·计算机工程与应用·

*K*子空间和时延自相关器的英汉音素识别

罗万伯[1]，罗霄岚[2]，陈 炜[2]，彭 舰[1]，吴端培[2]

(1. 四川大学计算机学院 成都 610064; 2. 美国克莱姆森大学电子与计算机工程学院, SC29634)

【摘要】提出了用于音素识别的*K*子空间和时延自相关器神经网络结构，用将时延设计加入线性自相关器，以扩展音素滤波神经网络的方法，产生*p*维子空间，并采用迭代过程修改划分，以便捕获语音信号中的时间序列信息。这种带不分类训练过程的体系结构提供了一种高识别性能的方法，没有大多数常规语音识别神经网络所常有的网络输出值不表示候选者似然性的缺陷。通过英语音素和汉语音素的初步试验，识别正确率为84.38%，比音素滤波神经网络方法好。

关 键 词 语音识别; 音素识别; 神经网络; 汉语音素; 时延自相关

中图分类号 TP 399; TN 912 文献标识码 A

English and Chinese Phonemes Recognition Using *K*-Subspaces and Time-Delay Auto-Associators

LUO Wan-bo[1]，LUO Xiao-lan[2]，CHEN Wei[2]，PENG Jian[1]，WU Duan-pei[2]

(1. Computer Science College, Sichuan University Chengdu 610064;

2. Electrical and Computer Engineering College, Clemson University, SC29634 US)

Abstract A neural network architecture, *K*-subspaces and time-delay auto-associators, is proposed for phoneme recognition. It extends the phoneme filter neural networks approach by adding linear auto-associators to create p-dimension subspace, and an iteration is employed to improve the decision. It is good to capture the time-sequence information in speech signal. The architecture proposed could provide a high recognition performance without traditional neural network's shortcoming. Some recognition simulations for both English and Chinese phonemes are conducted, and the recognition rate is 84.38% which is better than phoneme filter neural networks approach.

Key words speech recognition; phoneme recognition; neural network; Chinese phoneme; time-delay auto-associators

许多研究者研究了主元分析(Principal Component Analysis，PCA)的神经网络实现，并应用于数据压缩和特征抽取[1]。文献[2]指出自相关器与PCA非常相近；具有*p*个隐蔽单元的线性自相关器没有最小平方误差意义下的局部最小值，但存在唯一的全局最小值。在全局最小值点，具有*p*个隐蔽单元的线性自相关器形成一个*K*维子空间，它由与训练数据相关的协方差矩阵前面的*p*个主元本征矢量产生。对于子空间，新输入模式的失真由输入与子空间的距离决定，为该空间的生成问题提供了一个精确的定量答案。为了更好地适应多种模式的数据，文献[3]提出了一种称为向量量化PCA(Vector Quantization，VQPCA)的算法，训练数据首先被聚类进*N*个Voronois元，然后对每一个元形成一个局部PCA。

基于3个隐蔽层的非线性自相关器，文献[4]提出了一种用于音素识别的Nakamura法，称为音素滤波神经网络方法。该方法的基础是自相关器能从统计角度特征化训练数据并形成一个好的生成。在该方法中，每一个音素类与一个且只与一个自相关器相关联。每一个自相关器独立地用属于该音素类的语音数据训练。使用该方法进行音素识别的优点是：(1) 不像大多数传统的语音识别神经网络，该方法的网络输出值反映了候选者的似然性；(2) 当一个新音素类添加入准备分类时，已被训练的网络不需重训练。当然，这种网络缺乏捕获语音信号中固有的时间序列信息的能力。仿真表明，对于辅音识别，特别是像/b/，/d/和/g/之类的浊

收稿日期：2003 – 06 – 25

作者简介：罗万伯(1946 –)，男，硕士，教授，主要从事信息安全、多媒体技术和计算机仿真方面的研究.

