基于小波神经网络的与文本无关说话人识别方法研究

白 莹 赵振东 戚银城 王 斌 郭建勇
(华北电力大学电子与通信工程系 保定 071003)

摘 要 基于神经网络的说话人识别方法可以在一定程度上模仿人脑的功能，是说话人识别中的一种主要技术，但它通常难以确定隐层单元的数目，收敛速度慢，易于收敛到极小点。该文研究了一种用于说话人识别的小波神经网络模型，给出了网络结构和学习算法。采用Mel频率倒谱系数作为与文本无关的说话人识别的特征参数，并利用该模型进行了5个人的说话人识别实验，得到99.5%的识别率。实验结果表明，小波网络和传统的BP网络相比，训练速度和识别率都有了较大提高，具有良好的应用前景和进一步研究的价值。

关键词 说话人识别，小波神经网络，BP网络，Mel频率倒谱系数
中图分类号：TN391.42 **文献标识码**：A **文章编号**：1009-5896(2006)06-1036-04

Research on Text-Independent Speaker Recognition Methods Using Wavelet Neural Network

Bai Ying Zhao Zhen-dong Qi Yin-cheng Wang Bin Guo Jian-yong
(*Dept. of Electronic and Communication Engineering, North China Electric Power University, Baoding* 071003, *China*)

Abstract The approach for speaker recognition based on neural networks is able to emulate the function of human brain in some degree, so it is a main implementation technology in the speaker recognition. But it is difficult to determine the number of hidden layer neurons, slowly convergent and easy to fall into local minimum point. The model of wavelet neural networks is studied. The structure of the network and learning algorithm are given. The recognition correctness reaches to 99.5% for 5 speakers using Mel frequency cepstral coefficient as feature parameters. The experimental at results show that the learning rate and recognition correctness are improved much compared to the BP networks. It has a good application prospect and worth to research further more.

Key words Speaker recognition, Wavelet neural network, BP network, Mel frequency cepstral coefficient

1 引言

说话人识别就是用待识别语音和预先提取的说话人特征来鉴别出说话人身份的一种技术，是语音信号处理领域一个十分活跃的研究方向。从本质上讲，说话人识别是语音信号模式识别的问题。根据待识别说话人的语音材料，识别系统可以分为与文本有关的说话人识别(Text-dependent)和与文本无关的说话人识别(Text-independent)。与文本有关的说话人识别系统要求用户按照规定的内容发音，并根据特定的内容建立精确的模型，可以达到较好的识别率，若用户不按规定内容发音，则难以识别。与文本无关的说话人识别系统在训练和识别时的说话内容是任意的，建立精确的模型较为困难，这就要提取只和说话人语音特点有关的特征来建立模型，比较接近实用时的情况。

在说话人识别中，常用的分类技术有模板匹配法(Pattern Match, PM)、矢量量化法(Vector Quantization,VQ)、混合高斯模型法(Gaussian Mixture Model, GMM)、隐马尔可夫模型法(Hidden Markov Model, HMM)和人工神经网络法(Artificial Neural Network, ANN)。人工神经网络在数据处理中，具有抗干扰能力强，能自适应学习等优点，所以近年来各种类型的神经网络在说话人识别中得到应用，网络的训练方法一般采用BP训练方法。针对普通的BP网络固有的收敛速度慢，易陷入局部极小值的缺点，本文利用小波变换良好的时频局域特性和多分辨分析功能，把综合小波变换与神经网络优点的小波神经网络[1,2](Wavelet Neural Network,WNN)用于说话人识别。在说话人识别的测试中，与普通BP网络相比，网络训练收敛快，避免了局部最优的非线性优化问题，提高了识别率。

2 小波神经网络

小波神经网络由法国著名信息科学研究机构IRISA的Zhang和Benveniste于1992年首次提出，并用于逼近$L(R^n)$中的

2004-11-01 收到，2005-03-22 改回

函数$f(x)$。由于小波变换良好的时频局域特性和多分辨分析功能，使得小波神经网络表现出了良好的辨识性能和逼近任何函数的能力。

2.1 小波理论

定义1　设函数 $\Psi \in L^2(R) \cap L^1(R)$，并且由 Ψ 经伸缩平移得到一族函数:

$$\Psi_{a,b}(x) = |a|^{-\frac{1}{2}} \Psi\left(\frac{x-b}{a}\right), \qquad a,b \in R,\ a \neq 0 \tag{1}$$

