

一种基于高维空间覆盖动态搜索方法的 非特定人连续数字语音识别的研究

王守觉^{1,2}, 潘晓霞², 徐春燕², 陈旭¹, 安冬¹, 曹文明²

(1. 中国科学院半导体研究所, 北京 100083; 2. 浙江工业大学智能信息系统研究所, 浙江杭州 310014)

摘 要: 本文使用高维空间点分布分析原理, 在仿生模式识别高维空间点覆盖原理的基础上, 提出了一种基于高维空间点覆盖动态搜索理论的非特定人连续数字语音识别的新算法, 这种算法可以不经过端点检测和分割, 通过对被识别连续数字语音直接进行动态搜索, 得到被识别语音到各类高维空间覆盖范围的距离随时间变化曲线, 通过距离曲线上的极小值点进行识别。

关键词: 连续语音识别; 高维空间点覆盖; 非特定人语音识别

中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2005) 10-1790-04

Research on Speaker-Independent Continuous Figure Speech Recognition Based on High-Dimensional Space Covering and Dynamic Scanning

WANG Shou-jue^{1,2}, PAN Xiao-xia², XU Chun-yan², CHEN Xu¹, AN Dong¹, CAO Wen-ming²

(1. Lab of Artificial Neural Networks, Institute of Semiconductors, CAS, Beijing 100083, China;

2. Research Institute of Intelligent Information System, Zhejiang University of Technology, Hangzhou, Zhejiang 310014, China)

Abstract: Through analyzing samples distribution in high dimensional space, a novel algorithm for speaker-independent continuous figure speech recognition is presented based on high-dimensional space covering, biomimetic pattern recognition and dynamic scanning in this paper. Without endpoints detection and segmenting the continuous speech is dynamic scanned, then distance curves from the continuous speech to every kind of covering area in high dimensional space are obtained. The continuous speech is recognized by detecting the minimum values in the curves.

Key words: continuous speech recognition; high-dimensional space covering; speaker independent

1 引言

目前小词汇表非特定人的孤立词识别系统已经实用化, 但较好的实用连续语音识别系统依然很少, 因此 20 世纪 90 年代以来, 语音识别研究主要集中在提高非特定人的大词汇量连续语音识别 (Large Vocabulary Continuous Speech Recognition, 简称为 LVCSR) 的性能上^[1]. 非特定人的连续数字语音识别算法的研究, 不但可以作为非特定人大词汇量的连续语音识别新算法的探讨; 而且有非常可观的应用价值, 比如声控电话交换、语音拨号系统、电话查询系统、自动应答服务、数字报表等等。

传统的语音识别算法是先将语音通过自动切分机切分成语音段, 再对切分出来的固定的语音段, 采用隐马尔可夫的统计模型^[2] (Hidden Markov Models, HMM) 或者动态时间规整 (Dynamic Time Warping, DTW) 等方法进行识别, 这类方法的好处是可将连续语音的识别转变成较为简单的孤立词的识

别, 从而降低识别的难度。

HMM 模型是目前最流行的方法, 它将语音识别系统的训练模型分成声学模型和语言模型。声学模型的训练采用多步训练的方法; 语言模型的训练是通过大量的语料进行统计而实现的。因此系统的识别算法要根据语言的特点和统计模型的整体结构进行设计, 然后按照一定的搜索规则 (深度优先或宽度优先) 进行解码^[3]。对于孤立词或语速较慢的语音而言, 可以得到比较高的识别率, 而且技术也已比较成熟^[4]。

然而对于语速不定, 环境不定的类似自然口音语音的连续语音而言, 这种方法的识别效果就很不理想了, 这是因为这种算法在本质上还是将特征空间划分为几个状态, 识别结果在这几个状态之间进行比较、转移、搜索, 而不是像仿生模式识别^[5]的先认识再区别。因此, 由于过分的依赖语音的端点检测的正确率, 得不到真正稳健的识别效果。而对于一段语速不定的未知语音而言, 即使用最精确的端点检测方法也无法保证其 100% 的正确率。为了解决连续的语音识别对端点的依

赖性问题,本文提出了一种基于高维空间点覆盖动态搜索理论的非特定人连续数字语音识别的新算法,这种算法可以不经过端点检测和分割.先根据实际连续数字语音的各不同数字音节,构建连续语音中各不同数字音节的特征空间覆盖区.在识别时,利用高维空间点覆盖动态搜索理论进行识别,得到了较为满意的识别结果.

2 连续语音的特点

- (1) 语音单元(如音素)间存在协同发音(co-articulation);
- (2) 由于端点的交错融合,因此很难分割;
- (3) 语速不定,使得字长不定,从而很难找到统一的模版进行匹配;
- (4) 有类似自然口语语音^[6]的特点,比较随意,有少量犹豫、停顿和填音现象.用传统划分的方法无法去除这些不在预定状态范围内的多余音素或音节.

