DOI:10.3969/j.issn.1671-0673.2013.05.011

一种基于 RBM 的深层神经网络音素识别方法

陈 琦,张文林,牛 铜,李弼程

(信息工程大学,河南 郑州 450001)

摘要:为提高连续语音识别中的音素识别准确率,采用深可信网络提取语音音素后验概率进行音素识别。首先利用受限玻尔兹曼机的学习原理,对深可信网络进行逐层的预训练;然后通过增加一个"软最大化(softmax)"输出层,得到用于音素状态后验概率检测的深层神经网络,并采用后向传播算法进行网络权值的精细调整;最后以后验概率为 HMM 发射概率,使用 Viterbi 解码器进行音素识别。针对 TIMIT 语料库的实验结果表明,该系统的音素识别率优于 GMM/HMM,MLP/HMM 和 TANDEM 系统性能。

关键词:受限玻尔兹曼机;深可信网络;神经网络;音素识别

中图分类号:TN912.34 **文献标识码**:A **文章编号**:1671-0673(2013)05-0569-06

RBM-Based Phoneme Recognition by Deep Neural Network Based on RBM

CHEN Qi, ZHANG Wen-lin, NIU Tong, LI Bi-cheng

(Information Engineering University, Zhengzhou 450001, China)

Abstract: To improve the performance of phoneme recognition in automatic speech recognition, a phoneme recognition method is built based on phoneme posteriors which are extracted by deep belief networks. Firstly, a deep belief network is pre-trained and layered as RBM greedily, and a deep neural network is created by adding a "softmax" output layer to the network. Subsequently, discriminative fine-tuning by back-propagation is done to adjust the weights and to make them better at predicting the probability distribution over the states of monophone hidden Markov models. Finally the sequence of the predicted probability distribution is fed into a standard Viterbi decoder. It is found that the method performs better on the TIMIT dataset than GMM/HMM, MLP/HMM and TANDEM methods.

Key words: restricted Boltzmann machine (RBM); deep belief networks; neural network; phoneme recognition

0 引言

隐马尔科夫模型(HMM)是自动语音识别任务中最常使用的建模方法。通常,每个 HMM 状态使用一个高斯混合模型(GMM)对单帧语音数据的声学特征进行建模。尽管 GMM/HMM 方法一直以来都是语音识别的主流方法,但是这种方法存在下列局限性[1]:GMM 假设数据的分布满足 Gaussian 分布,而实际的语音特征参数并不是 Gaussian 分布;由于 HMM 中假设单个状态的观测概率是统计独立的,因此对每个

收稿日期:2013-03-28;**修回日期**:2013-05-06

基金项目:国家自然科学基金资助项目(61175107)

作者简介:陈 琦(1974-),男,讲师,博士生,主要研究方向为语音信号处理。

HMM状态所对应的GMM进行训练时,使用的数据仅是所有的训练数据中同该状态所对齐的那一部分数据,即训练时没有充分考虑跨状态的上下文信息;对GMM参数进行训练有时需要使用特征降维,而这样做可能会导致某些有用信息的丢失。

为了克服上述缺陷,有学者提出使用人工神经网络(ANN)代替GMM进行语音识别。目前,最常用的ANN方法是使用多层感知器(MLP)对一组基于状态的后验概率进行估计。对后验概率的使用分为两类:在混合系统中[2],这些后验概率作为HMM的状态输出概率,送入解码器进行解码;在TANDEM系统[3],则是将其作为输入特征送入典型的GMM/HMM系统。通常,混合系统的性能要稍低于GMM/HMM系统,而TANDEM系统的性能则优于GMM/HMM[4]。