称 $\{\Psi_{a,b}\}$ 为连续小波，称 Ψ 为基本小波或母小波。其中a为伸缩因子，b为平移因子。

定义2　设函数 $f(x) \in L^2(R)$，则 $f(x)$ 的小波变换定义为：以函数族 $\Psi_{a,b}(x)$ 为积分核的积分变换，即

$$W_f(a,b) = \int_{-\infty}^{+\infty} f(x)\Psi_{a,b}(x)\mathrm{d}x = \int_{-\infty}^{+\infty} f(x)|a|^{-\frac{1}{2}} \Psi\left(\frac{x-b}{a}\right)\mathrm{d}x \tag{2}$$

由式(2)可知，参数a既改变窗口的大小和形状，也改变小波频谱 $\hat{\Psi}_{a,b}(\omega)$ 的谱图。参数a的值减少时，$\Psi_{a,b}(x)$ 的支撑区也随之变窄，而 $\hat{\Psi}_{a,b}(\omega)$ 的频谱随之向高端展宽；反之，向相反方向变化。这就实现了窗口大小的自适应变化，使小波变换在低频部分具有较高的频率分辨率，在高频部分具有较高的时间分辨率，对信号有自适应特性。

定义3　离散小波变换中，当基本小波 $\Psi(x)$ 经伸缩与位移引出的函数族:

$$[\Psi_{jk}(x) = 2^{-\frac{j}{2}}\Psi(2^{-j}x-k) \mid j \in z^+, k \in z] \tag{3}$$

对任意函数 $f \in L^2(R^n)$，满足 $A\|f\|^2 \leq \sum_j \sum_k |<f,\Psi_{jk}>|^2 \leq B\|f\|^2, 0 < A \leq B < \infty$ 时，便称 $[\Psi_{jk}(x) \mid j \in z^+, k \in z]$ 构成一个框架。

定理1　对任意 $f(x) \in L^2(R^n)$，小波族 Ψ 满足框架条件[3]，存在小波展开系数 c_{jk}，且满足

$$f(x) = \sum_{j,k} c_{j,k}\Psi_{j,k}(x) \tag{4}$$

定理1表明，任意函数 $f(x)$ 都可以用小波族以任意精度逼近。

定理2　同族小波函数组成的小波神经网络，具有通用逼近性及L^2逼近性[4]。

所谓通用逼近性是指对于任何 $f \in C(U)$　($C(U)$指由在紧集U上的连续函数构成的函数空间)　存在函数序列 f_n，使得 f_n 一致逼近 f。

2.2 小波神经网络结构

小波神经网络应用在说话人识别中，语音的特征(网络输入)和说话人身份(网络输出)构成映射的过程。实际上是在小波特征空间中寻找一组合适的小波基，通过对小波参数或形状迭代计算以使其输出误差函数最小化来实现，也是实现对复杂函数逼近的一种形式。小波函数的选择应满足两个条件：第一，定义域是紧支撑的，即函数应有速降特性，在一个很小的区间之外，函数值为0，以便获得空间局域化；第二，应具有振荡性，即是一个波。本文选用Mexican hat小波函数。

小波神经网络是以小波基函数为神经元激励函数的前馈网络模型[5]，可以看做是径向基函数网络的推广，其结构如图1所示。本文中小波神经网络采用3层结构，输入层L个神经元，隐含层M个神经元，输出层S个神经元；V_{ji} 表示隐含层第j个神经元和输入层第i个神经元间的连接权值，W_{kj} 表示输出层第k个神经元和隐含层第j个神经元间的连接权值;第n个样本的输入为 X_i^n，其中 i=1,2,⋯,L，网络的输出为 Y_k^n，其中k=1,2,⋯,S, n=1,2,⋯,N，其中N为样本总数，对应的目标输出为 D_k^n，k=1,2,⋯,S，n=1,2,⋯,N。输入层激励函数为线性变换(输出＝输入)，隐含层激励函数为小波函数，输出层激励函数为Sigmoid函数。

