DOI: 10.3969/J. ISSN. 1672-6693. 2010. 04. 018

基于神经网络的语音识别研究*

滕 云, 贺春林, 岳 淼
(西华师范大学 计算机学院, 四川 南充 637002)

摘要: 由于具有良好的抽象分类特性,神经网络现已应用于语音识别系统的研究和开发,并成为解决识别相关问题的有效工具。为解决一般语音识别系统准确率较低的问题,本文分别给出了由循环神经网络(RNN)和多层感知器(MLP)组成识别模块的两种语音识别系统,并对二者识别的准确性进行了比较。介绍了特征提取模块的主要工作步骤并讨论了组成识别模块的上述两种神经网络结构。其中,特征提取模块利用线性预测编码(LPC)倒谱编码器,把输入语音翻译成 LPC 倒谱空间中的曲线;而识别模块完成对某个特征空间曲线之间的联系和单词的识别。实验结果表明,MLP 方法准确率高于 RNN 方法,而 RNN 方法准确率可达 85%。

关键词: 神经网络; 语音识别; 循环神经网络; 多层感知器; 线性预测; 矢量量化
中图分类号: TP391 **文献标识码:** A **文章编号:** 1672-6693(2010)04-0073-04

一个语音识别系统主要由两个不同的模块组成: 特征提取和识别[1],如图 1 所示。特征提取模块使用标准的线性预测编码(LPC)倒谱编码器,它把输入语音翻译成 LPC 倒谱特征空间中的曲线。这些在降维空间中的曲线能对说出的词汇提供可靠的表征,同时降低了训练的复杂度和减轻了识别的工作。

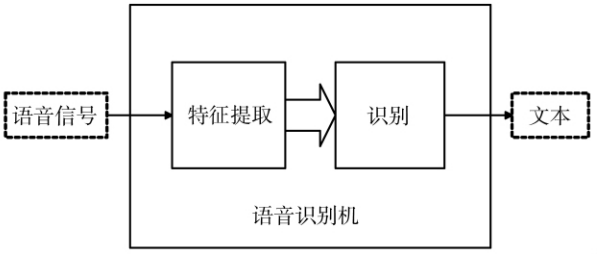

图 1 语音识别系统组成框图

特征提取模块的输出并不能得到由上述特征空间曲线表征的单词。它仅仅是把输入的语音压力波转换成某个特征空间的曲线。这些曲线之间的联系和单词的识别是由识别模块完成的。本文分别用二种神经网络[2-5]来构建识别模块,即循环神经网络(Recurrent neural networks, RNN)和多层感知器(Multi layer perceptrons, MLP)。

1 基本原理

在识别阶段使用神经网络,本文实现了一个简单的依赖于语音识别系统的扬声器,它能够识别单个的阿拉伯数字“0”~“9”。另有许多其他方法被有效地用于语音识别,比如,模式识别方法、隐马尔可夫模型(HMM)方法等[6],但本文使用的是神经网络。使用具有神经网络能力的模式识别以及其他的数学和信号处理工具,一个语音识别系统能正确地辨识出简单的字。系统可识别出经过训练的样本,且也能够归纳到同一个字的其他样本。当使用较大的词汇量时,系统识别的准确率将降低。

开发这种语音识别程序的第一步是设计一个特征提取器。依照在人类生物学研究中所取得的阶段成果及其发展,特征提取模块可被模型化[7]。它能够把输入的声音转换成内部的表示,而通过它可重建原始信号。这个阶段可依照听力器官功能模型化,它首先把输入的空气压力波转换成液体压力波,然后再把它们转化成特定的神经元放电模式。

特征提取模块的输出应能在后继阶段对这些数据开始工作之前降低问题的复杂度。此外,在输入空间中点序列之间存在的相关关系必须被保留在输出空间的点序列之中。特征提取模块对信号空间中

* 收稿日期: 2010-03-10
资助项目: 四川省教育厅重点科研项目(No. 08ZA018); 校级科研项目(No. 06A002)
作者简介: 滕云,男,讲师,硕士,研究方向为软件理论、算法理论和图形图像、信号处理。

