瞩目的焦点,越来越多的研究人员开始跻身于此项研究的行列当中。

1.2 隐马尔可夫模型概述

传统的 HMM 是一种统计学习的模型,这个过程通常是不能被观测的。这个过程可以看作为一种简单的动态贝叶斯网络。HMM 模型通常会被划归于这类网络。Baum 及其同事开发了基于 HMM 的数学模型。通多观察简单的马尔可夫模型,研究人员可以准确测定该模型的状态,故而状态转移概率作为模型仅有的参数,而在隐马尔可夫模型中,状态却非直接可见的,但是输出则依赖于状态,是可见的。每个状态在可能的输出令牌上有一个概率分布。如果想要获取相关状态序列的数据,需要由 HMM 生成令牌指令序列。形容词的隐藏并不是指描述模型的参数,而是模型之间互相传递的状态序列。即使这些参数是精确、且已知的,该模型仍将被称为隐马尔可夫模型。HMM 模型尤其适用于强化学习和模式识别。可以把隐马尔可夫模型视作原有模型的变化形式,从中选取一个隐藏的变量支配混合模型确定一个观察者。最近,隐马尔可夫模型已经推广到 Pairwise Markov 模型和 Triplet Markov 模型中,而这些模型能够支持更复杂的数据结构和非平稳数据的建模。

隐马尔可夫模型(HMM)由 5 种元素组成,其中含有 2 个状态集合以及 3 个概率矩阵,对此可写作如下数学形式:

$$H = \{S, O, \boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{B}\}. \quad (1)$$

其中,S 表示隐含状态。这部分状态是被隐马尔可夫模型不可被观察的一种状态,无法以观察者的身份进行观测;O 表示可观测状态,这部分状态通常能被观察者直接观测,并与 S 有一定联系;$\boldsymbol{\pi}$ 表示模型初始时期的概率分布,从而堆积而成的矩阵,设 $T = 1$,这个时期会得到一定概率,将其组成矩阵;$\boldsymbol{A}$ 表示模型在隐藏情况下产生的概率组成的矩阵,展示出隐马尔可夫模型互相之间信号传输的概率。并且,$\boldsymbol{A}_{ij} = P(S_j \mid S_i)$, $1 \leqslant i, j \leqslant N$;$\boldsymbol{B}$ 表示观测状态下的转移概率矩阵,假设 X 为隐含状态数,Y 为可被观测状态数,N 为被模型隐藏状态的数量,M 为模型中能被观测到的状态的数量,则:$\boldsymbol{A}_{ij} = P(O_j \mid O_i)$,$1 \leqslant i \leqslant X$, $j \geqslant 1$。其数学含义是:在 t 时刻模型的基本状态隐藏是 S_i 的前提下,模型状态的观察为 O_i 的概率。

综上论述可知,通常情况下,通过 $\lambda = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$ 这个三元组可以相对简化地描述出一个隐马尔可夫模型。隐马尔可夫模型是由马尔可夫模型经过演绎完善得到的一种新型模型,马尔科夫模型的状态集合不能以观察者的身份追踪监测与被隐藏的状态之间的概率联系,而隐马尔可夫模型却能做到这一点。当将 HMM 用作声学模型时,其设计结构如图 1 所示,表现为某状态转变为另一状态的转移概率。

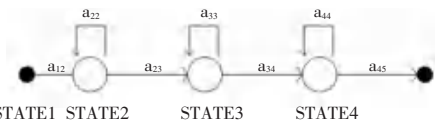

图 1 状态转移概率结构图

Fig. 1 State transition probability structure diagram

高斯混合模型(GMM)是一种参数概率密度函数,表示高斯分量密度的加权和。高斯混合模型是由方程给出的 M 分量高斯密度的加权和,其数学公式可表示为:

$$P(x \mid \lambda) = \sum_{i=1}^{M} \omega_i g(x \mid \mu_i, \sum_i), \quad (2)$$

其中，x 是 d 维连续值数据向量(即测量或特征)，$\omega_i (i = 1, \cdots, M)$ 是混合权重，$g\left(x \mid \mu_i, \sum i\right) (i = 1, \cdots, M)$ 是组件高斯密度，每个分量密度是形式的 d - 变量，语音输入信号的分布情况一般不能够用单高斯概率密度函数做出描述，大多数情况下是采用混合高斯函数表示输出概率。即：

$$g(x \mid \mu_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{D/2} \mid \boldsymbol{\Sigma}_i \mid^{1/2}} \exp\{ - \frac{1}{2} (x - \mu_i)' \boldsymbol{\Sigma}_i^{-1} (x - \mu_i) \}. \quad (3)$$