实验中采用的BP网络[6]与小波网络的结构相似，不同之处是BP网络隐含层激励函数采用Sigmoid函数。

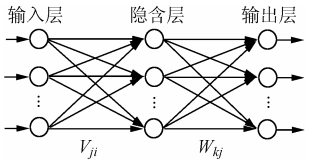

图 1　小波神经网络结构图

Fig. 1 Structure of wavelet neural network

2.3 小波神经网络学习算法

BP算法是目前应用广泛且比较成功的一种算法，网络中各个参数沿着误差能量函数梯度下降的方向调整，使输入值通过调整后的网络参数，最终输出满足误差能量函数要求的结果。算法如下[7,8]:

利用当前网络参数，可以得到网络的输出为

$$Y_k^n = \Phi\left[\sum_{j=1}^{M} W_{kj}\Psi\left(\frac{\sum_{i=1}^{L} V_{ji}X_i^n - b_j}{a_j}\right)\right] \tag{5}$$

其中 k=1,2,⋯,S，n=1,2,⋯,N。

实际输出与期望输出之间的误差能量函数为

$$E = \frac{1}{2}\sum_{k=1}^{S}(Y_k^n - D_k^n)^2 \tag{6}$$

若要求出满足误差能量函数E最小的网络参数，可以求出在第n个样本时E对每个网络参数的偏导数:

$$\frac{\partial E}{\partial W_{kj}} = (Y_k^n - D_k^n)\Phi(z)(1-\Phi(z))\Psi(T) \tag{7}$$

$$\frac{\partial E}{\partial V_{ji}} = \sum_{k=1}^{S}(Y_k^n - D_k^n)\Phi(z)(1-\Phi(z))\frac{\partial \Psi(T)}{\partial T}\frac{X_i^n}{a_j} \tag{8}$$

$$\frac{\partial E}{\partial a_j} = \sum_{k=1}^{S}(Y_k^n - D_k^n)\Phi(z)(1-\Phi(z))W_{kj}\frac{\partial \Psi(T)}{\partial T}\frac{(-T)}{a_j} \tag{9}$$

$$\frac{\partial E}{\partial b_j}=\sum_{k=1}^{S}(Y_k^n-D_k^n)\Phi(z)(1-\Phi(z))W_{kj}\frac{\partial \Psi(T)}{\partial T}\frac{(-1)}{a_j} \tag{10}$$

其中T为隐含层神经元的输入值， z为输出层神经元的输入值。

$$T=\frac{\sum_{i=1}^{L}V_{ji}X_i^n-b_j}{a_j} \tag{11}$$

$$z=\sum_{j=1}^{M}W_{kj}\Psi\left(\frac{\sum_{i=1}^{L}V_{ji}X_i-b_j}{a_j}\right) \tag{12}$$

隐层小波基函数采用Mexican hat 小波函数:

$$\Psi(t)=(1-t^2)\exp\left(-\frac{t^2}{2}\right) \tag{13}$$

$$\frac{\partial \Psi(t)}{\partial t}=-2t\exp\left(-\frac{t^2}{2}\right)-t(1-t^2)\exp\left(-\frac{t^2}{2}\right) \tag{14}$$

输出层神经元激励函数为

$$\Phi(t)=\frac{1}{1+\exp(-t)} \tag{15}$$

$$\frac{\partial \Phi(t)}{\partial t}=\frac{\exp(-t)}{(1+\exp(-t))^2}=\Phi(t)(1-\Phi(t)) \tag{16}$$

按梯度下降学习算法有

$$\Delta W_{kj}^{ii+1}=-\eta\times\frac{\partial E}{\partial W_{kj}}+\varepsilon\times\Delta W_{kj}^{ii} \tag{17}$$

$$\Delta V_{ji}^{ii+1}=-\eta\times\frac{\partial E}{\partial V_{ji}}+\varepsilon\times\Delta V_{ji}^{ii} \tag{18}$$

$$\Delta a_j^{ii+1}=-\eta\times\frac{\partial E}{\partial a_j}+\varepsilon\times\Delta a_j^{ii} \tag{19}$$

$$\Delta b_j^{ii+1}=-\eta\times\frac{\partial E}{\partial b_j}+\varepsilon\times\Delta b_j^{ii} \tag{20}$$

其中η为学习速率，只要η足够小，就可以实现真正的梯度下降，但容易陷入局部极小点；如果η取的太大，BP算法又可能产生震荡而不收敛，所以η的选取很重要。ε为动态参量, ε使每一步迭代权值变化不仅能反映梯度变化，而且能够跟踪误差曲面的变化趋势。ΔW_{kj}^{ii}为上次调整后的权值差值，ΔW_{kj}^{ii+1}为本次调整后的权值差值。式(18) – 式(20)中同理。

