RBF 神经网络及其在基于输出的客观音质评价中的应用

鄢田云, 云 霞, 靳 蕃, 朱庆军
(西南交通大学计算机与通信工程学院, 四川成都 610031)

摘 要: 针对汉语连续语音, 本文提出了采用径向基函数神经网络(RBFNNs), 对基于输出的语音质量进行客观评价的一种新方法))) RBFOBSQ(Output2Based Speech Quality Using RBFNNs). 该方法采用 Mel 倒谱对语音系统输出端的待测语音信号进行特征参数提取, 然后通过 RBF 神经网络完成特征参数到主观评价 MOS 分的非线性映射, 其映射值即为仅依赖于输出的客观音质评价结果, 其与主观评价 MOS 分的相关度, 当采用训练集样本时达到 0.92 以上, 而采用测试集样本时达到 0.88 以上.

关键词: RBF 神经网络; Mel 倒谱; 语音质量客观评价
中图分类号: TN912 文献标识码: A 文章编号: 0372-2112 (2004) 08-1282-04

RBF Neural Networks and Their Application to Output-Based Objective Speech Quality Assessment

YAN Tian-yun, YUN Xia, JIN Fan, ZHU Qing-jun
(School of Computer and Comm. Eng., Southwest Jiaotong University, Chengdu, Sichuan 610031, China)

Abstract: In dealing with continuous Chinese speech, this paper proposes a novel method (RBFOBSQ), using Radial Basis Function Neural Networks (RBFNNs), for output-based objective speech quality assessment. First the characteristics of speech signals at the output of the speech system were extracted by Mel cepstrum coefficients. Then the mapping from the characteristics to the Mean Opinion Score (MOS) was accomplished by the RBFNNs, and the average value of the RBFNNs. outputs was the result of output-based objective speech quality. The experimental results show that the correlation degree reaches more than 0.92 when using the train samples and attains more than 0.88 when using the test samples.

Key words: RBF neural networks; Mel cepstrum; speech quality objective assessment

1 引言

自 20 世纪 70 年代以来, 语音质量客观评价研究工作得到了迅速发展, 一些客观评价方法与主观评价结果达到了较高的相关度, 但这些工作主要集中于基于输入-输出方式的评价方法研究上[1, 2]. 但是基于输入-输出方式的评价方法要求参照输入端的原始语音, 而且输入输出两端的语音要求严格的内同步. 这些要求阻碍了该方法在实际中的广泛应用, 特别在现代军事领域、无线移动通信等领域在得不到原始语音的情况下也要能对语音质量进行评价, 即需要采用基于输出方式的客观音质评价方法. 在 90 年代中期, 基于输出的客观音质评价方法开始受到国内外学者的重视[2~4].

在国内外已提出的各种谱失真测度中, Mel 倒谱失真测度是一种弯折频率谱失真测度, 由于较充分地反映了人耳对频率及幅度的非线性感知特性, 以及人耳在听到复杂声音时所表现的频率分析和谱合成特性, 受到了广泛的重视和应用[5]. Mel 测度具有较好的主、客观相关性[6], 在语音质量客观评价和语音识别等方面得到了大量实际应用, 取得了较好的效果[5, 7, 8]. 因此, 本文采用 Mel 倒谱对汉语连续语音信号进行特征参数的提取.

RBF 神经网络除了具有一般神经网络的优点, 如多维非线性映射能力, 泛化能力, 并行信息处理能力等, 还具有很强的聚类分析能力, 学习算法简单方便等优点[9]; 它能将语音的动静态特性和听觉感知特性融合到网络特性之中, 用于客观音质评价时, 能使客观评测的结果与主观感知更接近. 因此, 本文采用径向基函数(RBF)神经网络完成语音信号特征参数到主观评价平均意见分(MOS)的映射, 得到基于输出的客观音质评价结果.

