粒子群优化径向基函数网络的语音转换

董添辉[1],张玲华[2]
(1. 南京邮电大学 通信与信息工程学院,江苏 南京 210003;
2. 江苏省通信与网络技术工程研究中心,江苏 南京 210003)

摘 要:径向基函数神经网络具有结构简单和学习速度快等特点,因此常被用作语音转换的模型。隐层核函数的中心是影响径向基函数神经网络性能的重要参数,而传统的 K-均值聚类算法受初值影响大,全局优化的效果不佳。所以,选择合适的优化算法来调整 RBF 网络核函数的中心参数,能改善整个网络的性能,从而提升语音转换的效果。而粒子群算法是一种基于迭代的优化算法,具有容易实现、算法参数少、收敛快和突出的全局寻优能力等特点。提出了一种改进的粒子群算法,优化了径向基函数的中心以提高网络性能,便于更准确地获得说话人与目标人之间谱包络的映射关系。实验结果表明,提出的方法能够有效提高神经网络的性能,使转换后的声音更接近于目标声音。

关键词:语音转换;径向基函数中心;改进的粒子群算法;径向基函数神经网络
中图分类号:TN912.3 **文献标识码**:A **文章编号**:1673-629X(2017)05-0064-05
doi:10.3969/j.issn.1673-629X.2017.05.014

Voice Conversion of Radial Basic Function Neural Network of Particle Swarm Optimization

DONG Tian-hui[1],ZHANG Ling-hua[2]
(1. College of Telecommunications & Information Engineering,Nanjing University of Posts and Telecommunications,Nanjing 210003,China;
2. Jiangsu Provincial Engineering Research Center of Telecommunications and Network Technology,Nanjing 210003,China)

Abstract:Due to simple structure and fast learning,Radial Basis Function (RBF) neural network is used commonly in voice conversion system. The center of kernel function in hidden layer is the important parameter of influencing the RBF neural network,but traditional K-means clustering algorithm relies on the initial value,which is ineffective in global optimization. Therefore,it is significance to select a suitable algorithm to modulate the center of function and enhance the effect of voice conversion. Particle swarm algorithm is an optimized one based on iteration,with the characteristics of easy implementation,much less parameters,fast convergence and better global optimization and so on. An improved particle swarm optimization is proposed to optimize the RBF's centers for improvement of the performance of RBF network,thus enhancing the transformation of speech parameters. The results acquired by modeling and simulation show that the proposed method has effectively improved the performance of neural network and the effect of converted voices is much closer to the goal.

Key words:voice conversion;centers of RBF;improved particle swarm optimization;radial basis function neural network

0 引 言

语音转换技术是指在不改变说话内容的前提下,转化源说话人声音的个性特征,使转换的语音更接近目标人。语音转换分为训练阶段和转换阶段。在训练阶段,通过模型对源说话人和目标说话人进行训练,得出相应的转换规则。在转换阶段,先提取源语音的个性特征,再根据训练阶段得到的转换规则进行转化,得到目标语音[1-2]。

常用的语音转换模型包括矢量量化法(Vector Quantization,VQ)、高斯混合模型(Gaussian Mixture

收稿日期:2016-06-06 **修回日期**:2016-09-21 **网络出版时间**:2017-03-13
基金项目:江苏省高校自然科学研究重大项目(13KJA510003);江苏高校优势学科建设工程(PAPD)
作者简介:董添辉(1991-),男,硕士,研究方向为语音信号的研究与应用;张玲华,博士生导师,通信作者,研究方向为语音信号的研究与应用、无线传感网络、数字助听器。

Model,GMM)、人工神经网络(Artificial Neural Network,ANN)等[3]。径向基网络作为一种简单的人工神经网络,具有计算量少、结果简单、学习速度快以及逼近任何非线性函数等特点[4]。重点研究人工神经网络在语音转换中的应用。

RBF 神经网络是一个类似于遗传网络的三层前馈型神经网络,该网络有三个非常重要的参数:隐层核函数的中心和宽度以及隐层到输出层的连接权值。针对径向基函数神经网络的核函数参数经传统 K -均值聚类算法训练存在收敛速度慢、易陷入局部最优、泛化性能不佳等问题,引入改进的粒子群优化算法(Particle Swarm Optimization,PSO)来训练隐层核函数的中心,并研究其在语音转换系统中起到的作用[5]。