2.4 小波神经网络初始值的设置

初始权值对神经网络的后续训练十分重要，初始权值设置的好可以大大加速收敛速度，初始权值设置的不好，学习次数会增加，甚至出现不收敛的情况。如果初始权值采用一定范围内的随机数，且和样本及网络结构没有联系，对不同的样本就难以得到最优的初始权值，从而导致多数训练结果不理想，需要大量的重复试验才能得到好的结果。本文采用与样本和网络结构相联系的权值选取方法来得到最优权值[9]。设置步骤如下：

(1)随机产生[–1,1]区间上均匀分布的随机数作为V_{ji}的初始值。

(2)再对V_{ji}进行归一化，即

$$V_{ji}=V_{ji}\Big/\sqrt{\sum_{i=1}^{L}V_{ji}^{\ 2}} \tag{21}$$

其中j=1,2,⋯ , M。

(3)然后使V_{ji}的值与输入层神经元数及隐含层神经元数相联系。

$$V_{ji}=C\times M^{1/L}\times V_{ji} \tag{22}$$

其中 j=1,2,⋯ ,M ， C的取值根据经验值取2.0～2.2间的任意数。

(4)最后与学习样本联系。若输入层第 i 个神经元的输入样本中最大值为 $X_{i\max}$，最小值为 $X_{i\min}$，则可以设定的值如下：

$$V_{ji}=\frac{2V_{ji}}{X_{i\max}-X_{i\min}} \tag{23}$$

$$a_j=\frac{\sum_{i=1}^{L}V_{ji}X_{i\max}-\sum_{i=1}^{L}V_{ji}X_{i\min}}{2\Delta\Psi} \tag{24}$$

$$b_j=\frac{\sum_{i=1}^{L}V_{ji}X_{i\max}(\Delta\Psi-t^*)+\sum_{i=1}^{L}V_{ji}X_{i\min}(\Delta\Psi+t^*)}{2\Delta\Psi} \tag{25}$$

其中 $\Delta\Psi$ ， t^* 分别为Mexican hat小波基函数的时域半径和中心。 $\Delta\Psi$ 取值1.08， t^* 取值为0。隐含层和输入层的权值用[–1,1]区间上均匀分布的随机数作为初始权值。

3 说话人识别系统

说话人识别的过程如图2所示。系统工作于两种模式: 和识别。在训练模式下，对训练语音进行端点检测等预处理，提取出说话人特征信息，把说话人特征信息送入小波神经网络进行训练，形成说话人的网络参数并进行保存；在识别模式下，对测试语音进行端点检测等预处理，提取出说话人特征信息，再根据每个人的网络参数进行匹配，并由判决逻辑给出比较判决的结果。

本实验所用语音数据来源于Timit数据库[10], 对不同人采用不同的连续语音训练，保证语音与文本无关。

特征提取时，首先对输入语音依次进行端点检测，预加重，分帧，加汉明窗，然后逐帧计算MFCC系数。C_0和C_1对说话人识别有负方面影响[11], 所以舍去，取C_2到C_{12}共11个系数作为说话人特征参数，形成网络输入的11阶语音特征。语音数据帧长为256点，帧移128点，采用$1-0.95Z^{-1}$网络进行预加重。该数据库中，一般20s语音经端点检测后取得2000帧，

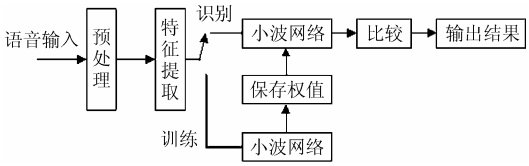

图 2　说话人识别系统图

Fig. 2　Diagram of speaker recognition system

所以每个人的语音特征为2000×11的矩阵。

网络训练时，因为特征值数目大，整体训练比单独训练的误差要大，所以采取逐行训练的方法。训练中，隐层神经元数的选取是决定网络性能的一个重要的因素。通过测试，隐层神经元个数根据经验值得到，采用11个神经元。实验中说话人总数为5个，则网络输出神经元取5个，目标输出为2000×5的矩阵，对第1个待识别人，网络的期望输出为1，0，0，0，0，对第2个待识别人为0，1，0，0，0，依次类推。网络采用11-11-5的前馈型结构，用BP算法对每个人逐个训练，达到期望误差精度后，保存网络的权值和域值。