3 连续数字语音识别的特征提取方法和高维空间分类覆盖区的神经网络构筑

3.1 构筑神经网络所用样本库的建立

本文所采用的连续语音是介于自然口语语音^[6](spontaneous speech,是指随意的、至少没有在讲话方式上经过特殊准备的语言,也就是人们在日常生活中通常所讲的“话”.它通常是不流畅的,包含许多随机事件)和“朗读式语音”(是指符合语法规则的,流畅的,讲话方式和讲话内容都经过特殊准备的语言)之间的一种语音.它在内容上尽量包括所有有连音的数字,以自然口语的语气和语速进行朗读.在相对安静的实验室背景下录音,采样频率为 8000Hz,位深度为 16bit.根据这些具有协同发音的很难分割的连续语音,以人工分段试听的方式找到听觉最佳的连续语音切分点,以此构建连续语音中的单音节样本库.必须强调指出,该单音节样本库不同于一般孤立语音的样本库.


3.2 构筑神经网络所用样本的特征提取方法

构筑神经网络所用样本的特征提取分两步进行:

第一步:将切分好的单数字样本按 Mel 倒谱方式提取特征向量,其过程简述如下:

(1) 对单数字语音样本经过预加重处理: $x(n) = x(n) - 0.935 * x(n-1)$.

(2) 再经过窗宽为 256, 帧移为 64 的汉明窗进行分帧处理: $x(n) = [0.54 - 0.46\cos(2\pi n/255)]x(n)$

(3) 每一帧数据再 C_1 C_n
经过有 24 个滤波器组 
的 Mel 倒谱变换,将得到的 24 个 Mel 倒谱系数^[10](MFCC)去掉第 1 个带有明显能量特征的系数,以及最后 7 个趋近于零的系数,留下 16 个系数作为特征参数.

第二步:将冗余数据剔除

(1) 将每 16 个特征参数组成一个向量 $C_i, i = 1, 2, 3, \dots, n$,如图 1 所示:

图 1 经过 MFCC 变换后的单数字

(2) 计算相邻两个 16 维向量之间的夹角 $\theta_j = \arccos \frac{C_j \cdot C_{j+1}}{|C_j| \cdot |C_{j+1}|}$,当该夹角小于统计实验数据 0.13rad 时,则删去 C_j 或 C_{j+1} 中的一个,直到相邻向量间的夹角都大于等于 0.13rad.

第三步:将数据压缩规整为一定的长度

(1) 从压缩完的每一类(即 ling 到 jiu 中间的一个数)MFCC 形式的单数字音节(以下称 MFCC 单数字音节)中选取最短的一个,对该 MFCC 单数字音节做如下截取:用人工试听的方法,挑选试听效果最佳的连续 8 个向量(共 16×8 个值),将这 128 个数值组成的高维特征向量作为这类 MFCC 单数字音节的参考标准.

(2) 将每类中所有 MFCC 单数字与该类标准比较,选取夹角最小的连续 8 个 16 维向量所组成的 128 维向量,作为该类 MFCC 单数字音节的各个样本特征向量,用以构造特征空间中的识别覆盖区.其比较过程如图 2 所示:

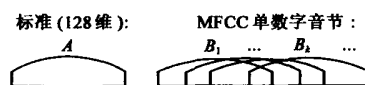


图 2 相似帧的选取

设: $\theta_k = \arccos \frac{A \cdot B_k}{|A| \cdot |B_k|}$, 若 $\min \theta_k = \theta_p = \min \theta_k, k = 1, 2, \dots, n$, 则 128 维向量 B_p 与标准 A 夹角最小.因此,选取 B_p 为 MFCC 单数字音节的一个特征向量,作为构造该类特征空间覆盖区的一个样本.

如上所述,特征提取的过程如图 3 所示:

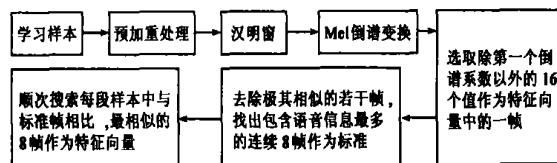


图 3 学习样本的特征提取框图

3.3 构造特征空间识别覆盖区^[5,7~9]

共有 11 类样本,“yi”和“yao”各一类其余每个数字一类,设这 11 类样本中的每一类样本(男女各 14 人,每人每类 10 个样本)组成的集合为 $S_i (i = 1, \dots, 11)$,对每一类样本进行学习,并采用作者提出的高维空间点覆盖方法^[5,7~9]构建一个神经网络.