为更准确地获得说话人与目标人之间谱包络的映射关系,提出了一种改进的粒子群算法,以优化径向基函数的中心并提高网络性能。对基于传统 GMM,基于 K -均值聚类、基于 PSO 算法以及基于改进 PSO 算法的 RBF 神经网络进行了语音转换实验,实验结果表明,提出方法相较于其他方法能够有效提高神经网络的性能,使转换后的声音更接近于目标声音。

1　RBF 神经网络

RBF 神经网络对任意的非线性函数具有良好的适应性,可以分析系统内一些难以解析的规律,具有良好的泛化和快速学习的能力[6]。因此,该网络被广泛用于语言转换领域。RBF 神经网络具有三层前向结构,包括输入层、隐层、输出层[7]。其结构如图 1 所示。

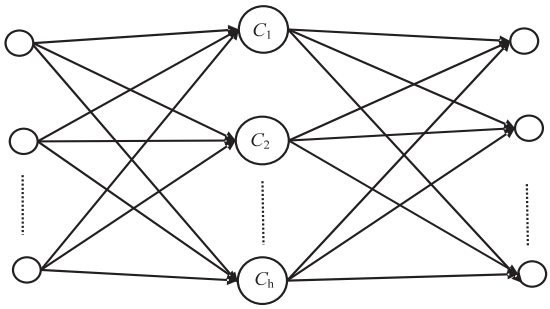

图 1　RBF 神经网络结构

一般采用 n 维的高斯函数作为径向基函数:

$$R(x-c_i)=\exp\left(-\frac{1}{2\sigma_i^2}\|\boldsymbol{x}-\boldsymbol{c}_i\|^2\right) \tag{1}$$

其中,$\|\boldsymbol{x}-\boldsymbol{c}_i\|$ 为欧氏几何范数,c_i 为第 i 个核函数的中心,σ_i 为第 i 个核函数的宽度。

RBF 神经网络的输出为:

$$y_i=\sum_{i=1}^{k} w_{ij}\exp\left(-\frac{1}{2\sigma_i^2}\|\boldsymbol{x}-\boldsymbol{c}_i\|^2\right),j=1,2,\cdots,n \tag{2}$$

其中,y_j 为输出向量;w_{ij} 为相对应的连接权值。

RBF 神经网络由两类参数组成:一类是隐层核函数的中心和宽度;另一类是隐层到输出层的连接权值[8]。常规 K -均值聚类算法的步骤如下:

Step1:初始化设置网络和各参数。

Step2:聚类中心 $c_i(i=1,2,\cdots,k)$ 由随机选择 k 个训练样本构成。

Step3:输入样本 x_p,按照近邻规则分组;根据欧氏距离(见式(3))将 x_p 分给离其最近的中心形成聚类。

$$d_i=\sqrt{\sum_{p=1}^{n}(x_p-c_i)^2},i=1,2,\cdots,k \tag{3}$$

Step4:重新调配聚类中心,计算每个聚类的均值来寻找新聚类中心。若随着迭代次数的增加聚类中心不再改变,则得到的聚类中心就是核函数的中心,否则返回 Step2。

Step5:计算核函数的宽度 σ:

$$\sigma_i=\frac{d_{\max}}{\sqrt{2k}},i=1,2,\cdots,k \tag{4}$$

其中,$d_{\max}$ 为所选中心的最大距离。

Step6:由最小二乘法可得隐层与输出层连接权值,计算公式如下:

$$w=\exp\left(\frac{k}{d_{\max}^2}\|\boldsymbol{x}_p-\boldsymbol{c}_i\|^2\right),p=1,2,\cdots,P \tag{5}$$

其中,P 为样本总数。

2　改进方法

隐层核函数的中心是影响 RBF 神经网络性能的重要参数,而传统 K -均值聚类算法受初值影响大,全局优化的效果不佳[9]。所以,选择合适的优化算法来调整 RBF 网络核函数的中心参数,能改善整个网络的性能。而粒子群算法是一种基于迭代的优化算法,具有容易实现、算法参数少、收敛快和突出的全局寻优能力等特点。因此,引入一种改进的 PSO 算法调整 RBF 神经网络的核函数中心,并将优化过的 RBF 神经网络应用于语音转换,以提高转换语音的相似度。

2.1　改进的粒子群算法

粒子群优化算法模拟鸟群觅食行为,通过粒子群的运动进行全局搜索。每个粒子都有一个相对应的适度值和速度矢量,分别表示距离及运动方向。在迭代算法中,通过比较每个粒子的全局极值 G_{best} 和个体极值 P_{best},对其位置和速度进行迭代更新[10]。