其中，μ_i 为均值向量，$\boldsymbol{\Sigma}_i$ 为协方差矩阵，混合权值满足约束条件 $\sum_{i=1}^{M} \omega_i = 1$。

1.3　深度学习的基本概念

在 20 世纪初，Hinton 等人发表了深度学习的构想，并提出了非监督逐层训练算法，这也是深层结构学习研究上的一个重要突破。同时又提出一种用于深层结构编码的编码器，能够利用空间关系，减少参数数目，从而优化训练性能[9]。深度学习是基于数据的一种更泛化的机器学习模式，而不是特定的某种算法。深度学习的架构就是深度神经网络，已经被大范围地应用到图像识别、语音识别、社交过滤、机器翻译、药物设计等领域中。在相当一部分领域，深度学习的能力要强于在该领域的专家。深度学习可以看作一种深度挖掘数据的新兴机器学习模式。通过采用级联多个非线性处理单元的方式进行特征的提取以及转换。每个连续层使用前一层的输出层作为输入层。在有监督的情况下学习(例如分类)或无监督(例如模式分析)。深度学习是一种对应多层的空间，同时与多个不同的隐藏层次进行映射。各个层次有机结合，从而组成一种概念层次结构。启用了一项坡度下降的模式，从而执行反向传播训练。深层学习中使用的层包括人工神经网络的隐含层和一组命题公式。此外，也可能包括深层生成模型中有组织变量的潜变量，如深层信念网络中的节点和深度玻尔兹曼机。

2　开发平台及数据来源

2.1　Kaldi 语音识别系统

Kaldi 语音识别系统是 Daninel Povey 等人使用 C++开发的一种语音识别系统。可在 GNU、Linux、BSD、OSX 10(8/9 等)、Windows (via Cygwin)等环境下运行。Kaldi 语音识别系统的开发能够提供一种兼具灵活性和可扩展性的语音识别系统开发平台。该系统支持线性变换，增加了 MMI 和 MCE 等基于特征空间的区分性训练的深层神经网络。Kaldi 语音识别系统在将深度学习与语音识别相结合的过程中表现出方便、且内容丰富的特点。该软件的成功开发是 2009 年约翰霍普金斯大学研讨会上的重要组成部分[10]。

2.2　深度学习的方法

2.2.1　循环神经网络

循环神经网络(Recurrent Neural Network，RNN)是一类人工神经网络在单位之间的连接形成一种循环，这使其可以表现出一种动态的时间行为。这个网络内部存在专属的存储器，可以用其输入需要的序列。这就使得该网络适用于多种任务，诸如分类、文字识别[11]或语音识别[12-13]等。卷积网络的灵感来自于生物过程，其中神经元之间的连接模式即是受到了动物视觉皮层组织的启发。

只有受约束的可观测区域会刺激独立的神经单位，这被称作感受野。多个独立的神经单位的感受野的局部会叠加，从而包含整片视野。与其它图像分类算法相比，CNN 进行预处理相对较小。在特征设计中，这种与先前知识储备无关的独立性是一个主要优点。循环神经网络有 2 种。一种是单向 RNN，另一种是双向 RNN。本次研究中可能用到的原理公式可分述如下。

(1)单向 RNN 的前层。具体公式如下：

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_i^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}, \quad (4)$$

$$b_h^t = \theta_h(a_h^t), \quad (5)$$

需要提及的是，在所有时间点中，隐层权重是共享的，所以需要将所有时间序列累加成和。研究中推得的隐层权重的偏导的计算公式为：

$$\frac{\partial Loss}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial Loss}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t, \quad (6)$$

$$\frac{\partial Loss}{\partial w_{hh'}} = \sum_{t=1}^{T} \frac{\partial Loss}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{hh'}} = \sum_{t=1}^{T} \delta_j^t b_{h'}^{t-1}, \quad (7)$$