辅音，识别精度相当差(见表1)。

本文提出了一种新的高性能的神经网络方法，该方法将时延设计引入线性自相关器，以扩展Nakamura法，捕获语音信号中的时间序列信息，并且让每一个音素类都使用*K*个自相关器，以便处理语音数据。

1　时延自相关器

图1是单隐蔽层时延自相关器(Time Delay Auto-Association，TDAA)的示意图。图中，虚线箭头线表示信号流方向，实线箭头线表示带权因子的连接。输入层定义一个*d*帧窗口，给出一个任意时刻*t*的*n*×*d*维输入向量，称为加窗向量。图1中，*n* = 3。在训练和识别两个阶段，语音数据以一个时刻一帧的方式，一步一步地进入输入层。

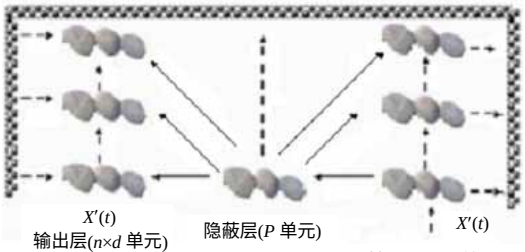

图1　时延自相关器结构

对线性单元用于隐蔽层和输出层，虽然一些研究者指出有3个隐蔽层的非线性自相关器可能比线性自相关器好，但未得到理论或实验证明。非线性自相关器存在局部最小值。当训练过程达到此值时，不能保证非线性自相关器仍比线性自相关器好。非线性相关器的另一缺点是收敛速率低，因为只存在唯一的全局最小值而不存在局部最小值[2]。训练线性自相关器的收敛速度通常比较快。使用Karhuman-Loève变换进行分析，可能得到权因子矩阵。所以，本文提出的方法采用线性自相关器。

用自相关器重构一个加窗向量的失真(distortion)定义为：

$$ds(a,t)=\sum_{i=0}^{d-1}\sum_{j=1}^{n}(X_j(t-i)-X_j'(a,t-i))^2 \tag{1}$$

式中　$X_j(t)$是*n*维输入向量的第*j*元素在时刻*t*的值；$X_j'(a,t)$是用自相关器*a*重构的$X_j(t)$；*d*是时间延迟数。加窗向量失真在训练和测试两个过程中都要被用到。

2　用于音素识别的带时延自相关器的*K*子空间

用于与讲话者无关的音素识别的语音数据数量应当足够大，以便能覆盖音素在各种语句环境的变化。因此，通常需要大量讲话者来生成语音数据。每一个讲话者的语句有不同的统计特性，只使用一个自相关器来刻画用于与讲话者无关的音素识别的语音数据的大量变异是不行的。

为了进一步改进用于音素识别的TDAA的性能，对每一个音素类都需构建*K*个自相关器，并用一种聚类过程对数据进行划分，聚类过程与*K*均值算法*K*元素均值类似，差别在于聚类器的每一数据均值用对应的带*p*个隐蔽单元的自相关器产生的*p*维子空间代替。训练过程不像TDAA或容忍移位的线性向量量化(Linear Vector Quantization，LVQ)[5]，而是采用一种无分类训练方式，因此每一个音素类可以单独训练。

本文提出的这种*K*子空间聚类过程可以视为是文献[3]中的VQPCA的扩展，采用误差后向传播算法训练的线性自相关器产生*p*维子空间，并采用迭代过程修改划分代替VQPCA中的一遍训练方案。*K*子空间聚类过程如下：

步骤 1　初始化步骤

(1) 用*K*均值算法将以加窗向量形式给出的每个音素类的训练数据集划分成*K*的聚类。文献[6]提出的修正的*K*均值算法被用于以欧几里德失真量度的仿真中。

(2) 建立*K*个线性自相关器，每个聚类器一个；用一个小的随机数初始化这些自相关器，并将每个自相关器的输出层偏差设置为相应聚类的数据均值[2]。

步骤 2　迭代步骤

(1) 使用相应聚类的数据，直到采用误差后向传播算法训练的每个线性自相关器收敛为止。候选的方法则是用Karhuman-Loève变换。在此训练期间，输出层偏差保持不变。

