DOI: 10.11992/tis.201710029
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20180408.1725.028.html

基于深度神经网络的蒙古语声学模型建模研究

马志强, 李图雅, 杨双涛, 张力
（内蒙古工业大学 数据科学与应用学院, 内蒙古 呼和浩特 010080）

摘　要: 针对高斯混合模型在蒙古语语音识别声学建模中不能充分描述蒙古语声学特征之间相关性和独立性假设的问题, 开展了使用深度神经网络模型进行蒙古语声学模型建模的研究。以深度神经网络为基础, 将分类与语音特征内在结构的学习紧密结合进行蒙古语声学特征的提取, 构建了 DNN-HMM 蒙古语声学模型, 结合无监督预训练与监督训练调优过程设计了训练算法, 在 DNN-HMM 蒙古语声学模型训练中加入 dropout 技术避免过拟合现象。最后, 在小规模语料库和 Kaldi 实验平台下, 对 GMM-HMM 和 DNN-HMM 蒙古语声学模型进行了对比实验。实验结果表明, DNN-HMM 蒙古语声学模型的词识别错误率降低了 7.5%, 句识别错误率降低了 13.63%; 同时, 训练时加入 dropout 技术可以有效避免 DNN-HMM 蒙古语声学模型的过拟合现象。

关键词: 语音识别; 声学模型; GMM-HMM; DNN-HMM; 监督学习; 预训练; 过拟合; dropout
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2018)03−0486−07

中文引用格式: 马志强, 李图雅, 杨双涛, 等. 基于深度神经网络的蒙古语声学模型建模研究[J]. 智能系统学报, 2018, 13(3): 486–492.
英文引用格式: MA Zhiqiang, LI Tuya, YANG Shuangtao, et al. Mongolian acoustic modeling based on deep neural network[J]. CAAI transactions on intelligent systems, 2018, 13(3): 486–492.

Mongolian acoustic modeling based on deep neural network

MA Zhiqiang, LI Tuya, YANG Shuangtao, ZHANG Li
(School of Data Science &Application, Inner Mongolia University of Technology, Hohhot 010080, China)

Abstract: Considering the difficulty of using the Gaussian mixture model (GMM) to adequately describe the correlation and independence hypothesis of the Mongolian acoustic features in the acoustic modeling of Mongolian speech recognition, this study investigates an acoustic model based on deep neural network (DNN). Firstly, using DNN, the internal structure of phonetic features were classified and learned to extract the Mongolian acoustic features, and a DNN-HMM Mongolian acoustic model was constructed. Secondly, a training algorithm was designed by combining unsupervised pre-training and supervised training tuning. In addition, dropout technology was added into the DNN-HMM Mongolian acoustic model training to avoid the over-fitting phenomenon. Finally, a comparative experiment was conducted for the GMM-HMM and DNN-HMM Mongolian acoustic models on basis of the small-scale corpus and Kaldi experimental platform. Experimental results show that the word recognition error rate of DNN-HMM Mongolian model was reduced by 7.5% and sentence recognition error rate was reduced by 13.63%. In addition, the over-fitting of DNN-HMM Mongolian acoustic model can be effectively avoided by adopting the dropout technique during training.

Keywords: speech recognition; acoustic model; GMM-HMM; DNN-HMM; supervised learning; pre-training; over-fitting; dropout

典型的大词汇量连续语音识别系统 (large vocabulary continuous speech recognition, LVCSR) 由特征提取、声学模型、语言模型和解码器等组成。声学模型是语音识别系统的核心组成部分, 基于 GMM 和 HMM 模型构建的 GMM-HMM 声学模型[1]一度是大词汇量连续语音识别系统中应用最广的声学模型。在 GMM-HMM 模型中, GMM 模型对语音特

收稿日期: 2017−10−31.　**网络出版日期:** 2018−04−09.
基金项目: 国家自然科学基金项目 (61762070, 61650205).
通信作者: 李图雅. E-mail: 2297854548@qq.com.

