优先度排序 RBF 神经网络在与文本无关说话人确认中的应用 [1]

邓浩江 王守觉* 杜利民

(中国科学院声学研究所语音交互技术研究中心 北京 100080)
*(中国科学院半导体研究所人工神经网络实验室 北京 100083)

摘 要 该文介绍了优先度排序径向基函数 (PORBF) 神经网络的结构与算法，并提出了将其应用于与文本无关说话人确认时的训练算法、似然度的计算方法以及识别规则。为了增强 PORBF 网络的泛化能力，该文用压缩矢量构造抑制样本集，提出了顺序选取、最近邻选取和最远距离选取等 3 种选择抑制样本集中说话人的方法，并对 PORBF 神经元的输出进行了等比递减加权。在相同条件下的与文本无关说话人确认实验中，传统的矢量量化方法的等差错率可达 10.56%，而基于 PORBF 网络的确认系统使用最近邻选择方法构造抑制样本集，其等差错率可达 6.83%，性能提高很多。

关键词 优先度排序，说话人确认，与文本无关， RBF 网络
中图号 TN912.3, TN-052

1 引 言

说话人确认是由说话人的一段语音判断其是否为所声言人的一种基于生物统计学的自动识别系统，属于说话人识别的范畴。说话人确认系统通常为每一位用户构造一个模型或分类器，确认时从语音信号中提取特征矢量序列输入到所声言人的模型或分类器中，得到一组表示匹配程度的数字，经过处理就可以得到一个表示该用户与所声言人相似程度的值 (似然值， Likelihood score)，然后根据似然值是否超过设定的阈值接受或拒绝该用户。

说话人确认模型的构造有两种方法：一种是无监督的学习方法，包括矢量量化 [1,2](VQ, Vector Quantization) 、混合高斯模型 [3](GMM, Gaussian Mixture Model) 、隐马尔可夫模型 [4](HMM, Hidden Markov Model) 等，这些模型的训练基本上都是建立在说话人自己的语音数据基础上；另一种是有监督的学习方法，每个说话人模型的训练样本不仅包括自己的特征矢量，还包括全部或部分其他人的特征矢量，其中人工神经网络是主要的一种方法。神经网络由大量简单处理单元 (神经元) 并行连接而成，能够自适应地学习输入空间到输出空间的复杂映射关系。对于那些基本统计特性了解得还不是很透彻的处理对象 (如语音信号) 来说，神经网络是一个很有用的工具。人工神经网络的许多模型，包括多层感知器 [5](MLP, Multi-Layer Perceptron) 、径向基函数 (RBF, Radial Basis Function) 网络 [6] 、预测神经网络 [7](PNN, Predictive Neural Networks) 和神经树网络 (Neural Tree Networks)[1,5] 等都在说话人确认中得到了广泛的应用。根据在训练和测试过程中的说话内容是否来源于相同的词汇表，说话人确认系统可以分为与文本有关 (Text-dependent) 和与文本无关 (Text-independent) 两种实现方式，文献 [2] 提出的基于神经计算机的说话人确认系统是与文本有关的，而本文研究的对象是与文本无关的。

2 优先度排序 RBF 神经网络

通用 RBF 网络 [8] 是一种三层前馈网络，隐层是 RBF 神经元，输出层是线性神经元。RBF 神经元先计算输入矢量与其中心之间的距离，然后再进行非线性变换，其传递函数为

$$h_i = \phi(\|\boldsymbol{X} - \boldsymbol{C}_i\|/\sigma_i) \tag{1}$$

[1] 2002-04-11 收到， 2002-09-16 改回

其中 $\boldsymbol{X}$ 是输入矢量，h_i, C_i, σ_i 分别是第 i 个 RBF 神经元的输出、中心矢量和半径。非线性函数 $\phi(\cdot)$ 可以有多种形式，最常用的是高斯函数 $\phi(x)=\exp(-x^2)$。从 RBF 神经元的传递函数可以看出，只有当输入矢量落在输入空间距神经元中心相近的局部区域时，RBF 神经元才产生非零响应，也就是说具有局部敏感性。因此 RBF 网络有时也称为局部感受野网络 [9]。