(2) 按式(1)给出的失真在整个训练集上修正划分，并计算每个聚类的新的数据均值。

(3) 如果在迭代步骤的(2)里划分没有变化，则停止迭代，训练过程完成；否则，继续迭代步骤(4)。

(4) 用相应聚类的新数据均值修改每个自相关器的输出偏差，然后转到迭代步骤(1)。

使用Karhuman-Loève变换，可保证以上的过程收敛，主要基于两点事实：(1)迭代步骤(2)的修正使整个训练集上的总失真总是减少或维持不变；(2)在每个聚类中，使用Karhuman-Loève变换，整个训练集上的总失真被最小化。因此，对任何新划分，每个聚类的总失真总是比用自相关器的旧的权矩阵计算出来的初始失真减少或维持不变。应该注意的是，迭代步骤(4)被用来改进神经网络训练。因为训练数据均值是找到最佳输出偏差向量的关键[6]，迭代步骤(4)将不会影响每个聚类的总失真的最小值。对于误差后向传播来说，冲量α 和学习率ε 需要仔细调整，以避免凹点。在本文的仿真里，选择α =0.9，并让ε 为迭代次数n的倒数(1/n)的函数。

3 音素识别仿真

本文进行了4种音素识别仿真。

第1种：仿真是TDAA英语音素识别，用于评价TDAA对英语音素/b/，/d/和/g/的识别性能。音素标记从TIMIT语音数据库中抽取，参与神经网络训练的是120个讲话者提供的623个音素标记，用另外120个讲话者提供的621个音素标记来测试。数据经20个通道的基于短时谱的滤波器组处理，频率范围从20 Hz到7 000 Hz，每10 ms一个抽样帧。

每一种音素都构造了3个TDAA，每一个TDAA具有相同的维数即(20×20)×8×(20×7)单元，并分别用对应的音素，采用标准误差向后传播学习进行训练，直到失真不再减少为止。

对于识别过程，仍采用文献[7]中提出的方案。特别是在时刻t，类别c的激活性由下式给出，即：

$$A(c,t)=1-\frac{ds(c,t)}{\sum_i ds(i,t)} \tag{2}$$

对每一个音素类来说，在音素的整个时段对每一时刻得到的激活性求和，具有最高总激活性的类被选为识别出的音素类。

表1列出了仿真的结果。为了方便比较，用Nakamera法中的带3个隐蔽层的非线性自相关器方法的仿真结果也列在了表1中。Nakamera法中的带3个隐蔽层的非线性自相关器不带时间延迟，结构维数是20×32×8×32×20单元。

表1 使用**TDAA**和**Nakamera**法的仿真结果

音 素	音素标记数		TDAA 识别率		Nakamera法识别率	
	训 练	测 试	训 练/ (%)	测 试/ (%)	训 练/ (%)	测 试/ (%)
/b/	221	244	87.78	84.02	71.95	58.19
/d/	263	248	87.45	84.05	71.10	59.27
/g/	139	129	82.01	73.64	65.47	60.47
总 计	623	621	86.36	80.68	70.14	59.10

从表1可以清楚地看出，TDAA法得到的识别精度远高于Nakamera法的不带时间延迟的方法，意味着结合时间延迟单元可以提高自相关器的音素识别性能。

第2种：仿真是用KS-TDAA的英语音素/b/，/d/和/g/识别，以便与TDAA对比识别性能。对于每个音素类，构建3个自相关器，每个自相关器的维数为(20×7)×8×(20×7)个单元，并用在第1种仿真中叙述的同样的数据和特征表示，采用误差后向传播算法训练这些自相关器。表2列出了用KS-TDAA的音素识别仿真结果。作为对照，采用文献[7]给出的结构的时延神经网络(Neural Networks Time Delay，TDNN)的识别结果也列在表2中。从表2可以看出，在小数据量集的情况下，KS-TDAA的识别结果比TDNN好。从表2还可看出，用KS-TDAA对这3个音素识别的总正确率为84.38%，比表1中的两种方法的结果都好。

第3种：仿真用大数据集进行。为评价KS-TDAA，仍选英语音素/b/，/d/和/g/，从TIMIT语音数据库中抽取420个讲话者的2 368个语音标记作为训练用数据，并抽取此420讲话者之外的另外210个讲话者的863个语音标记数据作为测试用数据，都使用大数据量集进行一系列仿真。识别的结果在表3中列出。

