第28卷 第3期 北京工业大学学报 Vol.28 No.3
2002年 9月 JOURNAL OF BEIJING POLYTECHNIC UNIVERSITY Sept. 2002

基于神经网络的汉语孤立词语音识别

孙光民，董笑盈
（北京工业大学 电子信息与控制工程学院，北京 100022）

摘 要：研究了基于神经网络的中文孤立词语音识别技术；将时间规整算法与神经网络相结合，组成一个混合级联神经网络语音识别系统．在这个模型中，第一级是时间规整神经网络．其作用是完成时间规整功能，从输入不等长的语音信号特征矢量序列中提取固定长度的特征矢量；然后将这组特征矢量馈入后一级 BP 网络完成语音识别．利用该方法对小词表汉语孤立词进行语音识别实验，获得了 98.25% 的正确识别率．实验结果表明，该系统不仅利用神经网络解决了语音识别中的时间规整难题，而且识别性能明显得到改善，识别率和训练速度均优于采用线性时间规整的神经网络语音识别方法．

关键词：语音识别；人工神经网络；时间规整
中图分类号：TN912.3 **文献标识码**：A **文章编号**：0254－0037(2002)03－0289－04

近年来，随着计算机技术的不断发展和广泛应用，语音识别技术得到了迅速发展，先后提出了动态规划(DP)、线性预测分析技术(LP)、动态时间规整技术(DTW)、矢量量化(VQ)和隐马尔可夫模型(HMM)理论等．尽管如此，目前语音识别研究中仍然存在着大量难题迫切需要解决．为了解决这些难题，研究人员又提出了各种新的修正方法[1,2]；所有这些努力都取得了一定成果，尤其是基于人工神经网络的语音识别系统受到人们的极大关注[3,4]．作者主要研究了基于神经网络的中小词表汉语孤立词语音识别技术，将人工神经网络中最经典的 BP 网络应用于语音识别，并针对 BP 网络用于语音识别时所遇到的时间对准问题，提出一种时间规整网络与 BP 网络混合而成的分类器模型，用于对小词表中文孤立词识别．

1 语音信号特征提取

特征提取是语音识别的基础．因为语音信号中含有丰富的信息，进行语音识别，首先要进行特征提取，将高维的原始信号空间变换到较低维的特征空间，即对语音信号进行分析处理，去除对语音识别无关的信息，提取有用的信息．对于非特定人语音识别，希望特征参数尽可能多地反映语义信息，减少个人信息．

线性预测(LP)分析法在语音识别时可用于特征提取．它通过采用一组简洁的语音信号模型参数来表达语音信号的频谱幅度，这些参数一般可看做是由线性预测系数推演而来．常用的参数包括倒谱系数、反射系数等．在语音识别中，最常用的是倒谱系数．p 阶线性预测倒谱系数的递推公式可表示为

$$c(n)=\begin{cases} c(1)=a_1 \\ c(i)=\sum_{k=1}^{n-1}(1-k/i)a_k\cdot c(i-k)+a_i \quad 1<i<p \end{cases} \tag{1}$$

倒谱 $c(n)$ 实际上是信号 Z 变换的对数函数的反 Z 变换，一般通过信号的傅里叶变换，对其模取对数，再求反傅里叶变换得到．因为线性预测分析法是一种谱估计方法，而且预测滤波器的频率响应 $H(e^{j\omega})$ 反映了声道的频率响应和信号的谱包络，所以用 $\log|H(e^{j\omega})|$ 作反傅里叶变换求出的倒谱系数便是一种描述语音信号的良好参数，并且计算量不大．往往只需要十几个倒谱系数就能较好地描述语音的共振峰特