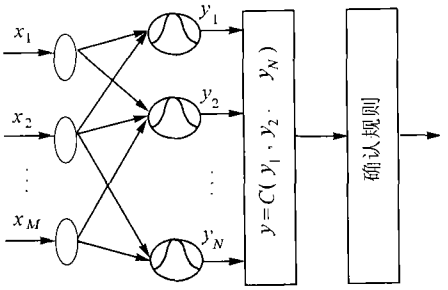

图 1 用于说话人确认的 PORBF 网络

优先度排序神经网络 (PONN, Priority Ordered Neural Network)[10] 由若干具有不同优先级的神经元或神经网络子模块组成，不同的神经元或子模块可以具有不同的结构。我们将优先度排序原理应用于 RBF 网络就构成了用于说话人确认的 PORBF 网络 (图 1)，每一个 RBF 神经元输出为 $y_n \in R^{N_n}(n=1,2,\cdots,N)$，其优先度为 p_n，整个网络的输出可以用下面的公式来描述：

$$\left.\begin{aligned} y &= C(y_1, y_2, \cdots, y_N) = y_s \\ s &= \min\{i | p_i = \max\{p_j | Q(y_j)=1\}\} \end{aligned}\right\} \tag{2}$$

其中 $C(y_1, y_2, \cdots, y_N)$ 是网络的决策函数，Q 是网络的条件映射 [10]，可以描述为

$$Q : \bigcup_{j=1}^{\infty} R^{N_j} \to \{0,1\} \tag{3}$$

在 PORBF 网络中，N_j 是每个 RBF 神经元输入矢量空间的维数。PORBF 网络的决策是在符合条件的 RBF 神经元中选择优先度最大的神经元输出作为最终输出。

3 训练与识别算法

3.1 PORBF 网络的训练

RBF 网络的训练通常分为两步：先确定隐层 RBF 神经元的中心和半径，再确定隐层到输出层的权值。中心的确定既可以从训练集的样本矢量中选取，也可以采用无监督的聚类方法来获取，而隐层到输出层权值的调整一般采用有监督的学习方法，如最小均方算法 (LMS, Least Mean Squares)。

与 RBF 网络的学习过程相比，PORBF 网络的训练只需确定 RBF 神经元的中心和半径即可。文献 [11] 提出了优先度排序方向基函数 (DBF, Direction Basic Function) 神经网络的训练算法，PORBF 网络训练的算法流程与之基本相同，但在计算 DBF 距离的地方都改为计算输入样本与候选中心的欧氏距离。该算法主要是重复进行下面的迭代过程，直至所有的样本都划分完毕，先训练出来的神经元具有较高的优先度：

(1) 选取神经元 n 的中心 $\boldsymbol{C}_n$，标记该神经元的所属类别；

(2) 计算半径 r_n；

(3) 删除以 $\boldsymbol{C}_n$ 为球心、以 r_n 为半径的超球体内与神经元 n 相同类别的样本，构造新的样本集。

中心的选取方法有从样本中顺序选取、随机选取以及遍历样本选取 [11] 等，其中遍历样本选取中心是以包含同类样本最多为原则，在一定程度上反映了样本集在特征空间的分布特征，

有利于提取与说话人有关的特征。用该方法训练得到的网络规模较为紧凑，泛化能力较强。可以看出，PORBF 网络的结构，即 PORBF 神经元的数目及参数是在训练中确定下来的，具有自组织、自适应的特点。