识别时，待识别语音和训练语音的处理过程相同，完成语音特征的提取后，调用每个人语音训练时的网络权值，逐个计算网络的输出值，找出与目标输出之间误差最小的，即为识别出的说话人。

取网络参数η=0.8，ε=0.1，训练误差精度取E=0.001。

4　实验结果

4.1　小波网络与BP网络训练速度的比较

采用上述网络结构和初始网络参数，分别训练两个网络，得到训练步数与误差值的数据，如图3所示。可见，小波网络的训练速度比BP网络要快。

4.2 说话人识别结果

利用设计的说话人识别系统进行了说话人识别的实验。训练时所用语音长度为20s，识别时采用不同的语音长度得到不同结果的识别率，如表1所示。可见，识别率随识别语音长度的减小而降低，在语音长度为10s时得到最好的识别结果，识别率为99.5%。在相同训练条件下，小波网络以较快的训练速度得到的识别率比BP网络有较大提高，证明了该系统的有效性。

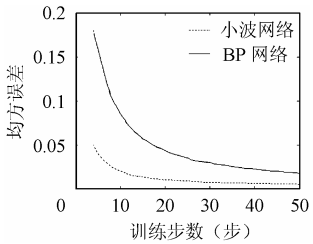

图 3　小波网络与 BP 网络训练速度的比较

Fig. 3　comparison of training speed between wavelet neural network and BP network

表 1　不同语音长度时的识别率

Tab.1 Recognition correctness with different speech length

识别用语音长度(s)	10	8	5	3
小波网络识别率(%)	99.5	98.6	96.2	90.4
BP网络识别率(%)	95.8	88.5	85.4	76.1

5　结束语

说话人识别是语音信号处理中的一个重要研究方向。本文把小波神经网络理论应用于说话人识别中，提出一种基于小波神经网络的说话人识别系统方案，建立了适用于说话人识别的小波神经网络模型，采用Mel倒谱系数作为说话人的特征，进行了说话人识别实验，与传统的BP网络相比，具有收敛速度快、识别率高的优点。

参 考 文 献

[1] Zhang Qinhua, Benveniste Al. Wavelet networks. *IEEE Trans. on Neural Networks*, 1992, 3(6)：889−898.

[2] Szu H, Telfer B, Kadambe S. Neural network adaptive wavelets for signal representation and classification. *Optical Engineering,* 1992, 31(9)：907−1016.

[3] 彭玉华. 小波变换与工程应用. 北京: 科学出版社, 2002：7−8

[4] Zhang J, Walter G. Wavelet neural networks for function learning. *IEEE Trans. on Signal Processing*, 1995, 43(6): 1485−1497.

[5] 李卫斌, 刘芳.小波神经网络的构造. 模式识别与人工智能, 2003, 16(4)：403−406.

[6] 焦李成. 神经网络的应用与实现. 西安：西安电子科技大学出版社, 1996, 第一章.

[7] Yoshihiro Yamamoto, Nikiforuk P N. A new supervised learning algorithm for multilayered and inter-connected neural networks. *IEEE Trans. on Neural Network*, 2000,11(1)：36−46.

[8] 李金平，王风涛，杨波. BP小波神经网络快速学习算法研究. 系统工程与电子技术，2001, 23(8)：72−75.

[9] 赵学智，邹春华，陈统坚. 小波神经网络的参数初始化研究. 华南理工大学学报(自然科学版), 2003, 31 (2)：77−80.

[10] Lamel L F, Kessel R H, Seneff S. Speech database development :Design and analysis of the acoustic-phonetic corpus. Proc.Speech Recognition Workshop(DARPA), 1986: 100−109.

[11] 甄斌，吴玺宏，刘志敏. 语音识别和说话人识别中各倒谱分量的相对重要性. 北京大学学报, 2001, 37(3): 371−378.

白　莹：　女，1980年生，硕士生，研究方向为信号处理、数字通信网技术等.

赵振东：　男，1956年生，教授级高级工程师，研究生导师，研究方向为数字通信系统等.

戚银城：　男，1968 年生，副教授，研究方向为信号与图像处理、数字通信网技术等.