(2)双向 RNN。具体公式如下。

① 激活前端网络隐藏层，其公式为：

$$a_h^t = \sum_{i=1}^{I} w_{ih} x_j^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t-1}, \quad (8)$$

$$b_h^t = \theta_h(a_h^t), \quad (9)$$

② 激活后端网络隐藏层，其公式为：

$$c_h^t = \sum_{i=1}^{I} w_{ih} x_j^t + \sum_{h'=1}^{H} w_{h'h} b_{h'}^{t+1}, \quad (10)$$

$$d_h^t = \theta_h(c_h^t), \quad (11)$$

③ 网络计算的输出,其公式为:

$$e_h^t = b_h^t + d_h^t, \quad (12)$$

需要提及的是,在所有时间点中,隐层权重是共享的,所以需要将所有时间序列累加成和。研究中推得的隐层权重的偏导的计算公式为:

$$\frac{\partial Loss}{\partial w_{ij}} = \sum_{t=1}^{T} \frac{\partial Loss}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{ij}} = \sum_{t=1}^{T} \delta_j^t b_i^t, \quad (13)$$

$$\frac{\partial Loss}{\partial w_{hh'}} = \sum_{t=1}^{T} \frac{\partial Loss}{\partial a_j^t} \frac{\partial a_j^t}{\partial w_{hh'}} = \sum_{t=1}^{T} \delta_j^t b_i^{t-1}. \quad (14)$$

其中,a_h^t 表示中间隐层为 h 的节点在 t 时刻的加权和,$b_{h'}^{t-1}$ 表示节点 h 在 $t-1$ 时刻的输出。

2.2.2 长短期记忆网络

长短期记忆网络(Long Short Term Memory,LSTM)是循环神经网络(RNN)的一层构建单元。RNN 对应单位常常叫做 LSTM 网络。LSTM 单元由 cell 组成。cell 负责“记住”在任意时间间隔值;所以,LSTM 中储存了大量“记忆”。在一个多层(或前馈)神经网络中,包含 3 种门。这 3 种门分别是:输入门(Input Gate)、遗忘门(Forget Gate)和输出门(Output Gate)。其中,每一种门可以被认为是一个“常规”人工神经元。也就是说,均可将其用来计算一个加权和的激活。直观地说,就是可以被认为是价值的流量调节器,经过严格的连接;因此,表示“门”。这些门和电池之间有联系。表达的短期是指虽是一个短期记忆可以持续长时间的模型。相对来说适合于分类和预测时间序列的时间滞后之间的重要事件,如未知的大小和持续时间。原始 RNN 存在一个不足,就是对深层节点的感知能力会逐渐下降,在深层网络将无法进行有效的训练。LSTM 的提出有利于代替 RNN 这种相对不敏感的学习方法。与此研究相关的原理公式详见如下。

(1)输入门。具体公式为:

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{hl} b_h^{t-1} + \sum_{c=1}^{C} w_{cl} s_c^{t-1}, \quad (15)$$

$$b_l^t = f(a_l^t), \quad (16)$$

其中,x_i^t 为输入;b_h^{t-1} 为上一时间的隐层的输出;s_c^{t-1} 为上一时间 cell 的输出。

(2)遗忘门。具体公式为:

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}, \quad (17)$$

$$b_\phi^t = f(a_\phi^t), \quad (18)$$

(3)输出门。具体公式为:

$$a_w^t = \sum_{i=1}^{I} w_{iw} x_i^t + \sum_{h=1}^{H} w_{hw} b_h^{t-1} + \sum_{c=1}^{C} w_{cw} s_c^t, \quad (19)$$

$$b_w^t = f(a_w^t). \quad (20)$$

其中,w_{ij} 表示一种从 i 到 j 的连接权重;a_j^t 表示时间 t 的网络输入;b_j^t 表示时间 t 的网络输出;l 表示输入门;Ø 表示遗忘门;w 表示输出门。

2.2.3 卷积神经网络

在机器学习领域中,卷积神经网络(Convolutional Neural Network,CNN)是一种属于深度学习网络范畴的前馈人工神经网络,非常适合应用于语音信号识别分析。CNN 的设计使用变化多层感知器,要求最小化的预处理[14]。故而也被称为平移稳定或空间稳定的人工神经网络(siann),且具有基于共同的权重结构和平移稳定性的特点[15-16]。这里,给出卷积神经网络的模型架构如图 2 所示。卷积网络的灵感来自于生物过程[17],这种网络不同神经元之间的连接方式是受到动物皮层组织之间的模式启发而设计的。如前所述,只有受约束的可观测区域会刺激独立的神经单位,将其称作感受野。这意味着深度网络引用了传统算法中手工设计的过滤器。在特征设计中,这种与先前知识和人工学习相分离的独立性是一个主要优点。现今,在图像识别、语音识别、推荐系统[18]和语言识别分析中都可见到其应用的实例[19]。