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

点序列处理的速度不必和特征空间中点序列产生的速度相同。这意味着,与输入信号空间的时间相比,输出特征空间中的时间可以发生在不同的时刻。

一旦特征提取模块完成了工作,识别模块就对其输出结果进行界定。识别模块把音素序列集成为单词,它把每个输入曲线界定为特定词汇表中的一个单词。这种把发音和它的符号表达关联起来,把口语翻译成书面语言的过程即称为语音识别[3]。

2 特征提取

在特征提取阶段,首先记录下单个说话人的语音样本,对每个字至少要用5个样本来训练神经网络。然后,提取每个字的LPC倒谱系数,应用矢量量化(Vector quantization)的K均值算法(K-means algorithm)[4]来得到简化的曲线。

特征提取主要由下面的步骤组成(如图2所示)。

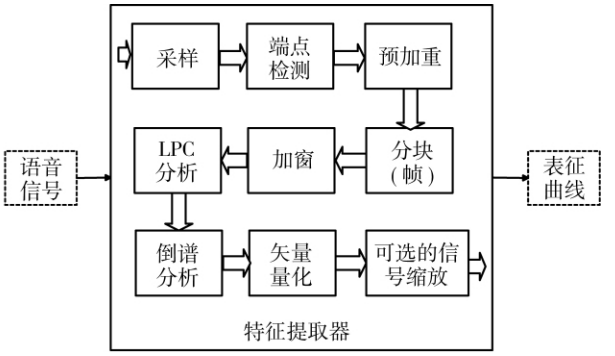

图2 特征提取原理框图

1) 语音采样:用相对便宜的动态麦克风和标准的PC机声卡来对语音记录和采样。采样频率为22 050 Hz,采样位数为16位。

2) 端点检测:一种快速的鲁棒技术被用来精确地定位孤立词汇的端点[0]。这种技术利用帧能来获得参考点。此算法取帧的大小为100个样本,并计算每个帧的能量且求所有帧能量的均值来得到能量的参考值。

每帧的能量用$p[] = \sum_{k=1}^{j} s[]^2$计算,其中$s[]$是帧中的语音数据。

对所有帧计算其p值,p的均值表示最终的帧能量值E。其中,$E = \sum_{k=1}^{m} p[]^2/m$。阈值设为$c^* E$($c$是常量),作为检测标准。

4) 预加重法:一个预加重滤波器被用来对数字语音的信号扁平化,并消除在后续的计算中有限数值精度的影响。这种类型的滤波器提升高频部分的量级,对低频部分则相对未处理。

5) 分帧加窗:信号采样之后,语音被分离,频谱更平滑,每个信号被分成数据块序列,每个数据块含300个样本,而数据块之间有100个样本。每个数据块乘上一个同样宽度的汉明(Hamming)窗,以减少频谱泄露的影响[1]。

6) LPC分析:用自相关法(德宾Durbin)对每个数据块得到一个12阶的LPC倒谱系数的矢量。

7) 矢量量化:用矢量量化技术对LPC倒谱矢量降维。经矢量量化共得到36个系数。在矢量量化中使用的是K均值算法。

含L个训练矢量的集合聚为含M个码本矢量的集合的方法如下。

1) 初始化:从初始的L个训练矢量集中任意地选取M个矢量,把它们作为码本中的初始码字集。

2) 最近邻搜索:对每个训练矢量,依照谱距,在当前码本中找到最接近的码字,然后把矢量分配给相应的胞腔(Cell)。

3) 距心更新:用分配给每个胞腔的训练矢量的距心对码字更新。

4) 迭代:重复第2步和第3步,直到平均距离降到预设的阈值以下。

在矢量量化阶段之后,只剩下3个矢量。最后阶段的输出就是用于识别阶段的最终特征。

3 识别

识别模块使用神经网络构建。使用的两种类型神经网络是多层感知器和循环神经网络。一个神经网络就是“神经元”层的集合,它模拟人类大脑结构。每个神经元从上一层的每个神经元取得信号输入(或如是第一层,则从外部世界取得输入)。然后,它对输入信号进行增强,并把它传到下一层。每个层之间的连接都有一定的权。神经网络每次处理一些输入,并调整这些权值,使得输出更接近给定的希望的值。几次重复之后(每次重复都是一遍循环),得到一个正确的输出,它松散地近似于输入[2]。