征向量进行概率建模, 然后通过 EM 算法生成语音观察特征的最大化概率, 当混合高斯分布数目足够多时, GMM 可以充分拟合声学特征的概率分布, HMM 模型根据 GMM 拟合的观察状态生成语音的时序状态[2-3]。当采用 GMM 混合高斯模型的概率来描述语音数据分布时, GMM 模型本质上属于浅层模型, 并在拟合声学特征分布时对特征之间进行了独立性的假设, 因此无法充分描述声学特征的状态空间分布; 同时, GMM 建模的特征维数一般是几十维, 不能充分描述声学特征之间的相关性, 模型表达能力有限。因此, 在 20 世纪 80 年代利用神经网络和 HMM 模型构建声学模型的研究开始出现, 但是, 当时计算机计算能力不足且缺乏足够的训练数据, 模型的效果不及 GMM-HMM[4-5]。2010 年微软亚洲研究院的邓力与 Hinton 小组针对大规模连续语音识别任务提出了 CD-DBN-HMM的混合声学模型框架[6], 并进行了相关实验。实验结果表明, 相比 GMM-HMM 声学模型, 采用 CD-DBN-HMM 声学模型使语音识别系统识别正确率提高了 30% 左右, CD-DBN-HMM 混合声学模型框架的提出彻底革新了语音识别原有的声学模型框架。与传统的高斯混合模型相比, 深度神经网络属于深度模型, 能够更好地表示复杂非线性函数, 更能捕捉语音特征向量之间的相关性, 易于取得更好的建模效果[7-12]。蒙古语语音识别研究主要借鉴了英语、汉语以及其他少数民族语言, 在语音识别研究上取得了成果, 因此蒙古语声学模型建模过程主要以 GMM-HMM 模型为基础开展研究, 也取得了一定的研究成果[13-16]。在特征学习方面 DNN 模型比 GMM 模型具有更大的优势, 所以本文用 DNN 模型代替了 GMM 模型来完成蒙古语声学模型建模任务。

1　蒙古语声学模型研究

在语音识别领域内, DNN 主要以两种形式被应用: 直接作为声学特征的提取模型, 但是这种应用方式仍需要借助 GMM-HMM 模型才能完成; 将 DNN 与 HMM 隐马尔科夫模型进行结合, 构成混合模型结构, 利用深度神经网络代替 GMM 高斯混合模型进行声学状态输出概率的计算[7-8]。与高斯混合模型相比, 深度神经网络有着更强的学习能力和建模能力, 能够更好地捕捉声学特征的内在关系, 有助于声学模型性能的提升, 所以本文通过使用深度神经网络模型对蒙古语声学特征逐层提取, 将分类与语音特征内在结构的学习进行了紧密结合, 有利于蒙古语语音识别系统正确率的提升。

1.1　DNN-HMM 蒙古语声学模型

DNN-HMM 蒙古语声学模型就是将深度神经网络技术应用到蒙古语声学模型中, 用 DNN 深度神经网络代替 GMM 高斯混合模型, 实现对蒙古语声学状态的后验概率估算。在给定蒙古语声学特征序列的情况下, 首先用 DNN 模型估算当前特征属于 HMM 状态的概率, 然后用 HMM 模型描述蒙古语语音信号的动态变化, 捕捉蒙古语语音信息的时序状态信息。DNN-HMM 蒙古语声学模型结构如图 1 所示。

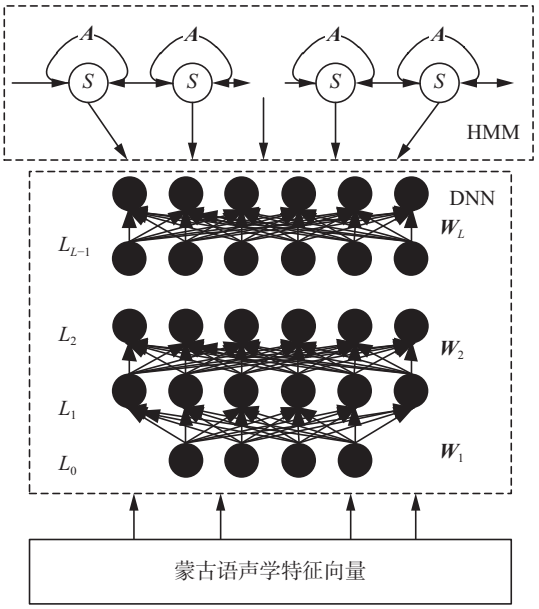


图 1　DNN-HMM 蒙古语声学模型

Fig. 1　The Mongolian acoustic model based on DNN-HMM.