收稿日期：2002-04-11．
作者简介：孙光民(1960–)，男，副教授，博士．

性，因此用于语音识别时，可以降低特征矢量维数，减少训练和识别时间．

作者选用加权的线性预测倒谱系数．首先应用 Burg 算法计算线性预测系数，然后递推求出倒谱系数，再经过带通提升(bandpass lifting)的视窗加权处理，视窗加权处理的目的是降低特征参数中剧烈变化的部分，而不改变语音信号的基频成分，使得倒谱分析能更好地表示语音信号中的波峰及细节变化特征．这样得到的倒谱系数称为加权倒谱系数．提取加权线性预测倒谱系数的全过程见图 1．

图1 提取加权LP倒谱系数的系统框图

语音信号是典型的非平稳信号，但可以假定语音信号是分段平稳的，即在 10～30 ms 其频谱特性和某些物理参数可以近似地看作不变．这样就可以采用平稳过程的分析处理方法来处理．因此，通过预处理和端点检测后，需要将语音信号分成各个短时段(分帧)进行处理．每秒的帧数视实际需要而定，一般为 33～100．本文中，语音信号采样率为 8 kHz；帧长取 30 ms，即 240 点；帧间重叠 15 ms，即 120 点．

2 基于神经网络的语音识别

2.1 语音识别系统方案

由于对不同的汉语孤立词，或不同人说相同的汉语词语时，发音长短、清浊音比例等都是变化的(即输入汉语语音词组信号的帧数不同)，而大多数神经网络分类器的输入结构是固定的，利用神经网络进行汉语孤立词语音识别时，存在着时间规整这一难题，这就意味着必须设法从可变长度的输入语音信号中提取相同维数的特征矢量，才能满足分类器的使用要求．为了解决时间规整问题，可以对语音信号进行线性时间规整，但这种方法可导致相同词组中的音素或类音素无法对准．另外，有人提出采用非线性时间规整算法来解决时间对准问题[5,6]．其中，动态时间规整(DTW)算法就是效果较好的一种非线性时间规整算法．其思路是采用动态规划技术，将一个复杂的全局最优化问题化为许多局部最优化问题，一步一步进行决策以寻找到最优路径，总的累计失真量最小时对应的路径就是最佳的时间规整函数．

作者将时间规整算法与神经网络相结合，组成一个混合级联神经网络语音识别系统．在这个模型中，第一级是时间规整神经网络，其作用是从输入不等长的语音信号特征矢量序列中提取固定长度的特征矢量；然后馈入后一级 BP 网络完成语音识别任务．相应的语音识别系统结构如图 2 所示．

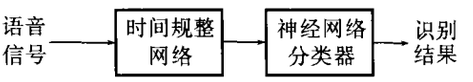

图2 识别系统框图

时间规整网络的另一个重要作用就是可大幅度降低特征矢量维数，从而减小后端识别神经网络输入层的节点数．这种网络输入矢量的降维处理是靠合并若干语音帧来实现的，并非依靠减少语音的特征参数，因此这种降维方法对语音特征描述所造成的损失较小．

2.2 时间规整网络结构

时间规整网络的结构如图 3 所示．若输入汉语语音词组信号的帧数为 n，而要求将其规整为 N 帧，则网络的输入层有 n 个节点，每个节点均有一个与之联系的输入矢量 $A_k^0(k=1,2,\cdots,n)$，A_k^0 是第 k 帧语音信号的特征矢量．进入下一层后，将欧氏距离最近的两相邻矢量加权平均合并，其余矢量不变，这样，下一层就具有 $n-1$ 个节点以及与之联系的 $n-1$ 个输出矢量 $A_k^1(k=1,2,\cdots,$

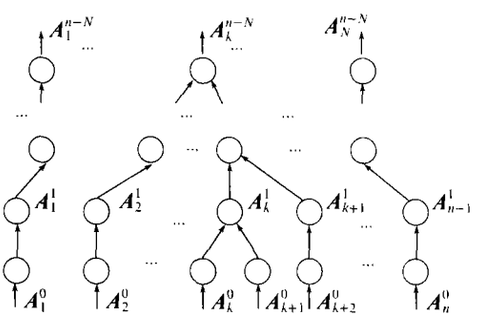

图3 时间规整网络的结构

$n-1)$，依此类推．在经过 $n-N$ 步合并后，最终网络的输出层具有 N 个节点以及与之联系的 N 个矢量 $A_k^{n-N}(k=1,2,\cdots,N)$．