2种不同的方法被用于语音识别。对每个字(实验采用数字“0”~“9”)使用5个不同的训练样

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

本来测试网络。对数字记录更多的样本来计算识别的准确性。

3.1 多层感知器(MLP)方法

图 3 是带有 2 层隐含层的多层感知器结构,这是一种多层前馈网络。图中所有连接均为相邻层之间的节点的连接,同层之间不连接。输入层不作任何运算,它只是将每个输入量分配到各个输入节点。

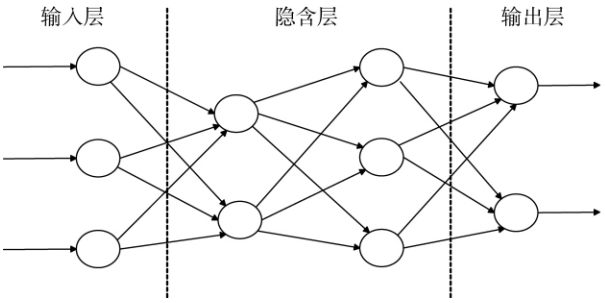

图 3　带 2 层隐含层的多层感知器的结构

这里设计的多层感知器有 36 个输入结点,36 个隐含的神经元和 1 个输出神经元。神经元的输出在区间 [-1,1] 内。每个神经元有一个额外的连接,其输入保持不变且等于 1(文献中,此连接通常指偏倚或阈值)。权初始化为在小区间内选定的随机值。使用误差反向传播方法(Error back propagation method) [3-4] 对 MLP 进行训练。表 1 是 MLP 方法对数字"0"~"9"测试的结果。

表 1　MLP 方法的测试结果　　%

数字	准确率	数字	准确率
0	90	5	80
1	100	6	90
2	100	7	80
3	100	8	100
4	70	9	80

3.2 循环神经网络(RNN)方法

图 4 是既有前馈通路又有反馈通路的循环神经网络的结构。反馈通路可将某一层神经元的激活输出经过一个或几个时间节拍之后送到同一层的神经元,或送到较低层的神经元。反馈通路使网络可"记忆"以前输入所引起激活特性 [5]。

在对网络训练时,会出现下面的情况。

对每个阶段:

1) 整个输入序列提交给网络,计算得到它的输出且和目标序列比较,产生一个误差序列。

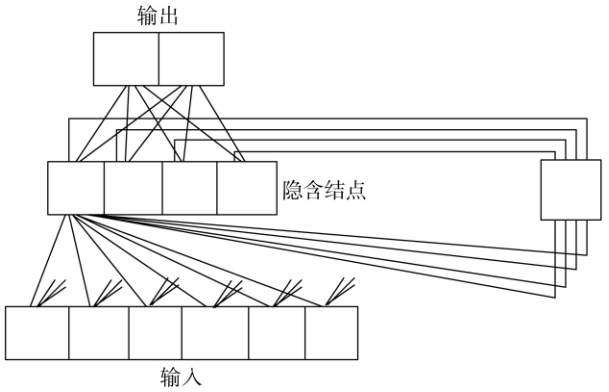

图 4　循环神经网络结构

2) 对每个时间步,误差被回传以找到每个权值和偏倚的误差梯度。由于权值和偏倚通过延迟的循环连接对误差的贡献常被忽略,这个梯度实际上是近似值。

3) 通过用户选择的回传训练函数,这个梯度用来更新权值。表 2 是用 RNN 方法对数字"0"~"9"测试的结果。

表 2　RNN 方法的测试结果　　%

数字	准确率	数字	准确率
0	80	5	90
1	80	6	90
2	100	7	70
3	80	8	100
4	70	9	90

4 结论

为解决一般语音识别系统识别准确率较低的问题,本文提出了分别由循环神经网络(RNN)和多层感知器(MLP)组成识别模块的两种语音识别系统。经实验表明,MLP 方法的识别准确率可达 89%,而 RNN 方法的准确率可达 85%。在性能上 MLP 方法优于 RNN 方法,但仍低于实际应用要求的限度,而 RNN 方法在空间需求上远低于 MLP 方法。下一步工作应继续分析各种数据之间的关系,对神经网络结构加以改进以进一步提高识别的实时性和准确率。

参考文献:

[1] Rabiner L R, Juang B H. Fundamentals of speech recognition [M]. Upper Saddle River, NJ: Prentice hall, 1993.