收稿日期: 2003-05-16; 修回日期: 2004-04-01
基金项目: 国防预研基金(No. 31161.2.6)

2　基于 Mel 倒谱的特征参数提取

Mel 倒谱失真测度建立在语音信号频域分析基础上, 而且根据人类听觉系统对频率及幅度的感知实验结果[10, 11], 将声音的频率非线性弯折到一个新的频率尺度, 在此尺度下提取语音特征参数. Mel 频率尺度可以比较准确地反映人的听觉系统对音高的感知与声音频率之间的关系; 而且 Mel 倒谱的计算量较小, 使得它在语音识别, 特别在连续语音识别等方面得到了更加广泛的应用[7, 8]. 对汉语连续语音进行 Mel 频率倒谱系数 (Mel Frequency Cepstrum Coefficient, MFCC) 的提取流程如图 1 所示.

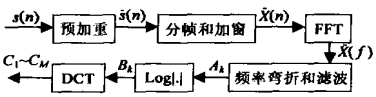

图 1　Mel 倒谱特征参数提取流程

¹ 对语音信号 s(n) 进行预处理: 包括预加重、分帧和加窗两部分. 输出端的待测信号 s(n) 经过预加重后的信号为 s(n); 本文研究过程中采用 8KHz 采样, 分析帧长取 25ms, 即 200 个采样点; 然后逐帧加窗, 窗函数为 Hamming 窗.

º FFT 变换: 对每一帧语音采样序列 X(n) 补 0 后进行 256 点 FFT 变换, 然后取模平方得功率谱 X(f) .

» 频率弯折和滤波: 进行频率弯折时, 1000Hz 以下采用线性频率弯折(warping), 而在 1000Hz 以上采用对数频率弯折. 用 mel 表示 Mel 频率, f 表示线性频率, Mel 频率与线性频率的映射关系按下式计算,

$$mel = 1000\log_2(1+ f/1000). \quad (1)$$

经频率弯折处理后, 将 X(f) 通过 Mel 测度三角带通滤波器组, 求得其通过每个数字滤波器的能量加权和 A_k.

¼ 对 A_k 求 Log| .| , 得到 B_k.

½ 对 14 点 B_k 作离散余弦变换(DCT), 得到 MFCC 系数 C_1~ C_M.

$$c_m = \sum_{k=1}^{M} B_k \cos[m(k- \frac{1}{2})\frac{P}{M}], \ m= 1, 2, \ldots, M= 14. \quad (2)$$

3　RBF 神经网络结构及算法

RBF 神经网络的工作原理分为两个阶段: 其一是学习阶段, 选定充足和高质量的训练语音样本训练 RBF 神经网络, 学习结果以权值的形式存储在网络结构之中; 其二是工作阶段, 当测试语音样本输入 RBF 神经网络时, 训练好且具有一定泛化性的网络将进行内插和外推等方式自适应完成特征匹配过程, 给出客观音质的评价结果. 其中训练语音样本和测试语音样本是不同失真条件下的汉语连续语音信号.

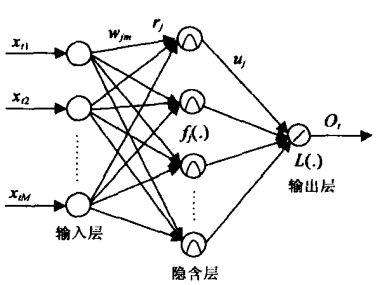

图 2　RBF 神经网络结构

根据 MFCC 的维数, 采用一个具有 14 个输入节点, J 个隐节点, 1 个输出节点的三层 RBF 神经网络 (图 2) 实现语音信号的 Mel 特征参数到 MOS 的映射. 设训练集的失真条件数为 D, 由仿真试验可知, J= D 时的仿真结果优于 J= 2D、3D、4D、5D 时的仿真结果, 而且 J 的取值越大训练时间越长, 所以在本文中隐节点数 J 等于训练集的失真条件数 D. 该网络的第 t 个输入为 $x_t= (x_{t1}, x_{t2}, \ldots, x_{tM})$, M= 14, 即待测语音信号的 Mel 特征参数矢量. 训练时取每一个 x_t 所对应的主观评价 MOS 分为其期望输出, 设某种失真条件的第 t 个输入矢量 x_t 相应的网络输出值为 O_t, 这种失真条件的 MOS 估值就是该失真条件下所有 O_t 的统计平均.