训练MLP时,网络的初始权值是随机设定的,并且由于其目标函数是非凸函数,使用随机梯度下降法进行训练时,有可能导致训练停滞于某个较差的局部最优点。当MLP仅包含1~3个隐含层,这个问题并不突出,但是当隐含层的层数超过3层时,该问题就显得十分突出[5]。近年来,有学者提出使用受限波尔曼兹机(restricted Boltzmann machine,RBM)对ANN的权值进行逐层无监督预训练[6-7]。该方法能够获得ANN初始权值的更优估计,使得构建含有多个隐含层的深层神经网络成为可能。

本文主要研究基于RBM的ANN训练方法,并在此基础上构建ANN/HMM音素识别器,实现连续语音的音素识别。

1 受限玻耳兹曼机

玻耳兹曼机(BM)可以看作一个由可见层和隐藏层组成的无向图模型,可见层由可见单元v构成,用于表示输入数据(观测值),隐藏层由隐藏单元h构成,h通过学习对输入数据的内在特征进行表示,v和h之间使用无方向的权值连接。当BM的可见层各单元之间以及隐藏层各单元之间没有内部连接时(如图1所示),就称为RBM。根据v和h所对应的建模单元类型的不同,RBM有不同的构成形式。当v和h都是随机的二值单元时,使用Bernoulli分布对二值变量进行建模,因此,称其为Bernoulli-Bernoulli RBM或二值RBM,这是最简单的一类RBM。下面分析二值RBM的学习原理,并将其扩展至输入为实值数据的情形。

对于二值RBM,定义如下的能量函数E:

$$E(v,h|\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}w_{ij}v_ih_j-\sum_{i=1}^{V}b_iv_i-\sum_{j=1}^{H}a_jh_j \tag{1}$$

其中,$\theta=(w,b,a)$,w_{ij}表示可见单元i和隐藏单元j之间的连接权值,且有$w_{ij}=w_{ji}$,b_i和a_j分别是可见单元和隐藏单元的偏移量,V和H分别表示可见单元和隐藏单元的数目。通过E可以定义可见单元和隐藏单元状态的联合分布概率,

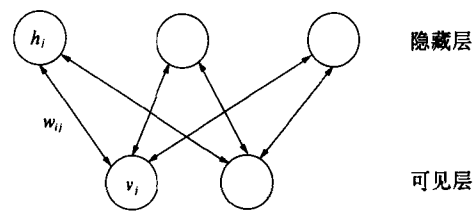

图1 RBM示意图

$$p(v,h|\theta)=\frac{1}{Z}e^{-E(v,h|\theta)},Z=\sum_{v'}\sum_{h'}e^{-E(v',h'|\theta)} \tag{2}$$

其中,Z称为配分函数(partition function)或归一化项,它的作用是保证$p(v,h|\theta)$是一个概率值。模型赋予可见矢量v的边缘分布概率等于对$p(v,h|\theta)$中所有隐藏矢量求和:

$$p(v|\theta)=\frac{1}{Z}\sum_h e^{-E(v,h|\theta)} \tag{3}$$

RBM采用梯度下降法进行极大似然学习[7]:

$$\frac{\partial\log p(v|\theta)}{\partial w_{ij}}=\langle v_ih_j\rangle_{\text{data}}-\langle v_ih_j\rangle_{\text{model}} \tag{4}$$

其中,$\langle v_ih_j\rangle_{\text{data}}$表示由输入数据所确定的期望,$\langle v_ih_j\rangle_{\text{model}}$表示对所有可能的$(v,h)$建立模型的期望。因此,可以得到RBM在训练集上执行对数概率随机梯度下降时的学习规则:

$$\Delta w_{ij}=\varepsilon(\langle v_ih_j\rangle_{\text{data}}-\langle v_ih_j\rangle_{\text{model}}) \tag{5}$$