3.2 样本集的制作

PORBF 网络用于说话人确认系统时，可以用全体用户的训练矢量组成原始样本集，然后将某一说话人的特征矢量的类别属性标记为 1，将其他说话人特征矢量的类别属性标记为 0，训练该说话人的 PORBF 网络，用相同的方法训练出所有用户的 PORBF 网络。遍历样本选取中心对于较多用户的网络训练就存在一个问题，只有少数训练矢量的类别属性为 1，大部分训练矢量的类别属性为 0，这就使得类别属性为 0 的神经元序号较小，优先度较高，因此网络倾向于做出否定的判决输出。这就存在一个样本平衡的问题，而解决的途径就是使类别属性为 0 和 1 的样本数大致相同。平衡样本结构的方法有两种：一种是加权说话人 j 的特征矢量，如噪声加权，以补偿激活样本的不足，但却扩大了样本集的规模，增加了 PORBF 网络的训练时间；一种是减少抑制样本的数量，其中既可以从抑制样本中抽取部分矢量组成新的样本集，也可以对原始矢量进行压缩。

针对语音信号特征矢量序列随时间变化的特性，我们采用一种时间轴上的动态平滑压缩方法，将选入抑制样本集的每位用户的特征矢量分别按一定比率压缩，组成与激活样本集规模相同的抑制样本集，压缩比 γ 等于选入抑制样本集的用户数。压缩方法如下：

将需要压缩的语音特征矢量序列 $x_i(i=1,2,\cdots,I)$ 按时间先后顺序随机或均匀分成 J 段 $T_1,T_2,\cdots,T_J$，其中 $J=I/\gamma$。计算每段的中心矢量 $\boldsymbol{C}_j=\frac{1}{N_j}\sum_{\boldsymbol{x}_i\in T_j}\boldsymbol{x}_i$(式中 N_j 是划分到第 j 段的特征矢量数)，并将每段的起始和终止矢量标记为 $\boldsymbol{x}_{js}$，$\boldsymbol{x}_{je}(j=1,2,\cdots,J)$。

然后按相邻最近邻原则重复下面的迭代过程，重新划分 $\boldsymbol{x}_{js}$，$\boldsymbol{x}_{je}$ 的归属，直至不再有 $\boldsymbol{x}_{js}$，$\boldsymbol{x}_{je}$ 移动为止：

依次计算 $d_{js}(\boldsymbol{x}_{js},\boldsymbol{C}_j)$ 和 $d_{(j-1)s}(\boldsymbol{x}_{js},\boldsymbol{C}_{j-1})$，$j=2,\cdots,J$，如果 $d_{(j-1)s}<d_{js}$，将 $\boldsymbol{x}_{js}$ 由第 j 段移入第 $j-1$ 段，重新标记 $\boldsymbol{x}_{js}$，$\boldsymbol{x}_{(j-1)e}$，并重新计算第 $j-1$，j 段的中心矢量 $\boldsymbol{C}_{j-1},\boldsymbol{C}_j$，$d(\cdot)$ 表示特征矢量到中心矢量的距离。

依次计算 $d_{je}(\boldsymbol{x}_{je},\boldsymbol{C}_j)$ 和 $d_{(j-1)s}(\boldsymbol{x}_{je},\boldsymbol{C}_{j+1})$，$j=1,\cdots,J-1$，如果 $d_{(j+1)e}<d_{je}$，将 $\boldsymbol{x}_{je}$ 由第 j 段移入第 $j+1$ 段，重新标记 $\boldsymbol{x}_{je}$，$\boldsymbol{x}_{(j+1)s}$，并重新计算第 j，$j+1$ 段的中心矢量 $\boldsymbol{C}_j,\boldsymbol{C}_{j+1}$。

最后得到的 J 个中心矢量 $\boldsymbol{C}_j$ 就是比率为 $1:\gamma$ 的压缩样本。

3.3 说话人确认系统的训练与判决规则

设系统的用户总数为 N，抑制样本集所包含说话人的数目为 $M(<N)$，当训练用户 j 的网络时，抑制样本集中说话人的选择有 3 种方法：

(1) 顺序选取，即从 N 个用户中顺次选取 M 个非 j 的说话人；

(2) 最近邻选取，当 $j\le M+1$ 时仍然采用顺序选取的方法，当 $j>M+1$ 时，将用户 j 的训练语句依次输入到前 $j-1$ 个用户已训练好的网络中去，按似然度从大到小排序，选取前 M 个说话人制作抑制样本集，似然度的计算将在下面介绍；

