多权值神经元网络仿生模式识别方法在低训练样本数量非特定人语音识别中与 HMM 及 DTW 的比较研究

覃 鸿, 王守觉
(中国科学院半导体研究所神经网络实验室, 北京 100083)

摘 要: 本文将基于多权值神经元网络的仿生模式识别方法用于连续语音有限词汇量固定词组识别的研究中, 并将其识别效果与 HMM 方法及 DTW 方法进行了比较分析. 以 15 个词组的词汇表做测试, 通过调整这三种识别算法的参数, 在它们的拒识率相同的情况下, 针对参加训练的词汇, 比较他们的错误识别率(某类误认为他类); 针对未参加训练的词汇, 比较他们的错误接受率(误认为某类). 结果表明, 在低训练样本数量的情况下, 仿生模式识别方法能获得更好的识别效果.

关键词: 仿生模式识别; 多权值矢量神经元; 语音识别; HMMs; DTW
中图分类号: TN912 文献标识码: A 文章编号: 0372-2112 (2005) 05-0957-04

Comparison of Biomimetic Pattern Recognition, HMM and DTW for Speaker-Independent Speech Recognition

QIN Hong, WANG Shou-jue
(*Lab of Artificial Neural Networks, Institute of Semiconductors, CAS, Beijing 100083, China*)

Abstract: The purpose of this paper is to compare the performance of three speech recognition methods, one based on Biomimetic Pattern Recognition (BPR) and the other two based on Hidden Markov Models (HMMs) and Dynamic Time Warping (DTW) respectively. As a general purpose model of pattern Recognition, BPR is realized by Multi-Weights Neuron Networks. For the 15 words vocabulary, we analyze the false recognition rate (ratio of accepting a trained word to another trained word) and false acceptance rate (ratio of accepting an untrained word to a trained word) respectively. Experiment results show that when the training data was not sufficient, the manner of BPR achieved a higher performance.

Key words: biomimetic pattern recognition; multi-weights neuron; speech recognition; hidden markov models; dynamic time warping

1 引言

语言是人类最自然的交流方式, 因而语音识别技术用途广泛, 潜力巨大. 然而, 由于语音信号的随机性、多变性和不稳定性, 语音识别技术跟其他识别技术相比发展缓慢. 近年, 语音信号特征参数提取技术的研究成果(LPCC 参数、MFCC 参数等) 促进了语音识别研究的快速发展, 出现了动态时间规整算法[1](DTW)、隐马尔可夫模型[2](HMM) 和人工神经网络[3](ANN) 等识别算法, 其中 HMM 在语音识别领域的成功应用[4], 使其成为语音识别的主要方法. HMM 方法关注的是模式的统计特性, 目的是将模型分成几类, 然后选择出与感知数据最接近的模型类别[5]. 它需要大量的训练样本才能获得满意的训练效果, 这往往很难做到. 实际应用中, 要识别的可能是一些词汇量较少, 并有可能变动的词汇集, 例如餐厅的语音点菜系统, 它要求新菜名的加入不影响已有词汇, 并且新菜名的训练过程中训练样本容易获得, 如果训练需要大量样本, 这就给系统的更新带来困难. DTW 方法是一种基于模板匹配的语音识别方法, 它应用动态规划方法成功解决了语音信号的参数序列在进行序列比较时时长不等的难题, 并在特定人语音识别和孤立词语音识别中得到了广泛应用. 仿生模式识别[6]自从被提出来以后, 已用于实物目标识别、人脸确认[7]与人脸识别[8]等的研究, 并在与传统模式识别的比较中, 显示出其在识别效果上的优越性. 本文利用仿生模式识别基本原理研究一种在低训练样本数量的情况下有限词汇量固定词组识别的新方法.

2 仿生模式识别简介

传统的模式识别或模式分类, 是模式识别长期发展过程中建立起来的经典方法, 它从"划分"的概念出发, 认为所有可用的信息都包含在训练集中, 同类事物之间的关系无任何先

收稿日期: 2004-03-22; 修回日期: 2004-12-13
基金项目: 国家自然科学基金(No. 60135010)

验知识存在. 其目标是如何对若干有限类样本在特征空间中进行最优分类划分, 或一类样本与有限类已知样本的区分.