在图 2 中, 第 j 个隐含层节点到第 m 个输入节点的连接权值为 w_{jm}, r_j 为隐含层第 j 个节点的高斯核宽度; 输出节点到第 j 个隐含层节点的连接权值为 u_j; L(.) 为线性函数; f_j(.) 为隐含层第 j 个节点的激励函数, 取高斯型函数, 其表达式为

$$f_j(x_t)= \exp\left[- \sum_{m=1}^{M} \frac{(x_{tm}- w_{jm})^2}{2r_j^2}\right], \ j= 1, 2, \ldots, J. \quad (3)$$

网络的最终输出 O_t 由下式得

$$O_t= \sum_{j=1}^{J} u_j f_j(x_t). \quad (4)$$

为了提高 RBF 神经网络的收敛速度, 将隐含层参数 w_{jm}、r_j 和输出层权值 u_j 分开进行训练. 对隐含层参数 w_{jm} 和 r_j 的训练采用一种新的聚类算法, 即改进最近邻聚类学习算法, 此算法具有学习时间短、计算量小、网络性能优良等优点[12]. 针对语音参数处理的特性, 对文献[12]中的自适应最近邻聚类学习算法进行改进; 对输出层权值 u_j 的训练采用梯度下降算法.

改进最近邻聚类算法过程:

¹ 在训练语音样本中, 共有 J 个失真条件; 采用计数器统计属于各类的样本个数.

º 对第 1 个失真条件的所有特征参数向量, 进行自适应最近邻聚类[12], 设聚成 L_1 类, 且第 l_1 个计数器的值最大, 则令第 l_1 个聚类中心为 $w_{1m}, l_1 \in [1, 2, \ldots, L_1]$. 计算第 1 个失真条件的所有特征参数向量与 w_{1m} 的范数, 则令其中最大的范数为 r_1.

» 对第 j 个失真条件的所有特征参数向量, 进行自适应最近邻聚类, 设聚成 L_j 类, 且第 l_j 个计数器的值最大, 则令第 l_j 个聚类中心为 $w_{jm}, l_j \in [1, 2, \ldots, L_j]$. 计算第 j 个失真条件的所有特征参数向量与 w_{jm} 的范数, 则令其中最大的范数为 r_j, 其中 j= 2, 3, , , J.

¼ 最后得到的向量($w_{1m}, w_{2m}, \ldots, w_{Jm}$) 及($r_1, r_2, \ldots, r_J$) 就是 RBF 神经网络隐含层参数 w_{jm} 和 r_j, j= 1, 2, , , J; m= 1, 2, . . . , 14.

梯度下降算法过程:

¹ 给 u_j 赋随机初值, j= 1, 2, , , J; 由改进最近邻聚类算法得到隐含层参数 w_{jm} 和 r_j 以及采用式(3) 和式(4) 计算神经网络的输出 O_t;

º 计算 MOS 分 Y_t 与 RBF 神经网络的输出 O_t 之间的误差为

$$e_t= Y_t- O_t= Y_t- \sum_{j=1}^{J} u_j f_j(x_t). \quad (5)$$

» 定义目标函数为

$$E(t)=\frac{1}{2}e_t^2. \tag{6}$$

¼ 在 t+ 1时刻, 输出层权值 u_j 按如下的规则更新, 其中 A为训练系数,

$$u_j(t+1)=u_j(t)+A\frac{\partial E(t)}{\partial u_j(t)} \tag{7}$$

4 基于输出的客观音质评价结果

对应的主观评价得分, 采用符合国家军用标准(SJ/T 20771- 2000) 的MOS 分. 进行MOS 评价按语音质量优、良、中、差、劣五个等级进行评定, 对五个等级的投票分别给予5、4、3、2、1 的加权, 最后按文献[6] 的加权平均公式计分.