[2] 孟显勇,袁丁. 多层 BP 神经网络用于破译椭圆曲线密码

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

[J]. 四川师范大学学报(自然科学版),2005,28(3):371-375.

[5] 张彤,肖南峰. 基于BP网络的指纹识别系统[J]. 重庆理工大学学报(自然科学版),2010,24(1):47-50.

[4] 高富强,邹恒,秦昌硕. BP和RBF神经网络在字母识别中的比较[J]. 重庆工学院学报(自然科学版),2009,23(9):77-80.

[5] 宋智,何嘉. 面向复杂问题的BP神经网络并行算法[J]. 西南师范大学学报(自然科学版),2009,34(3):103-106.

[6] 朱鑫森,刘顺承. 基于神经网络与改进D-S证据理论的目标识别[J]. 四川兵工学报,2009,30(7):67-69.

[7] Gandhiraj R, Sathidevi P S. Auditory-based wavelet packet filterbank for speech recognition using neural network[A]// Proc Int Conf Adv Comput Commun, ADCOM[C]. Institute of Electrical and Electronics Engineers Inc, 2007: 666-671.

[8] Mohammad I, Shah R S, Saad P D. Improving speaker independent speech recognition process using speech recognition engine[A]// Proc Int Conf Artif Intell, ICAI Proc Int Conf Mach Learn; Models, Technol Appl, MLMTA[C]. Las Vegas, NV, United states: CSREA Press, 2008: 870-875.

[9] Lee Chin H, Rabiner, Lawrence R. Directions in automatic speech recognition[J]. NTT Review, 1995, 7(2): 19-29.

[10] Lalith Kumar T, Kishore Kumar T, Soundar Rajan K. Speech recognition using neural networks[A]// Int Conf Signal Process Syst[C]. IEEE Computer Society, 2009: 248-252.

[11] Manish S, Richard M. Speech recognition using subword neural tree network models and multiple classifier fusion[A]// ICASSP IEEE Int Conf Acoust Speech Signal Process Proc Proceedings[C]. NJ, United States: IEEE, 1995: 3323-3326.

[12] Ronan F, Edward J. Combined speech enhancement and auditory modeling for robust distributed speech recognition[J]. Speech Communication, 2008, 50(10): 797-809.

[13] 李源,邓辉文. 新型前馈网络学习算法在语音识别中的应用[J]. 计算机科学,2008,35(8):122-124.

[14] 余华,李海洋,李启元. 基于径向基神经网络的数字"0"~"9"语音识别[J]. 江西师范大学学报(自然科学版),2009,(06):701-705.

[15] 曹平,陈盼,章文彬,等. 基于脉冲神经网络的语音识别方法的初步探究[J]. 计算机工程与科学,2008,30(4):139-141.

[16] 赵力. 语音信号处理[M]. 北京: 机械工业出版社,2009.

Research on Speech Recognition Based on Neural Networks

TENG Yun, HE Chun-lin, YUE Miao

(College of Computer, China West Normal University, Nanchong Sichuan 637002, China)

Abstract: Because of good characteristics of the abstract classification, neural networks have become an effective tool for resolving issues related to recognition, and have been applied to the research and development of speech recognition systems. A speech recognizer system comprises of two blocks, Feature Extractor and Recognizer. For increasing the recognition accuracy, this paper proposes two types of speech recognition system whose recognition block uses the recurrent neural network(RNN) and multi layer perceptron(MLP) respectively. Furthermore, the main work steps of Feature Extractor(FE) block is introduced and the structure of two types of neural networks mentioned above is discussed. Using a standard LPC Cepstrum, the FE translates the input speech into a trajectory in the LPC Cepstrum feature space. The recognizer block discovers the relationships between the trajectories and recognizes the word. The results show that the MLP's recognition accuracies were better than the RNN's, while the RNN's recognition accuracies achieved 85%.

Key words: neural networks; speech recognition; recurrent neural network; multi layer perceptron; linear prediction; vector quantization

(责任编辑 游中胜)

?1994-2018 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net