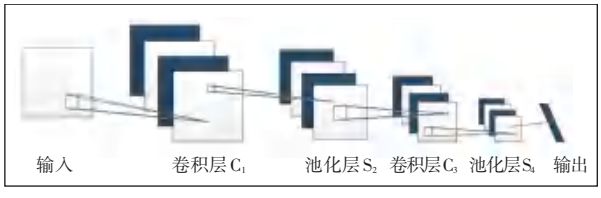

图 2 卷积神经网络模型

Fig. 2 Convolutional Neural Network Model

2.2.4 深度神经网络

传统语音识别基本上均使用 GMM-HMM 作为声学模型,而 DNN-HMM 的声学模型最显著的不同点就是使用深度神经网络(Deep Neural Network,DNN)将 GMM-HMM 中的高斯混合模型替换掉,对输入信号进行建模从而观察概率。DNN 模型输入的频谱特征与传统的方法有很大的区别。MFCC 比较常见。DNN 形成的语音波均经过加窗、分帧。如图 3 所示,DNN 与 GMM 的不同可表述为:DNN 会进行拼接帧的操作,而 GMM 只会采集单帧特征作

为输入。在本文会将相邻多帧拼接，从而得到汇集更多数据的输入量。采用拼接帧是优化效率的一种重要手段。DNN 是一个具有众多隐含层的多层网络，信号接送至输入层后分多条线路传输到隐含层，从采集的原始声音特征映射到新特征空间中，这种新特征空间是通过隐含层各节点构成的，从而得到一种新的特征表现形式。每一层隐含层都会对上一层的语音信号进行分解，并且在本层加以重组。当信号到达最后一个隐含层时，会通过深度学习网络映射到状态空间；绘制出 2 个模型的结构，从中可以看出深度神经网络包含多个高斯混合模型，一个高斯混合模型可以被当作仅含有一个隐含层的神经网络，各个高斯混合分量作为隐含层节点，由下层的各个混合分量经由线性组合而成的输出层可被当作 HMM 模型的状态。通过将采集的声学特征映射到 GMM 混合分量空间，从而在 HMM 模型中实现映射，最后在得到的状态中得到输出后验概率。基于前述研究可以看出，DNN 的建模能力要优于 GMM，因此 DNN-HMM 是一种更加高效的声学模型。

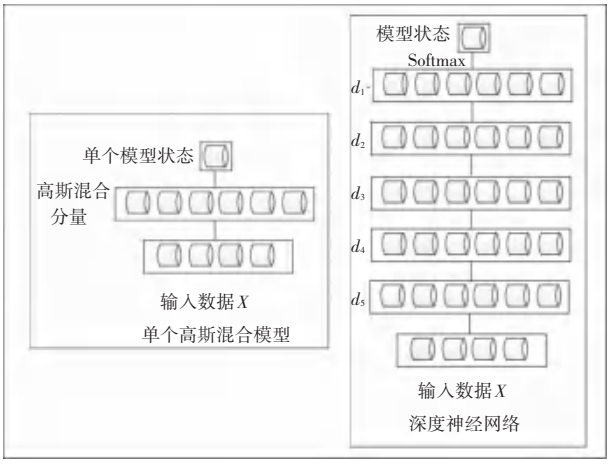

图 3　单个高斯混合模型与深度神经网络模型

Fig. 3　Single Gaussian Mixture Model and Deep Neural Network Model

DNN 将 GMM-HMM 模型中的 GMM 进行置换，从而计算 HMM 状态的后验概率。设给定时刻 T 的特征观察矢量是 O_π，在 DNN 中采用 Softmax 函数计算 HMM 状态出现的概率，状态为：

$$y_w(s) = P(s \mid O_{vt}) = \frac{\exp\{a_{vt}(x)\}}{\sum_{x'} \exp\{a_{vt}(x')\}}, \quad (21)$$