其中,ε表示学习率(学习步长)。

由于隐藏单元之间没有连接,因此条件分布$p(h|v,\theta)=\prod_j p(h_j|v,\theta)$,给定一个随机选择的训练样本

v,隐藏单元 h_j 的状态为1的概率:

$$p(h_j=1|v,\theta)=\text{logistic}(a_j+\sum_{i=1}^{V}w_{ij}v_i) \tag{6}$$

其中,$\text{logistic}(x)=(1+e^{-x})^{-1}$。

由于可见单元之间也不存在连接,条件分布 $p(v|h,\theta)=\prod_i p(v_i|h,\theta)$,给定随机选择的隐藏矢量 h,重构可见单元 v_i 的状态为1的概率由(7)式给出,

$$p(v_i=1|h,\theta)=\text{logistic}(b_i+\sum_{j=1}^{H}w_{ij}h_j) \tag{7}$$

由(6)式和(7)式可以得到$\langle v_ih_j\rangle_{\text{data}}$的无偏采样。

$\langle v_ih_j\rangle_{\text{model}}$的计算需要对训练集合中所有样本的对数概率进行求导。具体的实现方法是从可见单元的任一随机状态开始,长时间地重复执行交替Gibbs采样(alternating Gibbs sampling)。一次交替Gibbs采样的过程是:使用(6)式并行计算得到所有隐含单元的二值状态并对其更新;隐含单元的二值状态更新后,使用(7)式获得的概率更新每个可见层单元 v_i 的状态。由此可知,$\langle v_ih_j\rangle_{\text{model}}$的计算量随着训练样本的增加以指数级增长。

为了解决RBM的训练速度问题,Hinton[8]于2002年提出了一种快速训练方法-对比散度法(contrastive divergence-CD)。该算法首先将可见单元状态设置为当前的训练样本值,然后执行多步交替Gibbs采样,获得"重构(reconstruction)"的可见单元状态$\langle v_i\rangle_{\text{reconstruction}}$;最后,再次利用(6)式更新隐含单元状态,得到$\langle h_j\rangle_{\text{reconstruction}}$。因此,权值更新量的(5)式变为

$$\Delta w_{ij}=\varepsilon(\langle v_ih_j\rangle_{\text{data}}-\langle v_ih_j\rangle_{\text{reconstruction}}) \tag{8}$$

类似的,可见单元偏置和隐藏单元偏置的更新量公式分别为

$$\Delta b_i=\varepsilon(\langle v_i\rangle_{\text{data}}-\langle v_i\rangle_{\text{reconstruction}}) \tag{9}$$

$$\Delta a_j=\varepsilon(\langle h_j\rangle_{\text{data}}-\langle h_j\rangle_{\text{reconstruction}}) \tag{10}$$

CD算法虽然是对极大似然梯度的一种近似,但是它非常有效。如果在获取$\langle v_ih_j\rangle_{\text{reconstruction}}$的状态之前,执行更多步的交替Gibbs采样,则RBM可以学习到更好的生成模型。但是,本文使用RBM的主要目的是利用它进行深层神经网络初始权值的预训练,对于特征检测器的预训练目标(网络初始权值位于一个合适的起点)而言,多步交替Gibbs采样起到的作用很小[6]。因此,为了减少训练时间,本文的所有实验结果都是使用单步交替Gibbs采样(CD_1)得到的。CD_1是指:在初始更新隐含单元之后,执行一遍完整的交替Gibbs采样。为了抑制学习过程中的噪声,在重构时,通常使用实数概率值代替二值采样值,并将隐含单元的状态也设置为实数概率值;但是,在第1次计算隐含状态时,必须使用采样后的二值数据,因为此时的采样噪声起到了一个正则项的作用,可以有效防止训练中的过拟合现象。

以上对二值RBM的学习过程进行了描述,在音素识别应用中,神经网络的输入数据是类似于MFCC或PLP的底层声学特征,它们都是实数值的数据,使用二值分布对其进行建模是不合适的。为使RBM能够对底层声学特征进行学习,将RBM可见单元建模为具有Gaussian噪声的线性变量,而隐含层仍然由二值单元构成,这种类型的RBM称为Gaussian-Bernoulli RBM(GRBM),其能量函数定义为

$$E(v,h|\theta)=-\sum_{i=1}^{V}\sum_{j=1}^{H}\frac{v_i}{\sigma_i}w_{ij}h_j+\sum_{i=1}^{V}\frac{(v_i-b_i)^2}{2\sigma_i^2}-\sum_{j=1}^{H}a_jh_j \tag{11}$$