仿生模式识别是一种不同于传统模式识别的理论方法, 它从"认识"的概念出发, 认为两个同类事物之间至少存在一个渐变过程, 在渐变过程中间的各事物都是同属于该类的, 或者说特征空间中同类样本的全体是连续的. 因而其目标是如何对特征空间中的同一类样本作最佳覆盖. 关于仿生模式识别的详细叙述请参考文献[6].

3 基于仿生模式识别的语音识别

仿生模式识别以"同类样本的全体在特征空间中是连续的"作为样本点分布的先验知识, 即: 设特征空间 R^n 中所有属于 A 类事物的全体为集合 A, 若集合 A 中存在任意两个元素 x 与 y, 则对 ε 为任意大于零的值时, 必定存在集合 B, 使得:

$$B=\{x_1, x_2, x_3, \cdots, x_n \mid x_1=x, x_n=y, n\subset N,$$
$$\rho(x_m, x_{m+1})<\varepsilon, \varepsilon>0, n-1\geqslant m\geqslant 1, m\subset N\}, B\subset A$$

对应于学习过程, 就是针对同类事物的训练样本在特征空间中的分布, 选择一个或多个合适的封闭曲面, 形成一个高维空间的连续的复杂几何形体[9]来合理覆盖训练样本.

多权值矢量神经元[10]在一定的计算函数下, 能形成一个高维空间的封闭曲面.

一个多权值矢量神经元的通用表达式为: $Y=f[\Phi(W_1, W_2, \cdots, W_m, X)-\theta]$, 式中:

$W_1, W_2, \cdots, W_m$ 为 m 个权值矢量;

X 是输入矢量;

Φ 为由多权值矢量神经元决定的计算函数(多个矢量输入, 一个标量输出);

θ 为多权值神经元的激活阈值;

f 为非线性转移函数.

设特征空间是 n 维实数空间 R^n, 即 $X\in R^n$, 矢量函数方程:

$$\Phi(W_1, W_2, \cdots, W_m, X)=\theta \qquad (1)$$

可视为由 $W_1, W_2, \cdots, W_m$ 等 m 个权值矢量所决定的在特征空间 R^n 中 X 矢量的一种轨迹, 此轨迹为 R^n 空间中的 $(n-1)$ 维超曲面(或超平面), 它把 R^n 分成两个部分. 如果使公式(1)是一个封闭的超曲面, 则就在特征空间中形成了一个有限覆盖区域[11]. 改变神经元权值, 将得到具有不同形状的超曲面神经元. 对应于多权值神经元网络的学习过程, 就是针对不同类事物的训练样本在特征空间中的分布, 选择合适的多个多权值神经元, 他们的并在特征空间中表示为一个连通区域, 即对于某类 A, 则 $Y_A=\min(Y_{Aj})_{j=1}^{P}$, 其中 P 是覆盖 A 类全部训练样本所需的神经元的个数, 通过网络训练决定, $Y_{Aj}=f[\Phi(W_{Aj1}, W_{Aj2}, \cdots W_{Ajm}, X]-\theta_{Aj}]$ 是这些神经元的输出.

对应于识别过程, 则是判断被识别样本是否落在代表某类事物的多个神经元的超曲面围成的高维有限空间的并集中. 设被识别样本为 X, 则识别时的判别函数为:

$$R=F[\min(Y_A, Y_B, \cdots\cdots]$$

4 基于 HMM 的语音识别

20 世纪 80 年代初人们开始用 HMM(隐含马尔科夫模型)来描述语音信号后, 该模型不断得到发展, 成为目前语音识别的主流技术[12]. HMM 的算法是将语音看成是一连串特定状态, 这种状态不能被直接观测到, 而是以某种隐含的关系与语音的观测值(语音特征)相关联. 这种隐含关系在 HMM 中通常以概率形式表现出来, 模型的输出结果也以概率的形式给出.