其中，$\{a_{vt}(x)\}$ 为输出层 x 的激活概率。在此基础上，还将推得：

$$\log p(O_{vt} \mid x) = \log y_{vt}(x) - \log P(x), \quad (22)$$

其中，$P(x)$ 表示训练数据中状态 x 出现的先验概率。

在 DNN-HMM 模型中，DNN 的原理是将采集的输入信号的后验概率进行计算建模。对观察概率建模是传统 GMM 模型的模式。因此研究中既需要获取先验概率，又要获取后验概率，将二者相结合从而得到观察概率。设输出样本为 a，输出状态为 x，$P(x \mid a)$ 表示 DNN 后验概率，可采用式(23)进行计算：

$$P(x \mid a) = \frac{P(a \mid x)\, P(a)}{P(x)}. \quad (23)$$

通过以上模型观察，就可以得出概率并且利用 HMM 进行解码。

2.3　数据来源

本文使用了 863 汉语语音库。该语音库分别由 41 名男性和 41 名女性说话人组成训练集，42 名男性和 42 名女性说话人组成测试集。由上世纪 90 年代《人民日报》中选取千余句作为朗读文本。设置采样频率 16 KHz，分 3 类，并且选择其中一类进行阅读。

3　基于深度学习的语音识别声学建模实验

本节将深度学习技术建立声学模型，与传统声学模型进行对比，得到各模型的建模能力。实验中进行了如下的准备工作：本节在 863 语音库中分别选取 1 000 单词和句子作为样本。随机抽取 500 作为训练集合，500 作为测试集，将 Kaldi 搭载于 Linux 系统下。

3.1　实验步骤

3.1.1　特征提取

首先，要对传统 GMM-HMM 声学模型进行训练，从数据库中提取对应帧长为 20 ms、帧移为 10 ms 的语音数据对应 MFCC 特征，该提取特征有 40 维。将离散余弦变换阶数选取一个假定值，设为 13，经过一阶和二阶的差分后得到的是 39 维，在此基础上叠加帧能量，共得到 40 维。其次，分别选择长短期记忆网络、卷积神经网络以及深度神经网络声学模型进行训练，从数据库中提取对应帧长为 20 ms、帧移为 10 ms 的语音数据的特征。CNN-HMM 的语音数据对应的是 fBank 特征，提取的特征有 40 维。DNN-HMM 的语音数据对应的是 fBank 特征，提取的特征有 96 维。提取的特征需要分组捆绑，这是为了提升相关性分析的效率。

3.1.2　生成标签

在实验的过程中需要生成标签，用于监测

LSTM 网络、CNN 网络以及 DNN 网络的性能。搭建传统 GMM-HMM 声学模型,从中获取标签信息。将这个成熟的声学模型与有关信息相融合,再与初始文本标签进行对接。使用对接后的三因素模型用作声学模型训练的标签。

3.1.3 声学建模

3.1.3.1 LSTM-HMM 网络参数配置

该 LSTM-HMM 由 1 个输入层、3 个隐藏层、1 个输出层组成。输入层对应 75 维特征,扩展 3 帧,从而得到的节点数为 300。每个隐藏层对应 2 048 个节点,共计 6 144 个节点。输出层对应 36 016 个节点。使用 Softmax 函数用作输出层的分类,使用 Sigmoid 函数对隐藏层进行激活。

选取最小化交叉熵设定为目标函数,用来进行参数调优。设定起始学习率的值为 0.1,开始训练直到第 5 代,将学习率降低二分之一。接下来,每当迭代一次,学习率都将降低为上代的二分之一。一旦交叉验证值趋于平稳,结束实验。

3.1.3.2 CNN-HMM 网络参数配置

该 CNN-HMM 由 1 个输入层、2 个卷积层、2 个池化层和 1 个输出层组成。输入层对应 96 维特征,扩展 5 帧,从而得到节点数为 1 056。该网络的卷积层与池化层并不同时运行,而是在不同时间段内交替出现。C_1 是首次出现的卷积层,C_3 是第二次出现的卷积层。S_2 是首次出现的池化层,S_4 是第二次出现的池化层。使用 Softmax 函数用作层的分类。

参数调整与 LSTM 相同,选取最小化交叉熵设定为目标函数,用来进行参数调优。设定起始学习率的值为 0.1,开始训练直到第 5 代,将学习率降低二分之一。后续每当迭代一次,学习率都将降低为为上代的二分之一。一旦平均惩罚值维持在一定范围内趋于平稳,结束实验。