其中,σ_i 表示可见单元 i 的Gaussian噪声的标准差,相应地,用于CD_1学习的两个条件分布公式则分别修正为

$$p(h_j=1|v,\theta)=\text{logistic}\left(a_j+\sum_{i=1}^{V}w_{ij}\frac{v_i}{\sigma_i}\right) \tag{12}$$

$$p(v_i=1|h,\theta)=N(b_i+\sigma_i\sum_{j=1}^{H}w_{ij}h_j,\sigma_i^2) \tag{13}$$

其中,$N(\mu,\sigma^2)$表示均值为μ,方差为σ^2的Gaussian分布。

在使用CD_1进行GRBM训练时,需要在整个训练集上对输入数据进行归一化,使得GRBM输入数据的每一维都满足均值为0且方差为1的正态分布;在计算 $p(v_i=1|h,\theta)$ 时令 $\sigma=1$;在重构可见层数据时不使用二值数据,而使用概率值。采用这种处理方法可以避免在计算中对噪声水平进行估计。

2 基于RBM的深层神经网络音素识别

利用上述方法,可以逐层进行多个RBM的学习。当一个RBM的权值通过CD_1学习得到后,使用隐藏单元的状态值,作为训练下一个RBM的输入数据,继续对下一个RBM进行训练。根据需要,这一过程可以一直重复下去,这样就能够学习到任意多个RBM。将多个RBM自底向上依次堆积搭建起来,可以得到一个多层生成模型—深可信网络(deep belief network,DBN)。下面给出使用多个RBM构造用于音素识别的深层神经网络的详细步骤[7](见图2)。

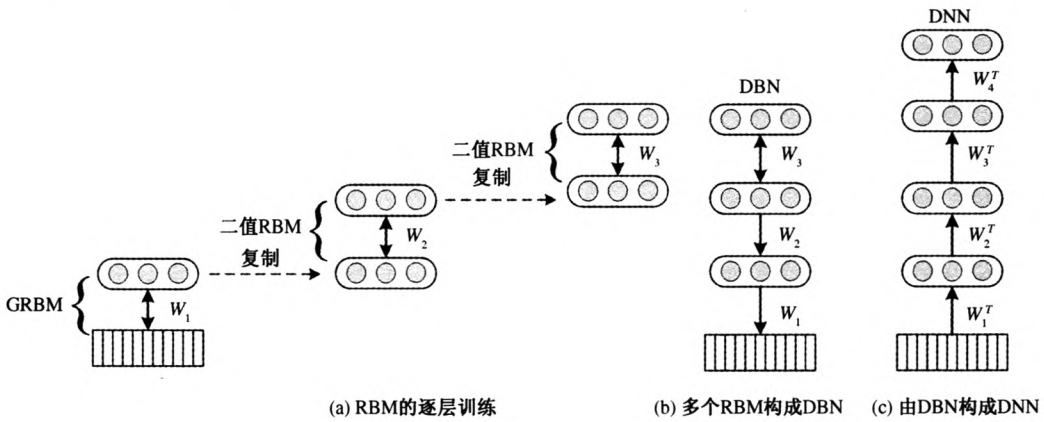

图2 利用RBM构造用于音素识别的深层神经网络

步骤1 RBM的逐层训练

①为了充分利用语音的上下文信息,将当前帧前后的多帧MFCC串接后作为GRBM的输入,使用CD_1算法训练GRBM,得到第1层GRBM的权值W_1;使用GRBM的输入和W_1,得到GRBM的隐藏单元的状态H_1。