HMM 由马尔科夫过程和概率输出函数集组成. 马尔科夫过程由状态及状态间的转移概率组成. 状态用 $S_l(l=1, 2, \cdots N)$ 表示, N 为状态的个数. 在时刻 n, 模型所处的状态用 x_n 表示, 并且有: $x_n\in\{S_1\sim S_N\}$, $\forall n$. 状态间的转移概率由矩阵 $A_{N\times N}$ 决定, 其中 a_{ij} 表示模型 n 时刻在 S_i 状态下, 则下一时刻 $(n+1)$ 转移到 S_j 状态的概率. 模型在初始化时, 有一个初始状态概率矢量 $\alpha=[\alpha_1, \alpha_2, \cdots, \alpha_N]$, 其中 α_i 表示初始时刻, 系统处于 S_i 状态的概率. 对应于每一个状态都有一个相关的概率输出函数, 用于估计观测值在该状态下的输出概率. N 个状态的概率输出函数集构成一个 N 维概率分布函数矢量 B, 具有离散分布的概率分布函数矢量构成一个矩阵, 而具有连续分布的概率密度函数矢量的每个元素为一个概率密度函数. 因此, 一个 HMM 的特性由其三个特征参数 α、A、B 来决定.

对应于模型的训练过程, 就是根据已知的训练样本集, 确定模型的三个特征参数 α、A、B, 使得该参数下的模型产生训练样本集中的每个学习样本的概率平均值达到最大. 对应于识别过程, 就是已知模型的输出 Y 以及模型的三个特征参数 α、A、B, 估计模型产生 Y 时最可能经历的状态序列 X, 或称为最优状态序列搜索.

5 基于 DTW(Dynamic Time Warping)的语音识别

把待识别的语音模式称为测试模板 T, $T=T(a_1, a_2, \cdots a_i, \cdots a_I)$, 把标准语音模式称为参考模板 R, $R=R(b_1, b_2, \cdots b_j, \cdots b_J)$, 识别时, 计算 T 与 R 之间的失真度. 失真度可用两者之间的欧几里得距离衡量, 在属于相同类别时, 距离小, 失真度小; 在属于不同类别时, 距离大, 失真度就大. 然而, 即使是同一个人说同一句话, 说话人每次发声的持续时间都会不一样, 这时, $I\neq J$, 或者说时间序列模型在整个时间段内发生了非线性时间伸缩. DTW 方法为了克服语速的差异, 用动态规划的方法, 将两个模式特征序列进行匹配, 即将 T 和 R 的帧号分别在坐标系的横轴和纵轴上标出, 坐标上某点 $D(i, j)$ 表示 T 的第 i 帧与 R 的第 j 帧的失真度. DTW 方法的目标就是要找出一条从起点 $(1, 1)$ 到终点 (I, J) 的最佳路径, 使得在路径失真度总和最小, 即总失真达到最小. 另外, 基于模板匹配的识别方法, 识别的性能与特征模板的可靠性有很大关系. 对于非特定人的语音识别, 需要先对训练样本进行聚类[13,14].

6 非特定人连续语音点菜系统的实验过程与结果

6.1 语音特征参数的提取方法

特征提取是语音识别的一个关键步骤, 本文实验用

MFCC(Mel 倒谱系数) 作为语音的特征参数.

人耳对不同频率的语音具有不同的感知能力, 人的听觉系统是一个特殊的非线性系统, 它响应不同频率信号的灵敏度是不同的. 为模拟人耳对不同频率语音的感知特性, 人们提出了Mel 频率的概念, MFCC 即为基于Mel 频率的概念而提出的, 由于它反映了人耳的听觉特征, 因而具有很好的性能及其鲁棒性. 实际提取过程中, 还需对语音信号进行预处理. 对语音的高频部分进行预加重来增加语音的高频分辨率; 根据语音具有短时平稳的特点, 对语音进行分帧操作以提取其短时特性. 在本文的实验中, 语音信号采用 22kHz, 16 位精度进行采样, 原始的一个语音信号经过端点检测、预加重后, 分为 32 帧, 每帧 16 个MFCC 参数, 映射成 512 维特征空间的一个样本点.

6.2 训练样本集与测试样本集

词汇表由 15 个汉语词组组成.

训练集 A 的发音人为来自不同地区的3 名男性和 3 名女性, 每人每个词组各发音 3 遍, 每个词组 18 个样本, 共 270 个样本用于训练. 训练集 B 将发音人数增加到 5 名男性和 5 名女性, 即每个词组 30 个样本, 共 450 个样本用于训练.