(3) 最远距离选取，方法同 (2)，但选择似然度最小的前 M 个说话人。

我们把待测语音序列与 PORBF 网络所代表的说话人之间的相似程度定义为似然度。进行说话人确认时，PORBF 网络的非线性函数 $\phi(\cdot)$ 为硬限幅函数，将待测语音特征矢量序列 $\boldsymbol{x}^{(t)}$ 依次输入声言人的 PORBF 网络模型中，得到输出序列 $y_{h^{(t)}}(t=1,\cdots,T)$，$h^{(t)}$ 是符合条件的 PORBF 神经元中优先度最大的神经元序号。该段语音对于第 j 个网络模型的似然度由下式计

算:

$$S_j = \sum_{t,l_{h(t)}=1} y_{h(t)} \Big/ \sum_{t,l_{h(t)}=0} y_{h(t)} \tag{4}$$

其中 $l_{h(t)}$ 是第 $h^{(t)}$ 个隐层神经元的类别属性，而神经元的类别属性在训练时就确定下来了 (3.1 节)。为了避免计算时溢出和便于比较，我们在实际系统中采用了对数似然度进行计算:

$$S_j^L = \log\left(\sum_{t,l_{h(t)}=1} y_{h(t)}\right) - \log\left(\sum_{t,l_{h(t)}=0} y_{h(t)}\right) \tag{5}$$

当似然值大于预先设定的阈值时就接受该说话人。

因为 PORBF 网络中各神经元的优先度不一样，序号小的神经元具有较高的优先度，所以优先度高的神经元产生错误输出时会影响到优先度低的神经元，也就是说优先度低的神经元产生错误输出的概率大。由 3.1 节的学习算法得到的 PORBF 网络对于训练样本集可以达到 100% 的正确识别率，所以对训练样本集不存在这个问题。但是一个人的语音受时间、环境、身体状况以及说话内容等条件的影响变化比较大，因此将该网络用于实际的与文本无关说话人确认中时容易产生较大的误识率。为此我们对 PORBF 神经元的输出进行加权，即 $\hat{y}_n = y_n \cdot w(n)$, 原则是加权系数随优先度的降低而减小，实验中我们采用了等比递减加权的方法，效果比较好。如此改进后，说话人确认系统的似然值可由下式计算:

$$S_j^{Lw} = \log\left(\sum_{t,l_{h(t)}=1} y_{h(t)} w(h^{(t)})\right) - \log\left(\sum_{t,l_{h(t)}=0} y_{h(t)} w(h^{(t)})\right) \tag{6}$$

其中 $w(h^{(t)}) = (1-\eta)^{h^{(t)}}$, $0 < \eta << 1$ 。

3.2 节采用的动态平滑压缩方法，用压缩矢量代替原始矢量组成抑制样本集，在一定程度上增强了样本的代表能力，虽然降低了训练样本集中抑制样本的正确识别率，但却提高了 PORBF 网络的推广泛化能力。

4 实验与结果

4.1 语音信号及预处理

我们使用的语音数据是在实验室环境下，用优质话筒以 8kHz 频率采样的单声道语音信号，包括 25 名青年男性。每个人在半年时间内分别采集两次，共 20 句，普通话，说话内容不限。其中， 5 句用来训练，其余 15 句用来测试，测试时是以每一句为单位进行的。

实验中我们使用从线性预测 (LP, Linear Predictive) 分析导出的 14 维倒谱系数作为特征矢量， LP 分析采用 30ms 的汉明 (Hamming) 窗，帧移是 10ms 。在提取特征矢量之前，我们首先在短时能量、短时过零率和一阶差分能量的基础上去除语音信号中的静音段和噪声段。