第4种：仿真是使用KS-TDAA的汉语音素/b/，/d/和/g/的识别。由于条件的限制，汉语音素识别只进行

了小数据量识别实验，参与神经网络训练的是5个讲话者提供的620个音素标记，用于测试的是另外5个讲话者提供的620个音素标记。同英语音素一样，这些数据也是经20个通道的基于短时谱的滤波器组处理，频率范围从20 Hz到7 000 Hz，每10 ms一个抽样帧。表4列出了汉语音素的识别结果。对比表2可以看出，汉语音素/b/，/d/和/g/的识别效果与英语的相应音素的识别结果没有多大差别，这可以用这两种语言的/B/，/D/和/G/3个音素非常相近来解释。

表2　使用**KS-TDAA**和**TDNN**的仿真结果

音素	音素标记数		KS-TDAA识别率		TDNN识别率	
	训　练	测　试	训 练/ (%)	测 试/ (%)	训 练/ (%)	测 试/ (%)
/b/	221	244	97.74	90.16	89.95	82.79
/d/	263	248	96.20	84.27	98.13	86.69
/g/	139	129	93.53	73.64	91.97	72.09
总计	623	621	96.15	84.38	90.93	82.13

表3　使用**KS-TDAA**的大数据集的仿真结果

音素	音素标记数		KS-TDAA识别率	
	训　练	测　试	训 练/ (%)	测 试/ (%)
/b/	844	342	93.13	91.81
/d/	1 017	341	88.10	84.75
/g/	507	180	91.12	83.89
总计	2 368	863	90.54	87.37

表4　使用**KS-TDAA**汉语音素识别的仿真结果

音素	音素标记数		KS-TDAA识别率	
	训　练	测　试	训 练/ (%)	测 试/ (%)
/b/	210	210	95.24	90.48
/d/	210	210	96.19	88.10
/g/	200	200	95.50	83.50
总计	620	620	95.64	87.42

4　结　论

本文描述了一种用于音素识别的高性能神经网络结构，即*K*子空间和时延自相关器。这种结构组合了用于音素识别的时延设计[8-9]和MLP自相关器技术[4]。对于每一种音素，*K*个时延线性自相关器使用属于该类音素的语音数据，按照本文提出的*K*子空间聚类过程进行构建和训练，过程与*K*均值算法非常类似。这种带不分类训练过程的体系结构提供了一种高识别性能的方法，且没有大多数常规语音识别神经网络所常有的网络输出值不表示候选者似然性的缺陷。

对于英语音素，用从TIMIT数据库中抽取出来的/b/，/d/和/g/3种音素试验，获得的识别率为84.38%。这个结果比Nakamura法采用的神经网络对同一任务所得的识别率59.10%高[4]。利用大数据集作相同试验，本结构的识别率为87.37%。

对于汉语音素，用自己准备的语音数据对“B”，“D”和“G”3种音素试验，识别率与英语音素相当。

参　考　文　献

[1] Haykin S. Neural networks –a comprehensive foundation[M]. New York: Macmillan College Publishing Company, 1994, 11-15.
[2] Balid P, Hornik K. Neural networks and principal component analysis: learning from examples without local minima[J]. Neural Networks, 1989, 2(3): 53-58.
[3] Kambhatla N, Leen T K. Fast non-linear dimension reduction[C]. ICNN, San Francisco, 1993, 3:1213-1218.
[4] Masami Nakamura. Phoneme recognition by phoneme filter neural networks[C]. Proc. IEEE ICASSP, Ontario Canada, 1991: 85-88.
[5] Bourlard H A, Kamp Y. Auto-association by multilayer and singular value decomposition[J]. Biological Cybernetics, 1988, 59(2): 291-294.
[6] Wilpon J G, Rabiner L R. A modified K-means clustering algorithm for use in isolated word recognition[J]. IEEE Trans. ASSP, 1985, 33(3):587-594.
[7] Leen T K, Rudnick M, Hammerstron D. Hebbian feature discovery improves classifier efficiency[C]. IJCNN, San Diego, 1990, 51-56.
[8] Waibel A. Phoneme recognition using time-delay neural networks[J]. IEEE Trans. on ASSP, 1989, 37(3):58-62.
[9] McDermott E. Shift-tolerant LVQ and hybrid LVQ-HMM for phoneme recognition[C]. Alex Waibel and Kai-Fu Lee, Reading in Speech Recognition, San Mateo: Morgan Kaufmann Publishers, 1990, 150-154.

编　辑　熊思亮