测试集由不是训练集发音人的其他多人的发音共 697 个样本构成, 其中 539 个样本属于词汇表中的词组, 构成测试集 A, 应能正确认识, 用以测试识别系统在一定拒识率下的错误识别率; 其余的 158 个样本为词汇表以外的词组, 构成测试集 B, 应能正确拒识, 用以测试它的错误识别率.

6.3 实验过程及结果

HMM 模型采用状态数 $N=5$ 的无跳转从左到右模型对每个词进行统计建模. 高斯概率密度函数混合数 M 的多少对识别有很大的影响, 混合数的个数较多, 系统的性能就更好, 但意味着需要训练的参数量增加, 因而需要更多的训练数据. 本文经实验比较, 在选择 $M=6$ 时系统获得最优的性能.

多权值神经元网络的输入层包括 512 神经元, 对应一个词组的 512 个特征值; 隐层由三权值神经元构成, 其神经元的数量和权值由训练过程决定; 输出层包括 15 个神经元, 每个神经元对应于一个要识别的语音. 整个网络的输出矢量为 $O=[O_1, O_2, \cdots, O_{15}]$ 若当前输入矢量属于第 i 类, $O_i=1$, 否则 $O_i=0$. 若 $O=[O_1, O_2, \cdots, O_{15}]=0$, 则认为当前输入矢量不属于已训练的任何类.

DTW 方法的模板通过 K 均值算法对训练样本进行聚类获得, 每个词组对应一个特征模板. 本文实验中 HMM 模型的训练及 DTW 的聚类用文献[15] 的算法及部分程序. 对于测试集 A, 为比较三种方法的效果, 实验中调整各自的判别函数阈值, 使这三种方法在拒识率相同的情况下, 比较他们误识率. 对于测试集 B, 则比较他们对未训练过的词汇的错误识别率. 识别结果如表 1 所示.

表 1 BPR、HMM、DTW 三种方法在不同规模训练集下的识别结果

识别方法	训练集 A(270 个样本)			训练集 B(450 个样本)		
	测试集 A		测试集 B	测试集 A		测试集 B
	误识率	拒识率	误识率	误识率	拒识率	误识率
BPR (仿生模式识别)	3.40%	5%	13.92%	1.48%	5%	13.29%
HMM	13.36%	5%	95.57%	2.60%	5%	89.24%
DTW	6.68%	5%	93.67%	4.3%	5%	78.48%

7 结果讨论与结论

HMM 在本质上是一种统计的方法, 因此需要大量的训练样本才能得到事物的统计特性, 参加训练的样本越多, HMM 的性能越好. 以上实验结果表明, 在低训练样本数量的情况下, 对于已训练词组的被识别样本, HMM 方法的误识率较高, 随着训练样本数量增加, 其误识率下降. 然而, 事实上, 大量的语音数据往往不容易获得, 面临的问题更多是在低训练样本数量时, 如何获得较高的识别率. DTW 方法在训练样本数量较少的情况下, 能获得教高的识别率, 但不论是 DTW 方法或是 HMM 方法, 它们对于词汇表以外的词误识率都很高.

仿生模式识别方法在训练样本数量较少的情况下, 就能获得很好的识别效果, 而且对于未训练过的词组有较高的正确拒识率. 这正是由于它是基于“认识”事物而不是基于“区分”事物为目的, 它更接近于人类“认识”事物的特性, 因而显示出其优越的效果.

参考文献:

[1] LR Rabiner, AE Rosenberg, SE Levinson. Considerations in dynamic time warping algorithms for discrete word recognition[J]. IEEE Transactions on Acoustics, Speech and Signal Processing, December, 1978, 26(6): 575- 582.

[2] Rabiner L, Juang B. An Introduction to Hidden Markov Models[M]. IEEE Acoustics, Speech & Signal Processing Magazine, JANUARY 1986. 4- 16.

[3] Gemello R, Albesano D, Mana F, Moisa. L Multi-source neural networks for speech recognition: a review of recent results[A]. IEEE-INNS-ENNS International Joint Conference on Neural Networks[C]. Italy: 2000. 265- 270.

[4] Lee KF. Automatic Speech Recognition: The Development of the SPHINX SYSTEM[M]. Kluwer Academic Publishers, Boston, 1989.

[5] Duda R O. 模式分类(原书第 2 版)[M]. 李宏东, 等, 译. 北京: 机械工业出版社, 2003. 9.