说话人确认系统的性能通常用等差错率 (EER, Equal Error Rate) 来衡量 [12], EER 是后验差错率，此时判决阈值的设定就是使错误接受 (FA, False Accept) 率等于错误拒绝 (FR, False Rejection) 率。下面我们将在 EER 的基础上通过实验研究抑制样本集的组成以及 η((6) 式) 的选择对说话人确认系统性能的影响。

4.2 抑制样本集的组成和 η 值影响识别率的实验

用 $N(=10,15)$ 个人的语音数据做封闭样本 (Closed set) 实验，从每个人的训练语句中取 8s 的语音数据作 LP 谱分析，提取倒谱矢量，按照第 3 节提出的训练识别算法，针对抑制样本集中所包含说话人的数目和选择方法的不同进行实验，测试时分别统计 N 个说话人相互之间的 FA 率和 FR 率，结果如表 1, 图 2 所示. 其中似然值用 (6) 式计算，训练时间是指单个网络的训练时间，使用计算机的 CPU 是 PIII-667 .

表 1　抑制样本不同选择方法与神经元数目、训练时间及 EER 之间的关系 ($N=15, \eta=0.001$)

选择方法	抑制样本集用户数 (M)	PORBF 神经元数目			训练时间 (s)			等差错率 EER(%)
		最多	最少	平均	最长	最短	平均	
顺序选取	5	223	126	174	11.42	6.43	8.98	7.95
	8	215	108	170	11.25	6.09	8.73	6.88
最近邻	5	226	124	184	11.43	7.8	9.79	7.53
	8	215	114	173	11.49	6.87	9.41	6.83
最远距离	5	211	86	161	11.86	5.82	8.62	8.55
	8	215	90	160	11.15	6.21	8.68	7.95

可以看出，PORBF 网络的训练速度比较快，最慢 11.86s, 最快 5.82s, 速度的差异是由激活样本与抑制样本在特征空间的不同分布造成的；3 种选择方法中最近邻选取能获得最佳的识别效果，最远距离选取的训练速度最快；随着选入抑制样本集的说话人数目 M 的增加，EER 不断降低，当人数超过 8 以后，EER 开始回升. 这是因为随着人数的增加，非声言人的测试数据就更接近于抑制样本集的矢量，但是抑制样本集人数与压缩比成正比，压缩比增加到一定程度，就会使动态平滑得到的压缩矢量趋向于均值矢量而失去代表性，最终影响到识别效果，所以选择 8 个人制作抑制样本集是比较合适的.

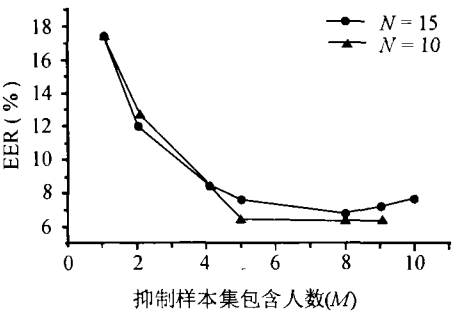

图 2　抑制样本集包含人数与 EER 的关系 (最近邻选取，$\eta=0.001$)

我们采用最近邻方法针对不同 η 值统计了 10 人和 15 人确认系统的 EER, 见表 2 . 可以看出，采用等比递减加权能够降低 EER, $\eta=0.001$ 时效果最好，此时 EER 比不加权 (即 $\eta=0$) 分别降低 6.4%($N=10$) 和 7.95%($N=15$) .