第 j 个失真条件对应的客观评价值Y(j) 与主观评价值 O(j) 之间的拟合关系采用二次多项式

$$Y(j)=a+b*O(j)+c*O(j)^2. \tag{8}$$

式(8) 中, a、b、c 为常数, 其中 j 是指第j 个失真条件, 设共有 J 种失真, j= 1, 2, . . . , J .

客观音质的评价值性能的好坏, 主要以其客观评价值与主观评价值的相关性高低来衡量. 相关系数 Q由下式计算,

$$Q=\sqrt{\sum_{j=1}^{J}[O(j)-Y]^2/\sum_{j=1}^{J}[Y(j)-Y]^2} \tag{9}$$

式(9) 中, Y 为所有Y(j) 取平均值. 采用两种形式的标准偏差, 其计算式为,

$$R_1=\sqrt{\frac{1}{J}[O(j)-Y(j)]^2}. \tag{10}$$

$$R_2=R\sqrt{|1-Q^2|} \tag{11}$$

式(11) 中, R 为主观MOS 分的标准偏差.

对于汉语连续语音数据样本, 从系统模型的基本功能考虑, 有选择性的采用军用标准 SJ/T 20771- 2000/ 军用通信系统音质MOS 评价法0 的标准测试语音数据库. 对于通信体制为模拟话音 SSB 调制的汉语连续语音, 有不同干信比的失真条件近 20 种, 作为数据集 1, 取其中一半作为训练集 1, 另一半作为测试集 1. 对于通信体制为 8Kb/ s 的 CS2ACELP(G. 729) QPSK 调制的汉语连续语音, 有不同干信比的失真条件共 40 余种, 作为数据集 2, 取其中一半作为训练集 2, 另一半作为测试集 2.

RBF 神经网络客观评 MOS 估值与主观评价 MOS 分的相关系数, 及其两种形式的标准偏差如表 1 所示. 训练集 1 和测试集 1 的实验结果分别如图 3(a) 与图 3(b) 所示; 训练集 2 和测试集 2 的实验结果分别如图 4(a) 与图 4(b) 所示. 从图 3、图 4 的二次拟合曲线(标示于各图题) 和表 1 可见, 采用 RBF 神经网络进行基于输出的客观音质评价结果与主观评价 MOS 分具有较高的相关度, 当采用训练集样本时相关度达到 0. 92 以上, 而采用测试集样本时达到 0. 88 以上, 而且标准偏差较小. 从本文与参考文献[2]、[4] 进行相关度方面的比较(见表 2) 可见, 由于 RBF 神经网络具有多维非线性映射性和很强的聚类分析能力, RBFOBSQ 方法的客观评价结果明显优于其它一些典型的基于输出的评价方法.

表 1 RBFOBSQ 评价的主客观相关系数及标准偏差

参数	训练集 1	测试集1	训练集 2	测试集 2
Q	0. 9265	0. 9058	0. 9224	0. 8894
R_1	0. 2021	0. 2453	0. 1068	0. 1492
R_2	0. 1163	0. 1741	0. 2020	0. 2216
R	0. 3091	0. 4110	0. 7386	0. 4849