[6] 王守觉. 仿生模式识别(拓扑模式识别) ——一种模式识别新模型的理论与应用[J]. 电子学报, 2002, 30(10): 1417- 1420.
WANG Shou jue. Bionic (Topological) pattern recognition—a new model of pattern recognition theory and its applications[J]. Acta Electronica Sinica, 2002, 30(10): 1417- 1420.

[7] 王守觉, 等. 基于仿生模式识别的多镜头人脸身份确认系统研究[J]. 电子学报, 2003, 31(1): 1- 3.
WANG Shou jue, XU jian, WANG Xiarr bao, QIN Hong. Multi-camera humarr face personal identification system based on the biomimetic pattern recognition[J]. Acta Electronica Sinica, 2003, 31(1): 1- 3.

[8] 王守觉, 等. 基于仿生模式识别与传统模式识别的人脸识别效果比较研究[J]. 电子学报, 2004, 32(7): 1057- 1061.
WANG Shou jue, QU Yarr feng, LI Wei-jun, QIN Hong. Face recognition: biomimetic pattern recognition vs. traditional pattern recognition

[J]. Acta Electronica sinica, 2004, 32(7): 1057- 1061.
[9] 模式识别中的非超球面几何形体覆盖方法[P]. 中国专利申请文件: 申请号 02124837.0
[10] 多权值突触的神经元构造方法[P]. 中国专利申请文件: 申请号 02122638.5.
[11] 王守觉, 等. 人工神经网络的多维空间几何分析及其理论[J]. 电子学报, 2002, 30(1): 1- 4.
WANG Shou-jue, WANG Bai-nan. Analysis and theory of high-dimension space geometry for artificial neural networks[J]. Acta Electronica Sinica, 2002, 30(1): 1- 4.
[12] L R Rabiner, B H Juang. Fundamental of Speech Recognition[M]. Englewood cliffs, NJ: Prentice-Hall, 1993.
[13] LR Rabiner, et al. Speaker-independent recognition of isolated words using clustering techniques[J]. IEEE Trans. Acoustics, Speech, and Signal Processing, 1979, ASSP-27(4): 336- 349.
[14] 蔡莲红, 等. 现代语音技术基础与应用[M]. 北京: 清华大学出版社, 2003. 243- 244.
[15] 何强, 何英. matlab 扩展编程[M]. 北京: 清华大学出版社, 2002.

作者简介:

覃 鸿 女, 中国科学院半导体研究所神经网络实验室在职研究生, 从事神经网络软、硬件研究工作. E-mail: qinh@red.semi.ac.cn.

王守觉 男, 历任中国科学院半导体研究所室主任、副所长、所长等职, 1980 年当选中国科学院院士. 现为半导体所神经网络实验室负责人.

电 子 学 报 2005 年第 5 期 Acta Electronica Sinica No. 5 2005

(总期 259 期) (Monthly) (Series No. 259)

主管单位	中国科学技术协会	China Association for Science and Technology
主办单位	中国电子学会	Published by the Chinese Institute of Electronics, Beijing
协办单位	南京才华科技有限公司	Nanjing Caihua Technology Co., Ltd.
编 辑	《电子学报》编辑委员会	Edited by Editorial Board of Acta Electronica Sinica
主 编	王 守 觉	Chief Editor: WANG Shou-jue
总 编 辑	刘 力	Director: LIU Li
通 信 处	北京 165 信箱 (邮政编码 100036)	Add: Editorial Office of Acta Electronica Sinica (POBox 165, Beijing 100036, China)
电 话	(010) 68279116, 68285082	Tel: 86-10-68279116, 68285082
传 真	(010) 68173796	Fax: 86-10-68173796

Home page: http://www.elecjournal.org; http://dzxu.chinajournal.net.cn

Email: cje@elecjournal.org; dzxu@chinajournal.net.cn

排版印刷	北京育兴达印刷厂	Printed by Yuxingda, Beijing, China
国内总发行	北京市报刊发行局	Distributed by Domestic: Beijing Baokan Faxingju, China
国外总发行	中国国际图书贸易总公司	Foreign: China International Book Trading Corporation
国内订购处	全国各邮电局	Subscription Office——All Local Post Offices in China

国内统一刊号: CN11- 2087/TN 邮发代号(国内/国外): 2- 891/M436 国内定价 ¥32.00