3.1.3.3 DNN-HMM 网络参数配置

该 DNN-HMM 由 1 个输入层、6 个隐藏层、1 个输出层组成。输入层对应 429 个节点,每个隐藏层对应 1 024 个节点,共计 6 144 个节点,输出层对应 1 366 个节点。使用 Softmax 函数用作输出层的分类,使用 Sigmoid 函数对隐藏层进行激活。

参数调整也是与 LSTM 相同,选取最小化交叉熵设定为目标函数,用来进行参数调优。设定起始学习率的值为 0.1,开始训练直到第 5 代,将学习率降低二分之一。后续每当迭代一次,学习率都将降低为上代的二分之一。一旦交叉验证值逐渐平稳,结束实验。

3.2 实验结果及分析

依照实验步骤分别在 Kaldi 系统中搭建 GMM-HMM 声学模型、LSTM-HMM 声学模型、CNN-HMM 声学模型以及 DNN-HMM 声学模型。最终得出结果见表 1。进而,研究中得到的各主要算法的实验仿真结果则如图 4~图 6 所示。

表 1 主要算法的单词句子正确率的测试结果

Tab. 1 Test results of the sentence correct rate of the main algorithms

测试方法	测试集	识别正确个数	词正确率/%	识别正确个数	句子正确率/%
GMM-HMM	500	407	81.4	369	73.8
LSTM-HMM	500	444	88.8	434	86.8
CNN-HMM	500	453	90.6	430	86.0
DNN-HMM	500	439	87.8	421	84.2

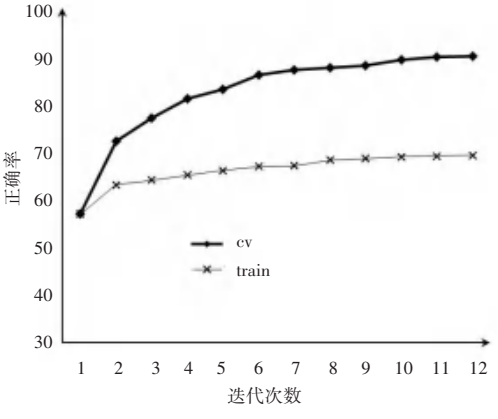

图 4 LSTM 正确率迭代次数变化图

Fig. 4 Iteration number change graph of LSTM correct rate

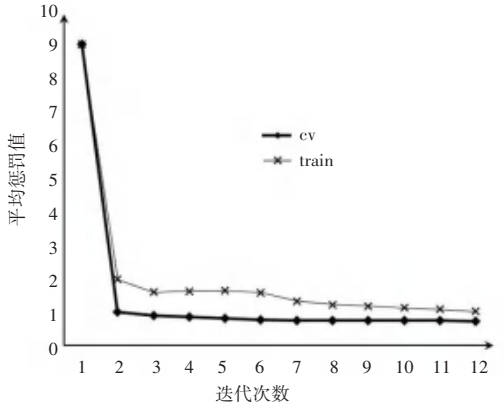

图 5 CNN 平均惩罚值迭代次数变化图

Fig. 5 Iteration number change graph of CNN average penalty value

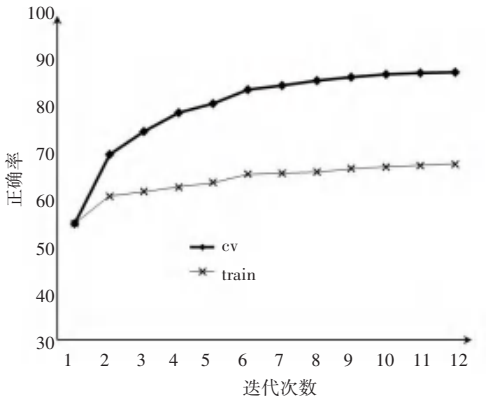

图 6　DNN 正确率迭代次数变化图

Fig. 6　Iteration number change graph of DNN correct rate

结合上述结果分析后可知，GMM 用作声学模型在网络学习中趋于浅层，而 LSTM、CNN、DNN 等模型属于深层学习范畴，虽然深度学习模型相较于传统声学模型更为复杂，但是能够显著提高语音识别正确率及效率。可以看出，基于深度学习的声学建模能力要普遍强于传统 GMM-HMM 声学模型。