2.3 时间规整网络算法

时间规整网络对特征矢量序列的聚类合并过程，从整体上看又是对输入语音信号的一个分段过程．具体的时间规整算法如下：设 $A_1^0, A_2^0, \cdots, A_n^0$ 是输入语音信号的特征矢量，以 m_k^i 表示矢量 A_k^i 所代表的语音帧数．其中 $i=0,1,\cdots,n-N$；$k=1,2,\cdots,n-i$．特别地，当 $i=0$ 时有 $m_k^i=1(k=1,2,\cdots,n)$．并以 d_k^i 表示矢量 A_k^i 与 A_{k+1}^i 之间的欧氏距离，即

$$d_k^i = \| A_k^i - A_{k+1}^i \| \tag{2}$$

从 $i=0$ 开始重复下述过程直至 $i=n-N-1$．

1) 计算 $d_k^i(k=1,2,\cdots,m-i-1)$，并找出 j，使得 $d_j^i < d_k^i$ 对所有的 $k \neq j$ 成立．

2) 计算 $A_k^{i+1}(k=1,2,\cdots,n-i-1)$

$$\begin{cases} A_k^{i+1} = A_k^i & k<j \\ A_k^{i+1} = \dfrac{m_k^i A_k^i + m_{k+1}^i A_{k+1}^i}{m_k^i + m_{k+1}^i} & k=j \\ A_k^{i+1} = A_{k+1}^i & k>j \end{cases} \tag{3}$$

3) 计算 $m_k^{i+1}(k=1,2,\cdots,n-i-1)$

$$\begin{cases} m_k^{i+1} = m_k^i & k<j \\ m_k^{i+1} = m_k^i + m_{k+1}^i & k=j \\ m_k^{i+1} = m_{k+1}^i & k>j \end{cases} \tag{4}$$

3 实验结果

为了验证本文提出的语音识别方法，共进行了两组实验：一组实验是采用经典的 BP 网络对经过线性时间规整的汉语孤立词特征进行识别，另一组实验是采用时间规整网络与 BP 网络组成的混合级联神经网络语音识别系统对汉语孤立词特征进行识别．实验所用语音数据的采样频率是 8 kHz，量化级是 256；词表中汉语孤立词共有 40 个；讲话者包括 5 个男性和 5 个女性，每人每个单词发音 3 遍．共采得 1 200 个语音样本，将每人的前两次发音共 800 个样本组成训练集，其余 400 个样本组成测试集．

在第 1 组实验中，经过端点检测后长度不等的语音信号，通过线性时间规整后进行分帧，每个孤立词分为 32 帧，每帧语音信号提取 10 阶加权线性预测倒谱系数作为特征参数，因此一个孤立词对应的特征矢量的维数为：$32\times10=320$ 维．用于语音识别的 BP 神经网络结构参数如下：输入层节点数为 320，隐层节点数为 15，输出层节点数为 40(词表中孤立词个数)．网络经过训练后，便可用于对测试集中汉语孤立词特征进行识别，实验结果如表 1 所示．实验结果表明，将 BP 网络用于小词表孤立词语音识别，可以得到较高的正确识别率．但由于每个汉语孤立词需要 30 多个语音帧来描述，所以特征矢量维数非常大，致使网络结构复杂，运算速度很慢；由于网络结构固定，识别前必须对语音信号进行时间规整．