表 2　η 值与 EER 之间的关系 ($M=8$)

η		0	0.0005	0.001	0.002	0.003
EER(%)	$N=10$	6.89	6.53	6.45	6.48	6.50
	$N=15$	7.42	7.22	6.83	6.94	7.01

4.3 与矢量量化方法的比较

我们将基于 PORBF 网络的说话人确认系统与 VQ 方法进行了比较，训练与测试样本与前面实验相同，即提取 8s 语音的特征矢量作为训练样本，用 LBG[13] 算法提取每个说话人的码本，计算测试语音特征矢量序列与所声言人码本之间的失真，并对测试矢量的数目求平均，然后再与一预先设定的阈值进行比较以决定是拒绝还是接受. 我们针对码本容量分别为 32, 64, 128 和 256 的 VQ 方法进行了实验，测试时针对不同判决阈值分别统计 15 人内部相互之间的 FA 率

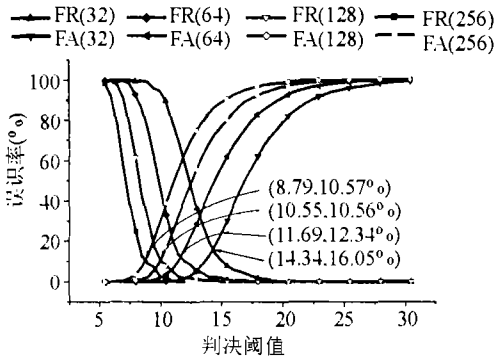

图 3 基于 VQ 说话人确认系统的误识率

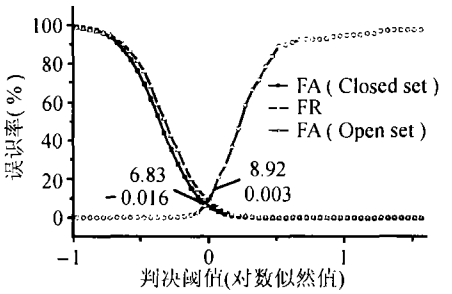

图 4 开放与封闭测试样本集的误识率

和 FR 率，结果如图 3 所示。由图可知，基于 VQ 的说话人确认系统性能随码本容量的增加而增强，当码本容量为 128 时，VQ 方法获得最佳性能，此时 EER 可达 10.56% 。而在相同的条件下的说话人确认实验中，基于 PORBF 网络的确认系统 EER 最低可达 6.83% 。可见，与传统的 VQ 方法相比，PORBF 网络的性能要提高很多。

4.4 开放样本 (Open set) 实验

实验中仍然采用 4.2 节训练得到的 15 人的 PORBF 网络，测试时分别统计 15 人内部相互之间的 FA 率和 FR 率，以及语音数据库中另外 10 人对这 15 人网络的 FA 率，结果如图 4 所示。可以看出，对于开放样本集，EER 阈值由 -0.016 上升到 0.003，而 EER 也由 6.83% 升高到 8.92% 。

5 结 论

本文介绍了 PORBF 神经网络的结构与算法，并将其应用于与文本无关说话人确认系统中。针对说话人确认，我们提出了用压缩矢量构造 PORBF 训练样本集的方法以及顺序选取、最近邻选取和最远距离选取等 3 种选择抑制样本集中说话人的方法。为了减少 PORBF 网络中优先度高的神经元产生错误输出时对优先度低的神经元的影响，我们采用了对 PORBF 神经元的输出进行等比递减加权的方法。通过实验可以得出下面的结论：

PORBF 网络神经元的数目及参数是在训练中确定下来的，不仅具有自组织、自适应的特点，而且训练速度比较快；3 种选择抑制样本集中说话人的方法中最近邻选取获得的识别率最高，随着抑制样本集中说话人数目 M 的增加，EER 先降后升，选择 $M = 8$ 是比较合适的；等比递减因子的采用能够降低 EER，当 $\eta = 0.001$ 时获得最佳效果；开放样本集的 EER 高于封闭样本集；与传统的 VQ 方法相比，PORBF 网络的性能要高很多。

参 考 文 献

[1] K. R. Farrell, S. Kosonocky, R. J. Mammone, Neural tree network/vector quantization probability estimator for speaker recognition, Proc. IEEE, 1994, 82(1), 279–288.