4　结束语

研究前文表 1 可以看出，传统基于混合高斯隐马尔可夫混合模型的建模方法显然要逊色于当下基于深度学习进行声学建模的方法。从词的正确率来看，LSTM-HMM 比 GMM-HMM 提升了 7.4%、CNN-HMM 比 GMM-HMM 提升了 9.2%、DNN-HMM 比 GMM-HMM 提升了 6.4%。从句子的正确率来看，LSTM-HMM 比 GMM-HMM 提升了 13%、CNN-HMM 比 GMM-HMM 提升了 12.2%、DNN-HMM 比 GMM-HMM 提升了 10.4%。分析如上数据可以得到，GMM 是一种趋于传统的浅层网络，虽然技术相对成熟，但是对海量数据的提取学习能力显得有些薄弱。而 LSTM、CNN、DNN 等深层网络适合于进行深度学习，而且非常适于在海量数据中进行数据提取。LSTM 模型能够对数据进行长期记忆，不会形成记忆断层，使其建模能力将会优于 DNN 模型。CNN 模型由于其独特的网络结构，根据得到的反馈来看，效果具有明显提升。因此可以得出以下结论：DNN 模型相对于 GMM 模型来说是一种建模能力更强的模型。使用 LSTM-HMM 模型建模得到的反馈略优于 DNN-HMM 模型，而 CNN-HMM 的建模能力却胜过其余 3 种模型。

随着人类社会的进步与现代化发展，语音识别能够将人机实现充分的有机结合。在海量数据的处理中以及提升准确性的前提下去优化其识别效率是每位相关研究人员的追求目标。本文希望通过基于深度学习的语音识别方法的研究，能够改善语音识别的声学模型，提高语音识别的准确性，优化其效率，从而有效满足多个领域对人工智能语音识别的各项丰富需求。

参考文献

[1] 赵力. 语音信号处理[M]. 3 版. 北京：机械工业出版社，2016.

[2] JUANG B H, RABINER L R. Automatic speech recognition-a brief history of the technology development[EB/OL]. [2015-01-17]. http://www.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf.

[3] BENESTY J, SONDHI M M, HUANG Yiteng. Springer handbook of speech processing[M]. Berlin/Heidelberg: Springer Science & Business Media, 2008.

[4] RABINER L R. First-Hand: The_Hidden_Markov_Model[EB/OL]. [2015-01-12]. https://ethw.org/First-Hand:The_Hidden_Markov_Model.

[5] PINOLA M. Speech recognition through the decades: How we ended up with Siri[EB/OL]. [2017-07-28]. https://www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html.

[6] KURZWEIL R. KurzweilAINetwork[EB/OL]. [2014-09-25]. http://www.kurzweilai.net/ray-kurzweil-biography.

[7] HERVÉ B, MORGAN N. Connectionist speech recognition: A hybrid approach[M]. Boston: Kluwer Academic Publishers, 1994.

[8] BAKER J M, LI Deng, GLASS J R, et al. Developments and directions in speech recognition and understanding, Part 1[J]. IEEE Signal Processing Magazine, 2009, 26(3): 75-80.

[9] 孙志军，薛磊，许阳明，等. 深度学习研究综述[J]. 计算机应用研究，2012，29(8)：2806-2810.

[10] Kaldi. History of the Kaldi project[EB/OL]. [2017-07-26]. http://www.kaldi-asr.org/doc/history.html.

[11] GRAVES A, LIWICKI M, FERNÁNDEZ S, et al. A novel connectionist system for unconstrained handwriting recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(5):855-868.

[12] SAK H, SENIOR A, BEAUFAYS F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling[C]//Interspeech 2014. Singapore: ISCA, 2014: 338-342.

[13] LI Xiangang, WU Xihong. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition[J]. arXiv preprint arXiv:1410.4281, 2014.

[14] LECUN Y. LeNet-5, convolutional neural networks[EB/OL]. [2013-11-16]. http://sites.google.com/site/chumerin/projects/mycunn.

[15] ZHANG Wei. Shift-invariant pattern recognition neural network and its optical architecture[C]//Proceedings of annual conference of the Japan Society of Applied Physics, 1988.

[16] ZHANG W, ITOH K, TANIDA J, et al. Parallel distributed processing model with local space-invariant interconnections and its optical architecture[J]. Applied Optics, 1990, 29(32): 4790-4797.