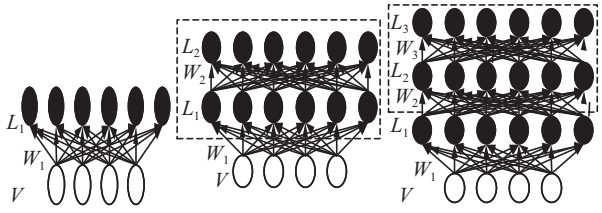
在 DNN-HMM 蒙古语声学模型中, DNN 网络是通过不断地自下而上堆叠隐含层实现的。其中, S 表示 HMM 模型中的隐含状态, $\boldsymbol{A}$ 表示状态转移概率矩阵, L 表示 DNN 深度神经网络的层数 (隐含层为 L-1 层, L_0层为输入层, L_L层为输出层, DNN 网络共包含 L+1 层), $\boldsymbol{W}$ 表示层之间的连接矩阵。DNN-HMM 蒙古语声学模型在进行蒙古语语音识别过程建模前, 需要对 DNN 神经网络进行训练。在完成 DNN 神经网络的训练后, 对蒙古语声学模型的建模过程与 GMM-HMM 模型一致。

1.2　DNN 网络的训练

蒙古语声学模型中的 DNN 网络的训练分为预训练和调优两个阶段。DNN 的预训练就是对深度神经网络的参数进行初始化。通常, DNN 深度神经网络的预训练方式分为生成式训练和判别式训练。逐层无监督预训练算法就是使用无监督学习方法对网络的每一层进行预训练, 它属于生成式训练算法[17]。在 DNN-HMM 蒙古语声学模型预训练中, 采用了逐层无监督训练算法。

DNN 模型是一个多层次的神经网络, 逐层无监督预训练算法是对 DNN 的每一层进行训练, 而且每次只训练其中一层, 其他层参数保持原来初始化参数不变, 训练时, 对每一层的输入和输出误差尽量减小, 这样就能够保证每一层参数对于该层来说都是最优的。接下来, 将训练好的每一层的输出数据作为下一层的输入数据, 那么下一层输入的数据就比直接训练时经过多层神经网络输入到下一层数据的误差小得多, 逐层无监督预训练算法能够保证每一层之间输入输出数据的误差都相对较小。

具体训练过程如图 2 所示, 训练算法见算法 1。



(a) 第1层隐含层训练　(b) 第2层隐含层训练　(c) 第3层隐含层训练

图 2　DNN-HMM 蒙古语声学模型预训练过程

Fig. 2　The pre-training DNN-HMM process for Mongolian acoustic model.

算法 1　逐层无监督预训练算法

输入　学习率 α, 最大迭代次数 T, 需要训练的层数 L; 各隐含层内的隐单元个数 $N=(n^1,n^2,\cdots,n^L)$; 训练数据按 mini-batch 划分后的序列 X^j, 其中 $j=(1,2,\cdots,\text{Max})$, 序列长度 Max。

输出　链接权重 $\boldsymbol{W}^i$, $\boldsymbol{i}=(1,2,\cdots,L)$; 偏执向量 $\boldsymbol{b}^i$, $i=(0,1,\cdots,L)$。

1) 初始化输入层的偏执向量;

2) For i in 1 to L do;

3) 初始化 $\boldsymbol{W}^i=0$, $\boldsymbol{b}^i=0$;

4) For t in 1 to T do;

5) For j in 1 to Max do;

6) mini-batch = X^j ;

7) DNNUpdate (mini-batch, α, W^i, b^i, b^{i-1});

8) End For;

9) End For;

10) End For;

其中 DNNUpdate 算法采用经典的对比散度算法 (contrastive divergence, CD-K), 具体见文献[7]。

通过逐层无监督预训练算法可以得到较好的神经网络初始化参数, 然后使用蒙古语标注数据 (即特征状态) 通过 BP(error back propagation) 算法进行有监督的调优, 最终得到可用于声学状态分类的 DNN 深度神经网络模型。有监督的调优算法采用随机梯度下降算法进行实现, 具体见算法 2。

算法 2　随机梯度下降算法

输入　训练集 set, 批量大小 batch_size; 学习率 α, 循环次数 epoch。

输出　模型参数 weight。

1) weight←initWeight();

2) For j in 0 to epoch do;

3) batch←randomSelect(set, batch_size);

4) weight←getWeightFromMaster();

5) ΔW←miniGradient(batch, weight);

6) weight←weight- α * ΔW;

7) End for;