4 基于高维空间点覆盖动态搜索理论的非特定人连续数字语音识别的新算法及其实现

4.1 被识别的连续语音样本的特征提取方法

根据 3.2 中的第一步和第二步,对长度不同的待识别的连续语音样本进行特征提取后,得到帧数 m 不同的 16 维向量串.再以连续的 8 个 16 维向量为新的帧长,作为一个窗口,构成一个 128 维的特征向量,作为一个待识别的点.以 1 个 16 维向量为新的帧移,移动窗口形成 128 维 $\times n (n = m - 7)$ 的长度不等的特征向量串(n 随待识别语音的长短的不同而不同,实验中最短为 207,最长为 465).

4.2 高维空间点覆盖动态搜索识别方法

在识别时,将待识别的 8 个数字连续语音(以“san wu ba er qi ling ling san”的连续语音串为例)所提取的长度不等的特征向量(128 维 $\times n$)作为高维空间的 n 个点,求出这 n 个点中依次各点到每一类覆盖范围的最小距离。

5 实验结果与讨论

5.1 实验结果

实验共采用了 32 人,每人说不同的连续 8 个数字的连续语音串作为待识别样本集,共计应识别数字 256 个,其中正确识别 218 个,误识 26 个。构筑网络所用连续语音样本是由 24 人采集。

下面以“san wu ba er qi ling ling san”的连续语音串为例讨论。识别结果曲线如图 5(a~f)。图中横坐标为时间轴,纵坐

标为各点(n 个点)依次离开各类(11 类)覆盖范围的最小距离。其中图 5(a)为离开“san”覆盖范围的距离随时间的变化曲线,明显的两个极小点显示了连续语音串中的两个“san”。图 5(b)为离开“wu”覆盖范围的距离随时间的变化曲线,其中极小点显示了语音串中的“wu”。图 5(c),5(d),和 5(e)分别为离开“ba”“er”“qi”覆盖范围的距离随时间的变化曲线,对应了连续语音串中的“ba”“er”和“qi”。图 5(f)是离开“ling”覆盖范围的距离随时间的变化曲线,其中第一个极小值由于时间上和图 5(c)中离“wu”覆盖范围的距离曲线的极小值重叠,而离“wu”覆盖范围的距离曲线的极小值相对较小,即极小值点离“ling”覆盖范围的距离大于离“wu”覆盖范围的距离,根据最近邻原则,因而是“wu”而不是“ling”。图中后两个极小值对应了两个“ling”。因而该识别结果为:“san wu ba er qi ling ling san”。

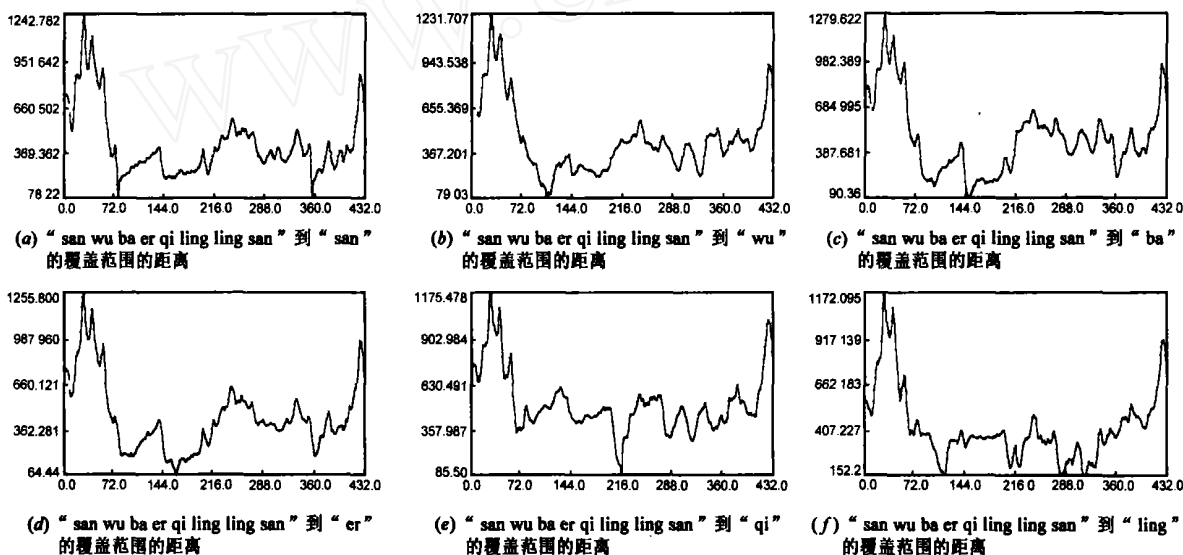


图 5

5.2 讨论

本文的目的是对非特定人连续自然语音识别的一种新方法的探讨,由于该方法不需要端点检测和分割,通过对被识别连续数字语音直接进行动态搜索,得到被识别语音到各类高维空间覆盖范围的距离随时间变化曲线,通过距离曲线上的极小值点进行识别。一般来说,语音中的噪音、停顿以及语音中的非关键信息对应于距离较大的部分,数字音节的关键信息对应于距离曲线的极小值部分,因此本方法具有很强的鲁棒性,较适用于长短不同语速不同的连续自然语言的识别。虽然目前实验效果还不十分理想,但可以看到该方法将是一种很有前途的连续自然语言识别新方向。