[17]MATSUGU M, MORI K, MITARI Y, et al. Subject independent facial expression recognition with robust face detection using a convolutional neural network [J]. Neural Networks, 2003, 16(5): 555-559.

[18]VAN DEN OORD A, DIELEMAN S, SCHRAUWEN B. Deep content-based music recommendation (PDF) [C]//NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, Nevada: ACM, 2013, 2: 2643-2651.

[19]COLLOBERT R, WESTON J. A unified architecture for natural language Processing: Deep neural networks with multitask learning [C]//ICML, volume 307 of ACM International Conference Proceeding Series. Helsinki, Finland: ACM, 2008: 160-167.

(上接第134页)

[4] CHANG Pichuan, GALLEY M, MANNING C D. Optimizing Chinese word segmentation for machine translation performance [C] // Proceedings of the Third Workshop on Statistical Machine Translation-StatMT'08. Columbus, Ohio: ACM, 2008: 224-232.

[5] TOUTANOVA K, MANNING C D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger[C] // Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. Hong Kong: ACL, 2000, 13: 63-70.

[6] TOUTANOVA K, KLEIN D, MANNING C D, et al. Feature-rich part-of-speech tagging with a cyclic dependency network [C] //Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-NAACL'03. Edmonton, Canada: ACL, 2003, 1: 173-180.

[7] FINKEL J, DINGARE S, MANNING C D, et al. Exploring the boundaries: Gene and protein identification in biomedical text[J]. BMC Bioinformatics, 2005, 6(SUPPL.1): S5.

[8] XIA F. The segmentation guidelines for the Penn Chinese treebank 3.0[R]. USA: University of Pennsylvania, 2000.

[9] Annotation guidelines for word segmentation on Chinese clinical text[EB/OL]. [2016]. https://github.com/WILAB-HIT/Resources/blob/master/segmentation_pos_parsing/annotation_guidelines/Seg.pdf.

[10]HE Bin, DONG Bin, GUAN Yi, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts[J]. Journal of Biomedical Informatics, 2017, 69: 203-217.

[11]World Health Organization. The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines[M]. Geneva: World Health Organization, 1992.

[12] National Library of Medicine. Medical subject headings. main headings, subheadings, and cross references used in the index medicus and the national library of medicine catalog [M]. Washington, DC: U.S. Department of Health, Education, and Welfare, 1960.

[13]DONNELLY K. SNOMED-CT: The advanced terminology and coding system for eHealth[J]. Studies in health technology and informatics, 2006, 121: 279-290.

[14]SAVOVA G K, MASANZ J J, OGREN P V, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications[J]. Journal of the American Medical Informatics Association, 2010, 17(5): 507-513.

[15] UZUNER Ö, SOUTH B R, SHEN S, et al. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text [J]. Journal of the American Medical Informatics Association, 2011, 18(5): 552-556.

[16]WEED L L. Medical records that guide and teach[J]. The New England Journal of Medicine, 1968, 278(11): 593-600.

[17]DE BRUIJN B, CHERRY C, KIRITCHENKO S, et al. Machine-learned solutions for three stages of clinical information extraction: The state of the art at i2b2 2010[J]. Journal of the American Medical Informatics Association, 2011, 18(5): 557-562.

[18]TANG Buzhou, CAO Hongxin, WU Yonghui, et al. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features[J]. BMC Medical Informatics and Decision Making, 2013, 13(SUPPL1): S1.

[19]LV Xinbo, GUAN Yi, DENG Benyang. Transfer learning based clinical concept extraction on data from multiple sources[J]. Journal of Biomedical Informatics, 2014, 52: 55-64.

[20]叶枫, 陈莺莺, 周根贵, 等. 电子病历中命名实体的智能识别[J]. 中国生物医学工程学报, 2011, 30(2): 256-262.

[21]LEI Jiangbo, TANG Buzhou, LU Xueqin, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814.

[22] WU Y, JIANG M, LEI J, et al. Named entity recognition in Chinese clinical text using deep neural network[J]. Studies in Health Technology and Informatics, 2015, 216: 624-628.

[23] WANG Hui, ZHANG Weide, ZENG Q, et al. Extracting important information from Chinese operation notes with natural language processing methods [J]. Journal of Biomedical Informatics, 2014, 48: 130-136.

[24] WANG Yaqing, YU Zhonghua, CHEN Li, et al. Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study[J]. Journal of Biomedical Informatics, 2014, 47: 91-104.