1.3　蒙古语语音数据识别

通过对 DNN 网络的预训练和调优后, 可以利用 DNN-HMM 声学模型对蒙古语语音数据进行识别, 具体的过程如下。

首先, 根据输入的蒙古语声学特征向量, 计算 DNN 深度神经网络前 L 层的输出, 即

$$\boldsymbol{v}^\alpha=f(\boldsymbol{z}^\alpha)=f(\boldsymbol{W}^\alpha\boldsymbol{v}^{\alpha-1}+\boldsymbol{b}^\alpha),0\leqslant\alpha<L \tag{1}$$

式中: $\boldsymbol{z}^\alpha$ 表示激励向量, $\boldsymbol{z}^\alpha=\boldsymbol{W}^\alpha\boldsymbol{v}^{\alpha-1}+\boldsymbol{b}^\alpha$ 且 $\boldsymbol{z}^\alpha\in R^{N_\alpha X1}$; $\boldsymbol{v}^\alpha$ 表示激活向量, $\boldsymbol{v}^\alpha\in R^{N_\alpha X1}$; $\boldsymbol{W}^\alpha$ 表示权重矩阵, $\boldsymbol{W}^\alpha\in R^{N_\alpha XN_{\alpha-1}}$; $\boldsymbol{b}^\alpha$ 表示偏执向量, $\boldsymbol{b}^\alpha\in R^{N_\alpha X1}$, N_α 表示第 α 层的神经节点个数且 $N_\alpha\in R$; $\boldsymbol{V}^0$ 表示网络的输入特征, $\boldsymbol{V}^0=\boldsymbol{o}\in R^{N_0X1}$。在 DNN-HMM声学模型中, 输入特征即为声学特征向量。其中 $N_0=D$ 表示输入声学特征向量的维度, $f(\cdot):R^{N_\alpha X1}\rightarrow R^{N_\alpha X1}$ 表示激活函数对激励向量的计算过程, $f(\cdot)$ 表示激活函数。

然后, 利用 L 层的 softmax 分类层计算当前特征关于全部声学状态的后验概率, 即当前特征属于各蒙古语声学状态的概率:

$$v^i=P_{\text{dnn}}(i|\boldsymbol{O})=\text{softmax}(i) \tag{2}$$

在 DNN-HMM 蒙古语声学模型中, DNN 深度神经网络用于估计每个 HMM 状态的后验概率, 所以 DNN 的输出是按照 HMM 隐含状态进行分类输出的, 实质上属于多分类任务, 因此 DNN 的输出层通常是 softmax 分类层。而且 softmax 分类层的神经单元个数与 HMM 声学模型中的隐含状态个数相同。在式 (2) 中, $i=1,2,\cdots,C$, 其中 C 表示声学模型的隐含状态个数, v^i 表示 softmax 分类层第 i 个神经单元的输出, 即输入声学特征向量 $\boldsymbol{O}$ 关于声学模型第 i 个隐含状态的后验概率。得到隐含状态的后验概率后, 利用维特比解码算法进行解码得到最优路径。在直接解码前需要根据贝叶斯公式, 将各个状态的后验概率除以其自身的先验概率, 得到各状态规整的似然值。隐含状态的先验概率计算较为简单, 仅通过计算各状态对应帧总数与全部声学特征帧数的比值即可得到。

2　蒙古语声学模型的调优训练

由于 DNN 模型在调优时需要对齐的语音帧标注数据, 同时标注数据质量往往影响 DNN 模型的性能, 因此, 在 DNN 网络调优阶段, 通过使用已训练好的 GMM-HMM 蒙古语声学模型生成对齐的蒙古语语音特征标注数据。

所以, DNN-HMM 蒙古语声学模型的训练过程为: 首先训练 GMM-HMM 蒙古语声学模型, 得到对齐的蒙古语语音特征标注数据; 然后在对齐语音特征数据的基础上对深度神经网络 (DNN) 进行训练和调优; 最后根据得到的蒙古语语音观察状态再对隐马尔科夫模型 (HMM) 进行训练。具体见 DNN-HMM 蒙古语声学模型训练过程。

DNN-HMM 蒙古语声学模型训练过程:

输入　蒙古语语料库。

输出　DNN-HMM 声学模型。

1) 进行 GMM-HMM 蒙古语声学模型训练, 得到一个最优的 GMM-HMM 蒙古语语音识别系统, 用 gmm-hmm 表示。

2) 利用维特比解码算法解析 gmm-hmm, 对 gmm-hmm 蒙古语声学模型中的每一个 senone 进行标号, 得到 senone_id。

3) 利用 gmm-hmm 蒙古语声学模型, 将声学状态 tri-phone 映射到相应的 senone_id。

4) 利用 gmm-hmm 蒙古语声学模型初始化 DNN-HMM 蒙古语声学模型, 主要是 HMM 隐马尔科夫模型参数部分, 最终得到 dnn-hmm1 模型。

5) 利用蒙古语声学特征文件预训练 DNN 深度神经网络, 得到 ptdnn。

6) 使用 gmm-hmm 蒙古语声学模型, 将蒙古语声学特征数据进行状态级别的强制对齐, 对齐结果为 align-raw。

7) 将 align-raw 的物理状态转换成 senone_id, 得到帧级别对齐的训练数据 align-frame。

8) 利用对齐数据 align-data 对 ptdnn 深度神经网络进行有监督地微调, 得到网络模型 dnn。

9) 根据最大似然算法, 利用 dnn 重新估计 dnn-hmm1 中 HMM 模型转移概率得到的网络模型, 用 dnn-hmm2 表示。

10) 如果 dnn 和 dnn-hmm2 上测试集识别准确率没有提高, 训练结束。否则, 使用 dnn-hmm2 对训练数据再次进行状态级别对齐, 执行 7)。

在训练过程中, 首先训练一个最优的 GMM-HMM 蒙古语语音识别数据准备系统, 目的是为 DNN 的监督调优服务。在训练 GMM-HMM 蒙古语声学模型时, 采用期望最大化算法进行无监督训练, 避免了对标注数据的要求; 然后利用蒙古语声学特征对深度神经网络进行预训练; 在深度神经网络训练的第二阶段 (即有监督调优阶段), 利用已训练的 GMM-HMM 蒙古语声学模型进行语音特征到状态的强制对齐, 得到标注数据; 最后利用标注数据对 DNN 深度神经网络进行有监督的调优。DNN 深度神经网络训练完成以后, 根据 DNN-HMM 在测试集上的识别结果决定其下一步流程。

3　实验与结果

3.1　实验方案设计

为了验证提出的 DNN-HMM 蒙古语声学模型的有效性, 设计了 3 组实验。在实验中, 将未采用 dropout 技术的 DNN-HMM 声学模型定义为 DNN-HMM, 将采用 dropout 技术的 DNN-HMM 声学模型定义为 dropout-DNN-HMM。

1) 开展 GMM-HMM、DNN-HMM 蒙古语声学模型建模实验研究, 主要观察不同声学建模单元对声学模型的性能影响, 以及对比不同类型声学模型对语音识别系统的影响。

2) 通过构建不同层数的深度网络结构的 DNN-HMM 三音子蒙古语声学模型, 开展层数对蒙古语声学模型, 以及对过拟合现象影响的实验研究。

3) 在构建 DNN-HMM 三音子蒙古语声学模型时, 通过采用 dropout 技术开展 dropout 技术对 DNN-HMM 三音子蒙古语声学模型过拟合现象影响的实验研究。

3.2　数据集

蒙古语语音识别的语料库由 310 句蒙古语教学语音组成, 共计 2 291 个蒙古语词汇, 命名为 IMUT310 语料库。语料库共由 3 部分组成: 音频文件、发音标注以及相应的蒙文文本。实验中, 将 IMUT310 语料库划分成训练集和测试集两部分, 其中训练集为 287 句, 测试集为 23 句。实验在 Kaldi 平台上完成。Kaldi 的具体实验环境配置如表 1 所示。

表 1　实验环境

Table 1　Experimental environment

项目	参数说明
操作系统	Ubuntu14.04
处理器	I5. 4×3.2 GHz
GPU 显卡	GTX 660ti 2 GB 显存
硬盘	SAT 硬盘 500 GB
Kaldi	0.9 版本
CUDA	6.5 版本

实验过程中, 蒙古语声学特征采用 MFCC 声学特征表示, 共有 39 维数据, 其中前 13 维特征由 12 个倒谱特征和 1 个能量系数组成, 后面的两个 13 维特征是对前面 13 维特征的一阶差分和二阶差分。在提取蒙古语 MFFC 特征时, 帧窗口长度为 25 ms, 帧移 10 ms。对训练集和测试集分别进行特征提取, 全部语音数据共生成 119 960 个 MFCC 特征, 其中训练数据生成的特征为 112 535 个, 测试数据生成的特征为 7 425 个。GMM-HMM 声学模型训练时, 蒙古语语音 MFCC 特征采用 39 维数据进行实验。单音子 DNN-HMM 实验时, 蒙古语 MFCC 语音特征为 13 维 (不包括一、二阶差分特征)。三音子 DNN-HMM 实验时, 蒙古语 MFCC 的特征为 39 维。