[2] Zhang Yiying, Zhang David, Zhu Xiaoyan, A novel text-independent speaker verification method based on global speaker model, IEEE Trans. on System, Man, and Cybernetics, 2000, 30(5), 598–602.

[3] 史静朴，等，用神经计算机的说话人确认系统及其应用，电子学报，1999, 27(10), 1999, 27–29.

[4] R. A. Sukkar, M. B. Gandhi, A. R. Setlur, Speaker verification using mixture decomposition discrimination, IEEE Trans. on Speech and Audio Processing, 2000, 8(2), 292–299.

[5] K. R. Farrell, R. J. Mammone, K. T.assaleh, Speaker recognition using neural networks and conventional classifiers, IEEE Trans. on Speech and Audio Processing, 1994, 2(1), 194–204.

[6] J. Oglesby, J. S. Mason, Radial basis function networks for speaker recognition, Proc. ICASSP, New Mexico, USA, 1990, 393–396.

[7] R. A. Finan, A. T. Sapeluk, Text-independent speaker verification using predictive neural networks, IEE Conf. Pub. of ANN, Cambridge, UK, July, 1997, 274–279.

[8] D. S. Broomhead, D. Lowe, Multivariable functional interpolation and adaptive networks, Complex Systems, 1988, 2(2), 321–355.

[9] J. Moody, C. J. Darken, Faster learning in networks of locally-tuned processing units, Neural Computation, 1989, 1(2), 281–293.

[10] Wang Shoujue, Priority ordered neural networks with better similarity to human knowledge representation, Chinese Journal of Electronics, 1999, 8(1), 1–4.

[11] 陈川，方向基函数神经网络模型与算法及其在模式识别中的应用，[博士论文]，北京，中国科学院半导体研究所，1999.

[12] Liu Chi-Shi, Wang Hsiao-Chuan, Lee Chin-Hui, Speaker verification using normalized log-likelihood score, IEEE Trans. on Speech and Audio Processing, 1996, 4(1), 56–60.

[13] Y. Linde, A. Buzo, R. M. Gray, An algorithm for vector quantizer design, IEEE Trans. on Commun., 1993, COM-28(1), 84–95.

TEXT-INDEPENDENT SPEAKER VERIFICATION USING PRIORITY ORDERED RADIAL BASIS FUNCTION NETWORKS

Deng Haojiang　Wang Shoujue*　Du Limin

(*SITR, Institute of Acoustics, Chinese Academy of Sciences, Beijing* 100080, *China*)

*(*Lab of Artificial Neural Networks, Institute of Semiconductors, Chinese Academy of Sciences, Beijing* 100083, *China*)

Abstract The structure and algorithm of Priority Ordered Radial Basis Function (PORBF) Networks is introduced. The concrete training algorithm, calculational methods of likelihood score and verification rule, used for text-independent speaker verification, are proposed. To enhance the generalization ability, the compressing vectors are applied to construct the inhibitory samples set and three methods including sequential selection, nearest neighbor selection and furthest distance selection are presented for the choose of anti-speakers. Moreover, the outputs of neurons are weighted by a descendent array. Using these algorithms and methods, the performance is examined by a series of experiments. The results show that under the identical experiment conditions, when the inhibitory set is composed of anti-speakers' compressing vectors selected using nearest neighbor method, the Equal Error Rate (EER) using PORBF networks can decreased to 6.83% from 10.56% using conventional VQ method. For speaker verification, the PORBF network provides better performance than the VQ classifier.

Key words Priority ordered, Speaker verification, Text-independent, Radial Basis Function networks

邓浩江：　男，1971 年生，中科院声学研究所博士后，主要从事语音信号处理、说话人识别以及神经网络信息处理等.

王守觉：　男，1925 年生，研究员，中科院院士，现从事半导体超高速电路与人工神经网络算法、模型、硬件和应用的研究.

杜利民：　男，1957 年生，研究员，主要从事语音信号与信息处理、语音识别和自然语言理解等方面的研究.