②使用H_1作为下一个RBM的输入数据,使用CD_1算法进行训练,得到该层RBM的权值W_2;使用H_1和W_2,得到该层RBM的隐藏单元状态H_2。

③重复执行②,直至达到所需要的层数。

步骤2 利用多个RBM搭建DBN

按照步骤1中RBM训练的先后次序,自底向上将多个RBM依次通过权值连接起来,构成DBN。在这个DBN中,最上面两层(最后训练得到的RBM)之间的连接是无方向的,DBN的其它层是自顶向下的有向连接。(使用该模型生成数据时,顶层RBM执行多步Gibbs交替采样,然后自顶向下通过其余各层,得到生成的数据。)

步骤3 利用DBN构造用于语音识别的深层神经网络

获得DBN之后,在其顶层之上,再增加一个"软最大化(softmax)"输出层,输出层的每个节点对应HMM中的一个状态。此时的网络称为经过预训练的深层神经网络(deep neural network,DNN)。

步骤4 DNN权值的精调整

通过RBM逐层预训练得到的DNN还需要进行区分性训练(精调整),才能够针对输入的声学特征生成音素状态的后验概率p(HMM状态|声学特征)。与预训练的无监督学习不同,精调整是有监督的学习。针对多帧的输入声学特征,取其中间帧所对应的HMM状态作为DNN的学习目标,在预训练获得的网络权值的基础上,采用某种训练准则,通过反向传播算法获得神经网络的最终权值。

经过精调整之后,DNN输出形式为p(HMM状态|声学特征)的概率。为了在HMM框架下进行Viterbi解码,还需要将DNN的输出转换为似然比p(声学特征|HMM状态)。根据贝叶斯公式可知:

$$p(\text{声学特征}\mid HMM\text{状态})=\frac{p(HMM\text{状态}\mid\text{声学特征})}{p(HMM\text{状态})}p(\text{声学特征}) \tag{14}$$

其中,p(HMM状态)可以通过训练集中的HMM状态进行强制对齐后统计获得或者简单的认为其是一个

固定值,p(声学特征)是未知的,但是它对于所有状态都是相同的,因此不会影响后续 Viterbi 解码的效果。

将经过(14)式计算得到的状态似然比送入 Viterbi 解码器[1]进行解码,即可以得到音素序列。

3　实验配置及结果分析

3.1　实验配置

本文实验在 TIMIT 语料库上进行,排除 SA1 和 SA2 中的语句之后,选择 462 个说话人的 3296 个语句作为训练集,选择 TIMIT 的核心测试集(24 个说话人的 192 个语句)作为测试集,在核心测试集之外再选择 50 个说话人的 400 个语句作为开发集。语音信号使用 Hamming 窗处理,帧长 25ms,帧移 10ms,预加重的系数为 0.97。声学特征参数使用 MFCC,包括语音对数能量、12 维 MFCC 参数及其一阶、二阶差分系数,共计 39 维特征参数。

MFCC 作为底层 RBM 的输入数据,在送入 RBM 训练前,需在整个训练集范围内对其进行归一化,使得每一维特征参数都满足均值为 0、方差为 1 的正态分布,开发集和测试集使用训练集的归一化参数进行归一化。TIMIT 语料库使用 61 个音素,每个音素对应 3 个状态,因此,DNN 的训练目标设置为 183 个音素状态类别。由于 TIMIT 库对语料的标注达到了音素级,最初的训练目标是将每个音素的持续时间等分为 3 部分,每部分对应 1 个状态;进行第 1 遍精调整并且解码之后,使用强制对齐,获得更新的训练目标,然后再次进行精调整以获得最终的 DNN 权值。

神经网络训练的参数设置[6]:

RBM 预训练　使用小批量(minibatch)的梯度下降算法,每个批量的规模为 128 个训练样本。对于 GRBM,学习率为 0.001,学习轮次(epoch)为 225;对于二值 RBM,学习率为 0.1,学习轮次为 75。冲量值(momentum)在最初 5 轮设为 0.5,然后增加至 0.9;权值衰减因子为 0.0002。

DNN 精调整　采用互熵误差准则,使用小批量的梯度下降算法,每个批量的规模为 512 个训练样本。冲量值在最初 5 轮设为 0.5,然后增加至 0.9;权值衰减因子为 0,学习率最初设为 0.008,在每一轮训练完成后,如果开发集的分类准确率下降,则将网络权值重置为当前轮次的初始值,并将学习率减半,继续训练。当学习率减半后,如果开发集的分类准确率仍没有提高,则停止训练。

对音素状态后验概率解码时不使用语言模型,获得音素序列后,将音素集由 61 个映射到 39 个[9],使用音素错误率对识别结果进行评价。

计算机配置:32G 内存,12 核 Xeon 3.07-GHz 处理器,NIVIDA Quadro 600 GPU。

3.2　实验结果及分析

DNN 的训练较为耗时,为减少训练时间,使用 GPU 进行预训练和精调整。对 8 个隐含层,每层节点数为 1024 的 DNN,输入帧数为 21,顶层 RBM 进行预训练时,对整个训练集进行一轮学习的时间约为 740s,进行 DNN 精调整时,一轮学习的时间约为 960s。

图 3 给出了隐含层节点个数为 1024 时,随着隐含层数的改变,不同输入帧数时的识别结果。从图中可以看出,输入帧数固定时,增加隐含层数能够提高识别性能,但是随着隐含层数的增加,识别性能提高的幅度有所下降。当 DNN 的隐含层个数超过 3 层时,增加输入帧数能够提高识别性能,当隐含层个数小于 3 时,输入帧数的增加并不能显著提高识别性能。出现这种情况的原因是,多帧输入可以使 DNN 有效利用语音的上下文特征,但是,当隐含层数较少时,DNN 对于这些特征的区分能力有限。当输入帧数为 21 时,隐含层数较少时,识别性能不及输入帧数为 11 及 15 的情形,这说明对于语音识别应用,过长的上下文信息对识别性能的贡献有限,当隐含节点数为 1024 时,合适的语音输入时长范围在 110ms ~ 210ms。

图 4 给出了输入帧数固定为 11 帧,隐含层节点数变化时,DNN 的识别结果。从图中可以看出,当隐含层数固定时,增加隐含层节点个数,可以提高音素识别性能。

表 1 给出了本文方法(输入帧数为 15,4 个隐含层,每层 1024 节点)同其它音素识别算法的比较。从表 1 中可以看出,本文的方法优于其中的 3 种音素识别方法,说明本

表 1　不同音素识别方法比较

方法	音素错误率
GMM/HMM[10]	33%
MLP/HMM[9]	30.49%
TANDEM[11]	28.4%
本文方法	22.8%

算法的有效性。

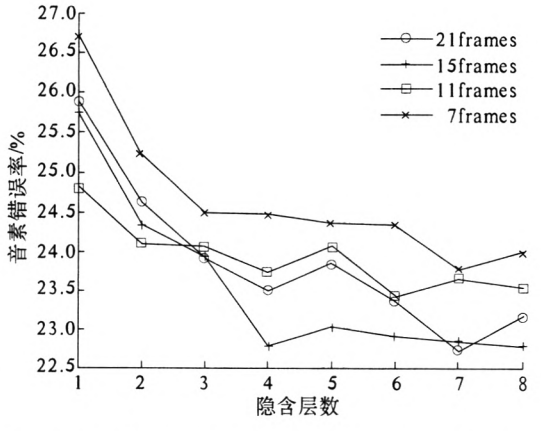

图3 输入帧数变化时的音素识别性能(隐含层节点数为1024)

