

为连续语音识别用的单词音节神经网络建模的研究

王守觉^{1,2}, 徐春燕², 潘晓霞², 安 冬¹, 陈 旭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: 本文主要研究连续语音中单词音节的神经网络建模问题. 采用了一种富有特色的特征提取方法, 并依据高维空间点覆盖理论, 对实际连续数字语音的各不同数字音节, 以人工切自连续数字语音中的 2640 个单字音节, 构建连续语音中各不同数字音节的特征空间覆盖区, 并使用 7308 个自连续数字语音中切分出的单字音节, 利用仿生模式识别原理, 进行了建模正确性验证. 验证结果正确率达到 97% 以上, 对同样数量的少量建模样本, 识别率优于 SVM 方法.

关键词: 连续语音; 单词音节; 高维空间点覆盖; 神经网络模型

中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2005) 10-1883-03

Single Figure Syllable Modeling Based on Neural Network for Continuous Speech Recognition

WANG Shou-jue^{1,2}, XU Chun-yan², PAN Xiao-xia², AN Dong¹, CHEN Xu¹, 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 310014, China)

Abstract: The single figure syllable modeling based on neural network for continuous speech recognition is discussed. A new feature extraction method is used which mainly includes compressing single figure frames according to a certain inter-frame angle, extracting representative information comparing to standard single figure of fixed length. 2640 single figure syllables made from continuous speech are used to construct each kind of high dimensional space covering area. By biomimetic pattern recognition theory 7308 single figure syllables made from continuous speech are used to confirm this model in CASSANN-II neural computer and get a quite good result. Experiments show the recognition rate is higher than SVM when the training samples are small.

Key words: continuous speech; high dimensional space covering; single syllable; neural network modeling

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^[1~3]. 传统的语音识别算法大都采用隐马尔可夫统计模型^[4] (Hidden Markov Models, HMM) 或者动态时间规整 (Dynamic Time Warping, DTW) 等方法进行识别. 但使用这两种方法进行语音识别时, 一般都需要进行端点检测和分割, 连续语音由于受到协同发音 (co-articulation) 的影响, 使得端点的检测和分割变得非常困难; 再加上语速的不固定使得模型的确定变得更加困难. 因此, 由于传统的识别方法过分依赖语音的端点检测和分割的正确率, 得不到真正鲁棒的识别效果.

鉴于上述原因, 作者提出了一种不依赖于端点检测和分割的非特定人连续语音识别方法. 在解决非特定人连续语音识别问题前, 必须对连续语音中的单词音节建立识别模型. 因

而, 本文主要研究连续语音中单词音节的神经网络建模问题, 以提供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^[11]中应用, 这将有助于不依赖于端点检测和分割的非特定人连续语音识别算法的实现. 本文采用介于自然口语语音和“朗读式语音”之间的一种连续语音, 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, 构建连续语音中的单词音节样本库. 为使单词音节样本在特征空间中具有相同的维数, 本文采用一种新的特征提取方法. 依据高维空间点覆盖理论, 对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, 并应用仿生模式识别原理, 用连续数字语音中 7308 个单词音节作为对神经网络建模正确性的验证, 取得了较好的效果, 对同样数量的少量建模样本, 识别率优于 SVM 方法.

2 连续数字语音中单音节样本库的建立

本文采用从 0 到 9 共 11 类数字 (其中 1 有“yi”和“yao”两种发音), 考虑了每两个数字排列的 112 种可能性, 共设计了

收稿日期: 2004-07-12; 修回日期: 2005-07-21

每 8 位一组共 18 组电话号码,包含尽可能多的数字间的排列组合.在多种非绝对安静的环境下,使用不同的录音工具,对不同年龄共 36 人录音,每人用自然语速把 18 组号码读 5 遍,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,以此构建连续数字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.该样本库共有 7308 个样本,其中 2640 个样本为训练样本(选不同年龄段的男女各 12 人(表 1),每人每类 10 个样本)

表 1 参与构建非特定人连续汉语数码语音高维空间点覆盖模型的人员组成

	<20 岁	20 - 30 岁	30 - 40 岁	40 - 50 岁	50 - 60 岁	>60 岁
男	2	2	2	2	2	2
女	2	2	2	2	2	2

3 构筑神经网络所用样本的特征提取方法,简述如下

新的样本特征提取方法可分三步进行,流程图如图 1 所示.