DNN 网络训练时, 特征提取采用上下文结合的办法, 即在当前帧前后各取 5 帧来表示当前帧的上下文环境, 因此, 在实验过程中, 单音子 DNN 网络的输入节点数为 143 个 (13×(5+1+5)), 三音子 DNN 网络的输入节点数为 429 个 (39×(5+1+5))。DNN 网络的输出层节点为可观察蒙古语语音音素个数, 根据语料库标注的标准, 输出节点为 27 个; DNN 网络的隐含层节点数设定为 1 024, 调优训练次数设定为 60, 初始学习率设定为 0.015, 最终学习率设定为 0.002。

3.3　实验和结果

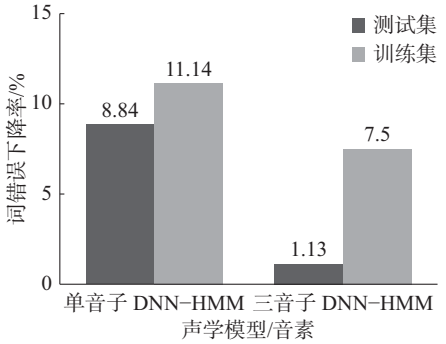
为了验证深度神经网络能够更好地捕捉蒙古语语音的声学特征, 具备更好地建模能力。本文设计了 4 个实验, 分别是单音子 GMM-HMM、三音子 GMM-HMM、单音子 DNN-HMM 和三音子 DNN-HMM 实验。采用 3.2 中的实验参数设置进行了实验, 实验结果数据见表 2。

表 2　GMM-HMM 与 DNN-HMM 蒙古语声学模型实验数据

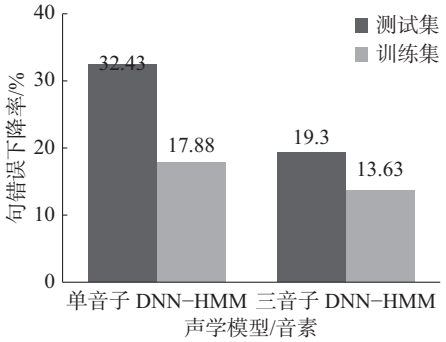
Table 2　The experimental data of Mongolian acoustic mode from GMM-HMM and DNN-HMM

声学模型	词错误率/%		句错误率/%	
	训练集	测试集	训练集	测试集
单音子 GMM-HMM	14.32	47.34	41.71	75.3
单音子 DNN-HMM	5.48	36.2	9.28	57.42
三音子 GMM-HMM	5.86	33.2	30.74	60.39
三音子 DNN-HMM	4.53	25.7	11.44	46.76

从图 3(a) 中可以发现, 相对于单音子 GMM-HMM 蒙古语声学模型, 单音子 DNN-HMM 蒙古语声学模型在训练集上的词错误率降低了 8.84%, 在测试集上的词识别错误率降低了 11.14%; 但是, 对于三音子模型来说, 三音子 DNN-HMM 蒙古语声学模型比三音子 GMM-HMM 蒙古语声学模型在训练集上的词错误率降低了 1.33%, 在测试集上的词识别错误率降低了 7.5%。由图 3(b) 发现, 单音子模型在训练集上的句识别错误率降低了 32.43%, 在测试集上的句识别错误率降低了 17.88%; 对于三音子模型来说, 三音子 DNN-HMM 蒙古语声学模型比三音子 GMM-HMM 蒙古语声学模型在训练集上的句识别错误率降低了 19.3%, 在测试集上的句识别错误率降低了 13.63%。



(a) 相对于GMM−HMM声学模型的词错误下降率



(b) 相对于GMM−HMM声学模型的句错误下降率

图 3　相对于 GMM-HMM 声学模型的实验对比结果

Fig. 3　The experimental results are compared with the GMM-HMM acoustic model

从以上分析可以得出: 单音子 DNN-HMM 蒙古语声学模型明显优于单音子 GMM-HMM 蒙古语声学模型; 对于三音子模型来说, 三音子 DNN-HMM 蒙古语声学模型比三音子 GMM-HMM 蒙古语声学模型的识别率还要高。