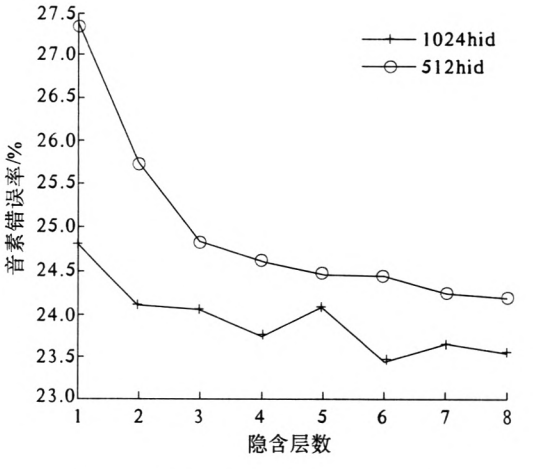

图4 隐含层节点数不同时的音素识别性能

4 结束语

本文主要研究分析了RBM的学习原理,利用RBM的逐层无监督学习构建深可信网络,实现了利用深层神经网络进行音素识别的方法。实验表明,深层神经网络能够充分利用语音底层声学特征的上下文信息,有效降低音素识别错误率。本文为进一步研究深度学习理论在语音识别中的应用奠定了基础。

参考文献:

[1] Mohamed A, Sainath T N, Dahl G, et al. Deep Belief Networks Using Discriminative Features for Phone Recognition[C]// IEEE International Conference on Acoustic Speech and Signal Processing. 2011:5060-5063.

[2] Bourlard H, Morgan N. Connectionist Speech Recognition: A Hybrid Approach[M]. Norwell, MA: Kluwer, 1993.

[3] Ellis D P W, Singh R, Sivadas S. Tandem Acoustic Modeling in Large-Vocabulary Recognition[C]//IEEE International Conference on Acoustic Speech and Signal Processing. 2001:517-520.

[4] Sainath T N, Kingsbury B, Ramabhadran B, et al. Making Deep Belief Network Effective For Large Vocabulary Continuous Speech Recognition[C]//IEEE Automatic Speech Recognition and Understanding Workshop. 2011:30-35.

[5] Dahl G, Yu D, Deng L, et al. Context-Dependent Pre-Trained Deep Neural Networks for Large Vocabulary Speech Recognition[J]. IEEE Trans. Audio, Speech, Lang. Process, 2012, 20(1):30-42.

[6] Hinton G E. A Practical Guide to Training Restricted Boltzmann Machines[EB/OL]. [2013-04-28]. http://www.cs.toronto.edu/~hinton/absps/guideTR.pdf.

[7] Bengio Y, Lamblin P, Popovici D, et al. Greedy layer-wise training of deep networks[C]// Advances in Neural Information Processing Systems. 2007,19:153-160.

[8] Hinton G E. Training products of experts by minimizing contrastive divergence[J]. Neural Computation, 2002, 14(8): 1711-1800.

[9] Lee K F, Hon H W. Speaker-independent phone recognition using hidden Markov models[J]. IEEE Trans. Acoustic, Speech, Signal Process, 1989, 37(11):1641-1648.

[10] Sha F, Saul L. Large margin Gaussian mixture modeling for phonetic classification and recognition[C]// IEEE International Conference on Acoustic Speech and Signal Processing. 2006:265-268.

[11] Joel P, Sivaram G, Mathew M D, et al. Analysis of MLP-Based Hierarchical Phoneme Posterior Probability Estimator[J]. IEEE Trans. Audio, Speech, Lang. Process., 2011, 19(2):225-241.

一种基于RBM的深层神经网络音素识别方法

作者： 陈琦， 张文林， 牛铜， 李弼程， CHEN Qi， ZHANG Wen-lin， NIU Tong， LI Bi-cheng
作者单位： 信息工程大学,河南郑州,450001
刊名： 信息工程大学学报
英文刊名： Journal of Information Engineering University
年，卷(期)： 2013,14(5)