参考文献:

- [1] 张雄伟,陈亮,杨吉斌.现代语音处理技术及应用[M].北京:机械工业出版社,2003.202.
- [2] L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257 - 286.
- [3] 刘加.汉语大词汇量连续语音识别系统研究进展[J].电子学

报,2000,28(1):85 - 91.

- Liu Jia. Research on large vocabulary mandarin Chinese continuous speech recognition system[J]. Acta Electronica Sinica, 2000, 28(1): 85 - 91. (in Chinese)
- [4] 李虎生,刘加,刘润生.高性能汉语数码语音识别算法[J].清华大学学报(自然科学版),2000,40(1):32 - 34.
- Li Husheng, Liu Jia, Liu Runsheng. High performance digit mandarin speech recognition[J]. Tsinghua Univ(Sci &Tech), 2000, 40(1): 32 - 34. (in Chinese)
- [5] 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. Bionimetic (topological) pattern recognition —— a new model of pattern recognition theory and its applications[J]. Acta Electronica Sinica, 2002, 30(10): 1417 - 1420. (in Chinese)
- [6] 冯俊兰,杜利民.自然口语语音识别研究概况[J].电子科技导报,1999,(9):3 - 7.
- [7] 王守觉,王柏南.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 Sini-

- ca, 2002, 30(1): 1 - 4. (in Chinese)
- [8] 王守觉, 徐健, 王宪宝, 覃鸿. 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, Qin Hong. Multi-camera human-face personal identification system based on the biomimetic pattern recognition[J]. Acta Electronica Sinica, 2003, 31(1): 1-3. (in Chinese)
- [9] 王守觉, 等. 通用神经网络硬件中神经元基本数学模型的讨论[J], 电子学报, 2001, 29(5): 577 - 580.
Wang Shou-jue, Li Zhao-zhou, Chen Xiang-dong, Wang Bai-nan. Discussion on the Basic Mathematical Models of Neurons in General Purpose Neurocomputer[J]. Acta Electronica Sinica, 2001, 29(5): 577 - 580. (in Chinese)
- [10] R J Mammone, X Zhang, R P Ramachandran. Robust speaker recognition: A feature-based approach, IEEE Signal Processing[J]. 1996(13): 58 - 71.
- [11] 王守觉, 徐春燕, 潘晓霞, 安冬, 陈旭, 曹文明. 为连续语音识别用的单词音节神经网络建模的研究[J]. 电子学报, 2005, 33(10). Wang Shou-jue, Xu Chun-yan, Pan Xiao-xia, An Dong, Chen Xu, Cao Wei-ming. Single figure syllable modeling based on neural network for continuous speech recognition[J]. Acta Electronica Sinica, 2005, 33(10): 1883 - 1885. (in Chinese)

作者简介:

王守觉 男, 1926 年生于江苏, 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, 现为半导体神经网络实验室负责人, 是我国半导体器件与微电子奠基人之一, 为研究解决我国发展两弹一星所需高速计算机的半导体器件与集成电路作出过重要贡献, 在国家“八五”、“九五”科技攻关中, 承担了半导体神经网络的实现和应用技术的攻关工作, 研制了我国最早可商品化的神经计算机 CASSANN-1、CASSANN-2 等, 提出了新的“仿生(拓扑)模式识别”理论和“高维空间复杂几何形体覆盖”识别方法, 为模式识别开辟了一个崭新的理论与实现的新途径, 《电子学报》和《CE》主编, 中国计算机学会 CAD 与图形学专业委员会名誉主任.

E-mail: wsjue@semi.ac.cn.

潘晓霞 女, 1979 年生于浙江, 研究生, 主要研究模式识别、智能信息处理等.

徐春燕 (见本期第 1885 页)

陈旭 女, 1978 年出生于陕西, 博士生, 主要研究仿生模式识别.

安冬 女, 1977 年出生于河北, 博士生, 主要研究仿生模式识别.

曹文明 男, 1965 年 11 月出生于江苏, 教授、硕士生导师, 研究方向: 神经网络在模式识别和控制中的应用.