在第 2 组实验中，经过端点检测后长度不等的语音信号，直接进行分帧处理，同样每帧语音信号提取 10 阶加权线性预测倒谱系数组成特征矢量，然后送入时间规整网络，通过对特征矢量聚类合并完成时间规整或分段处理．分段数的选择对后端识别神经网络的性能有很大影响，研究表明，段数的选取与词汇表中的单词有关，若词表中词汇所含音素都比较少，且词汇间差异较大，则只需较小的分段数(4～6)即可；若词表中词汇所含音素较多，且有些单词比较接近，则分段数需取较大的值．由于本文所用孤立词表不大，并且词表中的词汇也没有太多易混淆词，分段数选为 6，即将每个孤立词对应的特征矢量的维数变为 $6\times10=60$ 维．相应地，后端用于识别的 BP 网络结构变为：输入层节点数为 60，隐层节点数为 10，输出层节点数为 40．可见，由于前端采用了时间规整神经网络，致使后端识别网络结构大为简化．将训练样本集和测试样本集分别通过时间规整网络进行处理，然后对后端 BP 网络进行训练和测试，实验结果如表 2 所示．

表1 BP神经网络识别结果

词组数	测试样本数	正确识别词组数	正确识别率	训练时间/s
40	400	378	94.5%	228

表2 混合级联神经网络识别结果

词组数	测试样本数	正确识别词组数	正确识别率	训练时间/s
40	400	393	98.25%	76

4 结 论

采用时间规整网络级联BP神经网络分类器构成的语音识别系统，综合了传统识别方法的时间规整思想，很好地解决了将神经网络应用于语音识别时所遇到的时间对准困难，即神经网络要求输入矢量维数固定不变的问题，同时很好地利用了神经网络用于模式识别的各种优越性能．实验结果表明，在对小词表中文孤立词识别中，可获得98.25%的正确识别率，训练时间仅需76 s，可见，无论从正确识别率还是从训练速度来看，均比单采用BP网络的识别系统性能优越；当词表很大，且其中有很多的易混淆词汇时，需要对该方法及系统性能作进一步的研究．

参考文献：

[1] ZHU S, CHEN D, HUANG T. Feature parameter curve method for high performance NN based speech recognition[A]. Proceedings of ICASSP[C]. New York: IEEE Press, 1996. 1-4.

[2] 胡光锐，周浩，严永红. MHMM和ANN法结合的语音识别方法[J]. 应用科学学报，1995，13(3)：314-318.

[3] LIPPMANN R P. Review of neural networks for speech recognition[J]. Neural Computation, 1989, 1(1): 1-38.

[4] MORGAN N, BOURLAND H A. Neural networks for statistical recognition of continuous speech[J]. Proceeding of IEEE, 1995, 83(5): 742-770.

[5] 顾明亮，王太君，何振亚. 语音信号时间动态规正新方法[J]. 东南大学学报，1998，28(2)：10-14

[6] 史笑兴，顾明亮，王太君，等. 一种时间规整算法在神经网络语音识别中的应用[J]. 东南大学学报(自然科学版)，1999，29(5)：47-51.

Neural Networks Based Phonetic Recognition of Isolated Chinese Phrase

SUN Guang-min, DONG Xiao-ying

(College of Electronic Information & Control Engineering, Beijing Polytechnic University, Beijing 100022, China)

Abstract: The technique of phonetic recognition of isolated Chinese phrase is studied. By combining the time alignment algorithm with neural network technique, a phonetic recognition system based on mixed cascade neural networks is established. The system is composed of two different neural networks. The former is a time alignment network used to solve the time alignment problem and to extract fixed dimension feature vectors from input speech signal. And the latter is a BP network used as the neural network classifier. In the experiment of phonetic recognition for isolated Chinese phrase, the correct recognition rate of this method is up to 98.25%. The experimental results demonstrate that the method can not only solve commendably the time alignment problem by using a neural network but also improve obviously recognition performance. Compared with other ANN based phonetic recognition systems by using the linear time alignment method, both the correct recognition rate and the training speed of the proposed method are improved.

Key words: phonetic recognition; artificial neural network; time alignment