另外, 为了研究隐含层层数、dropout 技术[18-20]对 DNN-HMM 三音子蒙古语声学模型的影响, 本文以未采用 dropout 技术的 4 层三音子 DNN-HMM 蒙古语声学模型为基准实验, 分别进行了关于隐含层层数以及 dropout 技术的对比实验, 实验结果数据见表 3。

表 3　三音子 DNN-HMM 声学模型上 dropout 实验

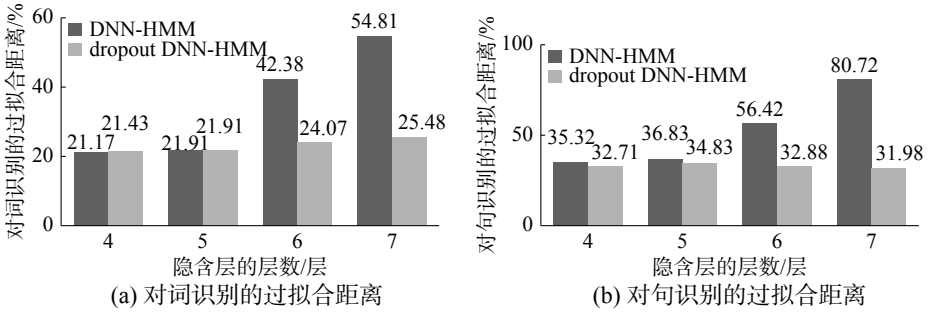
Table 3　Dropout experiment on Triphone DNN-HMM acoustic model

声学模型	隐含层层数	词错误率/%		句错误率/%	
		训练集	测试集	训练集	测试集
三音子 DNN-HMM	4	4.53	25.7	11.44	46.76
	5	4.49	26.4	12.19	49.02
	6	3.23	45.7	9.81	66.23
	7	2.11	56.92	7.2	87.92
三音子 dropout-DNN-HMM	4	4.67	26.1	12.27	44.98
	5	4.49	26.4	12.19	47.02
	6	5.27	29.34	15.33	48.21
	7	8.32	33.8	19.1	51.08

为了表示过拟合现象的程度, 本文定义了一个模型的过拟合距离, 在语音识别中, 过拟合往往是通过训练集和测试集上的识别率来进行判断的, 当数据在训练集上的识别率很高, 而在测试集上的识别率很低时, 那么, 就表示该模型有着严重的过拟合现象, 我们用模型在测试集上的评价指标和模型在训练集上的评价指标的差值的绝对值来表示过拟合现象的程度, 所以, 将它的计算公式定义为

$$\text{模型的过拟合距离} = |\text{模型在测试集上的评价指标} - \text{模型在测试集上的评价指标}| \tag{3}$$

从图 4 深色部分中可以发现, 在未采用 dropout 技术训练得到的 DNN-HMM 蒙古语声学模型中, 当隐含层网络层数由 4 层增加至 7 层时, 对词识别的过拟合距离从 21.17% 增长到了 54.81%; 对句识别的过拟合距离从 35.32% 增长到了 80.72%。由此可以看出, 随着隐含层网络层数的增加, 模型的过拟合距离越来越大, 过拟合距离的变大说明 DNN 网络构建的蒙古语声学模型已经严重过拟合, 那么, DNN-HMM 的表现就会越来越差。



(a) 对词识别的过拟合距离　　(b) 对句识别的过拟合距离

图 4　dropout 技术和隐含层层数对 DNN-HMM 模型过拟合距离的影响

Fig. 4　Influence of dropout technique and hidden layers on the over - fitting distance of DNN-HMM model

在图 4 中, 通过深浅两种颜色的对比可以看出, 采用 dropout 技术后, 当隐含层网络层数由 4 层增加至 7 层时, 对词识别的过拟合距离分别是 21.43%、21.91%、24.07% 和 25.48%。而未采用 dropout 技术, 对词识别的过拟合距离分别是 21.17%、21.91%、42.38%、54.81%。由此可知, 采用 dropout 技术后的过拟合距离要比未采用 dropout 技术后的过拟合距离小, 这一点, 在对句识别的过拟合距离上同样存在。所以, 在加入了 dropout 技术后, 有效地缓解了因隐含层数增加而导致的过拟合现象, 从而提高了模型的识别性能。