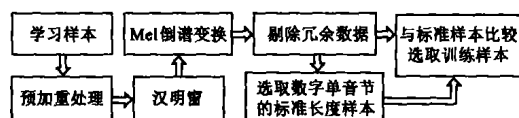


图 1 特征提取框图

第一步:将人工切分好的单数字音节按 Mel 倒谱^[5]方式提取特征参数

人的听觉系统是一个特殊的非线性系统,它对不同频率信号的灵敏度是不同的,基本上是一个对数的关系. Mel 倒谱参数 (Mel-Frequency Cepstrum Coefficients, MFCC) 能够比较充分反映人耳这种特殊的感知特性,所以本文采用 MFCC 为基本的特征参数,其实现步骤简述如下:

(1) 对单数字语音样本经过预加重处理:

$$x(n) = x(n) - 0.9375 * x(n-1).$$

(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 (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

将每 16 个特征参数组成一个向量 $C_i, i = 1, \dots, n$, 计算相邻两个向量之间的夹角 $\theta_i = \arccos \left(\frac{C_i \cdot C_{i+1}}{|C_i| \cdot |C_{i+1}|} \right)$, 当该夹角小于统计实验数据 0.13rad 时, 则删去 C_i 或 C_{i+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 MFCC 形式的单数字音节 (以下简称 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.

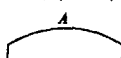
(2) 将每类中所有

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

标准 (128 维):



MFCC 单数字音节:



图 2 相似帧的选取

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:

$$\text{设: } \theta_k = \arccos \left(\frac{A \cdot B_k}{|A| \cdot |B_k|} \right), \text{ 若 } \theta_{\min} = \theta_p = \min \{ \theta_k, k = 1, 2, \dots, n \}, \text{ 则 128 维向量 } B_p \text{ 与标准 } A \text{ 夹角最小. 因此, 选取 } B_p$$

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.

4 构造特征空间识别覆盖区^[6,8~10]

对 11 类样本中的每一类样本进行学习, 采用作者提出的高维空间点覆盖方法^[6,8~10]构建非特定人连续汉语数字语音模型 $C(i), i = 1, \dots, 11$, $C(i)$ 为第 i 类非特定人连续汉语数字语音模型.

表 2 各类样本到其他类样本覆盖区的平均距离

被识别类别 \ 网络类别	ling	yi	er	san	si	wu	liu	qi	ba	jiu	yao
ling	1	1.6	2.1	2.0	2.0	1.8	1.7	1.9	2.7	1.7	2.0
yi	1.4	1	2.9	2.3	2.1	2.1	1.9	2.0	3.0	1.8	2.5
er	2.5	3.2	1	1.8	2.1	2.2	2.3	2.7	1.4	2.6	1.9
san	2.2	2.8	1.7	1	2.0	2.3	1.9	2.5	1.7	2.1	1.6
si	2.4	2.8	2.6	2.0	1	3.2	2.3	1.6	2.7	2.0	2.4
wu	2.1	2.8	2.4	2.2	2.7	1	2.5	3.4	2.4	2.6	2.3
liu	1.5	1.8	2.0	1.7	1.9	2.2	1	2.1	2.1	1.7	1.6
qi	2.4	2.6	3.1	2.5	1.5	3.8	2.6	1	3.2	2.1	2.9
ba	2.6	3.4	1.4	1.9	2.2	2.1	2.5	2.8	1	2.8	1.9
jiu	1.5	1.9	2.5	1.9	1.7	1.8	1.7	1.7	2.7	1	1.9
yao	1.9	2.5	1.6	1.6	2.1	1.7	1.6	2.5	1.5	2.0	1

5 实验与结果

用 11 类构造好的高维空间点覆盖模型 $C(i), i = 1, \dots, 11$, 来识别 MFCC 序列样本库中的全部样本 $A(j), j = 1, \dots, 7308$. 模型输出值为样本 $A(j)$ 到高维空间点覆盖区域 $C(i)$ 的最小距离 $(A(j), C(i))$. 由 $[A(j), k] = \min_{i=1, \dots, 11} (A(j), C(i))$, 求出 $A(j)$ 到 11 类模型 $C(i)$ 的最小距离, 最小距离所对应的第 k 类 $[A(j), k]$, 即为 $A(j)$ 所属的类别.