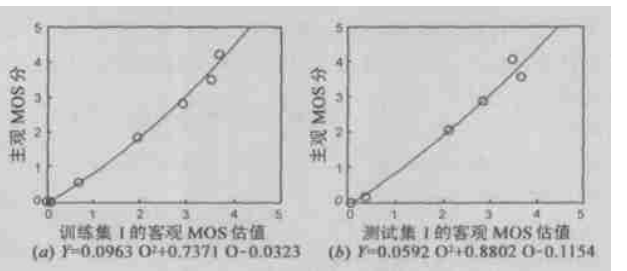

图 3 数据集1 的客观MOS 估值与主观评价MOS 分的关系

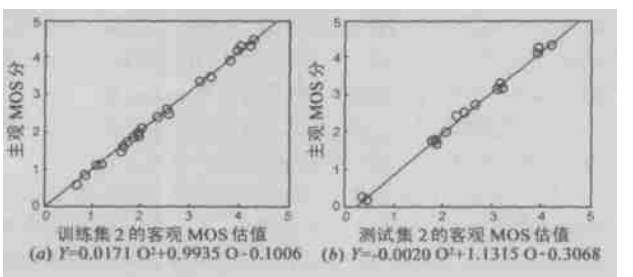

图 4 数据集2 的客观MOS 估值与主观评价MOS 分的关系

表 2 本文与参考文献[2] 和[4] 进行相关度比较

文 献	采用方法	平均相关度	最大相关度
本 文	RBFOBSQ	0. 9110*	0. 9265
文献[2]	OMBFD	0. 6926	0. 7553
文献[4]	PLP	0. 7525	0. 8200
	PLP2CD	0. 8037	0. 8900
	PLP2DCD	0. 7875	0. 8500

注: * 是对表 1 中的 Q 取平均值.

5 结论

本文提出了采用 RBF 神经网络, 对基于输出的汉语连续语音质量进行客观评价的一种新方法, 即 RBFOBSQ 方法. 首先采用 Mel 倒谱对待测语音信号进行特征参数提取, 然后利用 RBF 神经网络的多维非线性映射原理, 完成特征参数到主观评价 MOS 分的映射. 不但免去了基于输入输出的客观音质评价要解决原始信号和待测信号起点严格对齐的内同步关键技术和平均失真测度的复杂计算, 而且可以充分地反映人耳听觉系统的感知特性及计算简便. 主客观评价结果的相关性实验表明, 当采用训练集样本时相关度达到 0. 92 以上, 而采用测试集样本时达到 0. 88 以上, 这表明 RBFOBSQ 方法具有明显的优越性.

参考文献:

[1] Quackenbush S Q, Barnwell Ó T P, Clements M A. Objective Measures of Speech Quality[M]. Englewood Cliffs, NJ: Prentice Hall, 1988.

[2] 陈国, 胡修林, 张蕴玉, 朱耀庭. 多标度分形理论及其在语音质量客观评价中的应用[J]. 声学学报, 2002, 27(6): 531- 535.
[3] 陈国, 胡修林, 张蕴玉, 朱耀庭. 语音质量客观评价方法研究进展[J]. 电子学报, 2001, 29(4): 548- 552.
[4] Jin C, Kubichek R. Vector quantization techniques for output- based objective speech quality[A]. Proc of IEEE ICASSP[C]. Geogia, USA: IEEE, 1996. 1. 491- 494.
[5] 付强, 易克初, 田斌, 田红心. 一种采用余弦镶边临界带滤波器组的弯折谱失真测度[J]. 西安电子科技大学学报, 1999, 26(6): 823- 827.
[6] 黄惠明, 王瑛, 赵思伟, 张知易. 语音系统客观音质评价研究[J]. 电子学报, 2000, 28(4): 112- 114.
[7] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences[J]. IEEE Trans on ASSP, 1980, 28(4): 357- 366.
[8] Steve Y. A review of large2vocabulary continuous2speech recognition [J]. IEEE Signal Processing Magazine, 1996, 13(5): 45- 57.
[9] 靳蕃. 神经计算智能基础原理与方法[M]. 成都: 西南交通大学出版社, 2000. 101- 173.
[10] Stevens S S, Volkman J. The relation of pitch of frequency: a revised scale[J]. Am J Psychol, 1940, 53: 329- 353.
[11] Zwicker E, Flottorp G, Stevens S S. Critical bandwidth in loudness sum2mation[J]. J Acost Soc Am, 1947, 19: 90- 119.
[12] 周俊武, 孙传尧, 王福利. 径向基函数(RBF)网络的研究及实现[J]. 矿冶, 2001, 10(4): 71- 75.

作者简介:

鄢田云　女, 1975 年生于湖南新化, 现为西南交通大学计算机与通信工程学院博士研究生, 方向为智能信息处理, 目前主要研究领域为遗传算法、神经网络和客观语音质量评价等.

云　霞　女, 1980 年生于四川泸州, 现为西南交通大学计算机与通信工程学院硕士研究生, 方向为语音通信, 目前主要研究领域为客观语音质量评价.