4　结束语

在蒙古语语音识别声学建模中, 本文给出了 DNN-HMM 蒙古语声学模型、无监督与监督算法相结合的蒙古语声学模型的训练算法以及以 GMM-HMM 为基础的 DNN-HMM 蒙古语声学模型的训练过程。在 Kaldi 实验平台上使用小规模的蒙古语语音语料库 IMUT310 开展了实验研究, 实验结果表明: 1) 在不同建模单元 (单音子和三音子) 下, DNN-HMM 蒙古语声学模型不论词错误率还是句错误率都优于 GMM-HMM 蒙古语声学模型, 具体表现为三音子 DNN-HMM 声学模型比三音子 GMM-HMM 模型在测试集上的词识别错误率降低了 7.5%, 句识别错误率降低了 13.63%; 2) 在训练 DNN-HMM 三音子蒙古语声学模型时, 加入 dropout 技术可以有效避免随着隐含层层数增加带来的过拟合影响。

参考文献:

[1] 马志强, 张泽广, 闫瑞, 等. 基于 N-Gram 模型的蒙古语文本语种识别算法的研究[J]. 中文信息学报, 2016, 30(1): 133–140.
MA Zhiqiang, ZHANG Zeguang, YAN Rui, et al. N-Gram based language identification for Mongolian text[J]. Journal of Chinese information processing, 2016, 30(1): 133–140.

[2] RABINER L R. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257–286.

[3] RABINER L, JUANG B H. Fundamentals of Speech Recognition[M]. Upper Saddle River, USA: Prentice-Hall, 1993.

[4] RENALS S, MORGAN N, BOURLARD H, et al. Connectionist probability estimators in HMM speech recognition[J]. IEEE transactions on speech and audio processing, 1994, 2(1): 161–174.

[5] LI Deng, HINTON G, KINGSBURY B. New types of deep neural network learning for speech recognition and related applications: an overview[C]//Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. Vancouver, Canada, 2013: 8599–8603.

[6] HINTON G, DENG Li, YU Dong, 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.

[7] YU Dong, DENG Li, DAHL G E. Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition[C]//Proceedings of NIPS Workshop on Deep Learning and Unsupervised Feature Learning. 2010.

[8] DAHL G E, YU Dong, DENG Li, et al. Large vocabulary continuous speech recognition with context-dependent DBN-HMMs[C]//Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing. Prague, Czech Republic, 2011: 4688–4691.

[9] DAHL G E, YU Dong, DENG Li, 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.

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

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

[12] BENGIO Y, LAMBLIN P, POPOVICI D, et al. Greedy layer-wise training of deep networks[M]//SCHÖLKOPF B, PLATT J, HOFFMAN T. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2007: 19–153.

[13] HINTON G E. A practical guide to training restricted Boltzmann machines[R]. Toronto: University of Toronto, 2010: 926–927.

[14] KHALTA B O, FUJⅡ A. A lemmatization method for Mongolian and its application to indexing for information retrieval[J]. Information processing & management, 2009, 45(4): 438–451.

[15] JAIMAI P, ZUNDUI T, CHAGNAA A, et al. PC-KIMMO-based description of Mongolian morphology[J]. International journal of information processing systems, 2005, 1(1): 41–48.

[16] GAO Guanglai, BILIGETU, NABUQING, et al. A Mongolian speech recognition system based on HMM[C]//Proceedings of 2006 International Conference on Intelligent Computing. Kunming, China, 2006: 667–676.

[17] 飞龙, 高光来, 闫学亮, 等. 基于分割识别的蒙古语语音关键词检测方法的研究[J]. 计算机科学, 2013, 40(9): 208–211.
FEI Long, GAO Guanglai, Yan Xueliang, et al. Research on Mongolian spoken term detection method based on segmentation recognition[J]. Computer science, 2013, 40(9): 208–211.

[18] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. arXiv: 1207.0580, 2012.

[19] SRIVASTAVA N. Improving neural networks with dropout[D]. Toronto: University of Toronto, 2013.

[20] DENG Li, YU Dong. Deep learning: methods and applications[J]. Foundations and trends in signal processing, 2014, 7(3/4): 197–387.

作者简介：

马志强，男，1972 年生，教授，主要研究方向为机器学习、语音识别、自然语言处理。发表学术论文 30 余篇，被 EI 检索 10 余篇。

李图雅，女，1993 年生，硕士研究生，主要研究方向为机器学习、语音识别、自然语言处理。

杨双涛，男，1990 年生，硕士研究生，主要研究方向为机器学习、语音识别、自然语言处理。