为了解各类单数字音节样本相互间的距离分布情况, 对所有已知的切分好的单数字样本进行识别验证, 以各类样本到本类覆盖区的平均距离为 1, 各类样本到非本类覆盖区的平均距离如表 2 所示.

本文用所建立的单词音节神经网络识别模型对连续数字语音中 7308 个单词音节进行了识别实验作为对神经网络建模正确性的验证. 验证结果说明, 这样的神经网络建模结果对

连续语音中的单词音节区分的平均正确率达到 97.4 %。说明特征提取方法和神经网络模型是有效的。全部 7308 个样本的识别验证结果如表 3 所示。

表 3 全部 7308 个样本的识别验证结果

类别	ling	yi	er	san	si	wu	liu	qi	ba	jiu	yao
总数量	463	534	555	650	950	719	471	1087	315	888	676
识别结果	ling	445	7	0	1	0	17	4	0	4	0
	yi	3	527	0	0	0	1	0	0	0	0
	er	0	0	518	0	0	1	0	2	0	0
	san	0	0	0	643	0	0	0	0	0	1
	si	1	0	0	1	932	0	34	0	6	0
	wu	4	0	1	1	0	700	0	0	0	0
	liu	8	0	0	1	2	0	459	0	11	4
	qi	0	0	0	0	14	0	0	1050	0	1
	ba	0	0	35	1	0	0	0	313	0	5
	jiu	1	0	0	0	2	0	3	3	0	865
	yao	1	0	1	2	0	1	4	0	0	1
识别率	0.961	0.987	0.933	0.989	0.981	0.974	0.975	0.970	0.994	0.974	0.985

从表 2 和表 3 都说明了在识别连续数字语音中的 er 时, 被误识为 ba 的可能性最大(6.3 %), 其占全部误识情况的 18.4 %, 因而, 在进一步提高连续数字语音识别率的工作中, 对于发音 ba 的清辅音“b”的捕捉与识别将是工作中的一个重要方面。尽管根据表 2, 单词音节 er 到 ba 网络覆盖区的相对平均距离与单词音节 ba 到 er 网络覆盖区相对平均距离相差不远, 但是表 3 中, 连续数字语音中的 er 被误识为 ba 的可能性远大于 ba 被误识为 er。这是因为在高维特征空间中, 连续数字语音中 er 到 ba 网络覆盖区的距离分布比较分散, 而 ba 到 er 网络覆盖区的距离分布较为密集。

支持向量机法^[12](SVM)是建立在统计学习理论基础上的机器学习方法, 使用最优分类面作为样本分类的基准面, 兼顾到经验风险和置信范围, 具有相对优良的性能指标。采用径向基函数为核函数的 SVM 方法和

高维空间点覆盖的神经网络方法, 用不同数量的少量训练样本建模, 对测试集的识别率(图 3)上后者优于前者, 建立高维空间覆盖区比最优分类面有更好的识别效果。

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]. 机械工业出版社, 2003, 202.
- [2] 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)

- [3] 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)
- [4] 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.
- [5] R J Mammone, X Zhang, R P Ramachandran. Robust speaker recognition: A feature-based approach[J]. IEEE Signal Processing, 1996(13): 58 - 71.
- [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. (in Chinese)
- [7] 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.
- [8] 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 Sinica, 2002, 30(1): 1 - 4. (in Chinese)
- [9] 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)
- [10] 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)
- [11] 王守觉, 等. 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[J]. 电子学报, 2005, 33(10): 1 - 4.
- Wang Shou-jue, Pan Xiao-xia, Xu Chun-yan, Chen Xu, An Dong, Cao Weir-ming. Research on speaker-independent continuous figure speech recognition based on high-dimensional space covering and dynamic scanning[J]. Acta Electronica Sinica, 2005, 33(10): 1 - 4. (in Chinese)
- [12] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273 - 297.

作者简介:

王守觉 (见本期第 1793 页)

徐春燕 女, 1969 年生于山东, 研究生, 主要研究模式识别、智能信息处理等, E-mail: xchunyan @126.com.

潘晓霞 (见本期第 1793 页)

安冬 (见本期第 1793 页)

陈旭 (见本期第 1793 页)

曹文明 (见本期第 1793 页)

图 3 高维空间点覆盖的神经网络方法与 SVM 比较