假设粒子群中有 N 个粒子,则第 i 个粒子在 D 维度空间中的位置表示为 $X_i=(x_{i1},x_{i2},\cdots,x_{iD}),i=1,2,\cdots,N$,速度记为 $V_i=(v_{i1},v_{i2},\cdots,v_{iD}),i=1,2,\cdots,N$。

通过每一次的迭代寻找 P_{best} 和 G_{best},找到极值后再根据式(6)更新粒子的位置和速度。

$$v_{id}^{k+1}=v_{id}^{k}*w_k+c_1\text{rand}()(p_{id}^{k}-x_{id}^{k})+$$

$$c_2 \text{rand}( ) (p_{gd}^k - x_{id}^k) \tag{6}$$

$$x_{id}^{k+1} = x_{id}^k + v_{id}^k \tag{7}$$

其中，$i = 1,2,\cdots,N$，$d = 1,2,\cdots,D$；k 为迭代次数；p_{id} 和 p_{gd} 分别为粒子个体极值和全局极值的位置；c_1，c_2 为加速因子；rand( ) 为 0 到 1 之间的随机数；w 为惯性权值，通过合适的调节方法可以在局部寻优与全局寻优之间找到平衡，惯性权值越小则局部寻优能力增强，全局寻优能力减弱，惯性权值越大则效果相反[10]。

采用一种非线性策略来调整 w，从而改进粒子群算法。

$$w_k = w_{\min} + (w_{\max} - w_{\min}) * \exp\left[-4.5 * \left(\frac{k}{k_{\max}}\right)^3\right] \tag{8}$$

其中，$w_{\max}$，$w_{\min}$ 分别为惯性权值的初始值和迭代结束值；$k_{\max}$ 为最大迭代次数；k 为当前迭代次数。

当最优位置的适度值符合最小适应阈值或迭代次数等于最大值时，该 PSO 算法结束[11]。

2.2 基于改进粒子群算法的 RBF 神经网络

将核函数的聚类中心 c_i 看作是 PSO 算法的粒子，通过 PSO 算法优化网络，从而提高网络性能[12]，步骤如下：

Step1：初始化网络。设定粒子个数及每个粒子大小并随机初始化各个粒子的位置和速度，设置惯性权值的初始值和结束值，最大迭代次数。

Step2：粒子空间位置优劣只能由适应度函数衡量，函数决定着整个算法的优化效果，根据实际问题，采用的适应度函数为：

$$f = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ji}' - y_{ji})^2 \tag{9}$$

其中，N 为训练样本数；M 为输出层节点数；y_{ji}' 为实际输出；y_{ji} 为预期输出。

Step3：将通过 RBF 网络计算得到的样本实际输出与相应的预期输出代入式(9)，得到该粒子的个体极值；计算各个粒子的适度值并进行比较，得出最优值并将其作为粒子的全局极值。

Step4：通过式(6) ~ (8)分别更新粒子的速度、位置和权值。

Step5：如果重新计算更新后粒子的适应度值优于以前位置的适应度值，则新位置取代以前位置成为下次迭代的起点，否则下一次迭代的起点不变。

Step6：若全局极值满足小于设定的阈值或者迭代次数达到最大，则改进 PSO 算法结束。否则，转至 Step3，继续进行迭代。

Step7：将改进粒子群算法得到全局最优值的位置作为 RBF 神经网络的核函数中心。

2.3 基于改进粒子群优化径向基函数神经网络的语音转换

在语音转换系统中，常用提取加滤波的短时线谱模型来计算声音参数，从而得到线性预测系数(Linear Predictive Coefficient，LPC)。这些系数通常转化成其他形式的参数，以适应所需的性质。线谱频率(Line Spectrum Frequency，LSF)参数是通过一系列的计算由 LPC 参数得来的[13]。LSF 参数能够客观反映共振峰的位置和带宽，具有良好的插值特征，并且特征参数的某一部分失真对合成谱参数影响较小，因此广泛用于语音信号处理[14]。实验采用自适应加权谱内插(STRAIGHT)模型来获得 LSF 参数和基音频率，以及合成转换语音。

语音转换系统由训练阶段和转换阶段两部分组成。在训练阶段，提取源和目标说话人声音的基频和线谱频率参数；再运用动态时间规划将源与目标的特征参数对齐；将源声音的参数作为 RBF 网络的输入，目标声音参数作为输出，通过人工神经网络建立转换规则[15]。在转换阶段，将源测试声音同样通过 STRAIGHT 模型提取 LSF 参数和基频，再利用训练阶段得到的转换规则进行转换。最后，利用 STRAIGHT 模型合成声音。图 2 为语音转换框图。

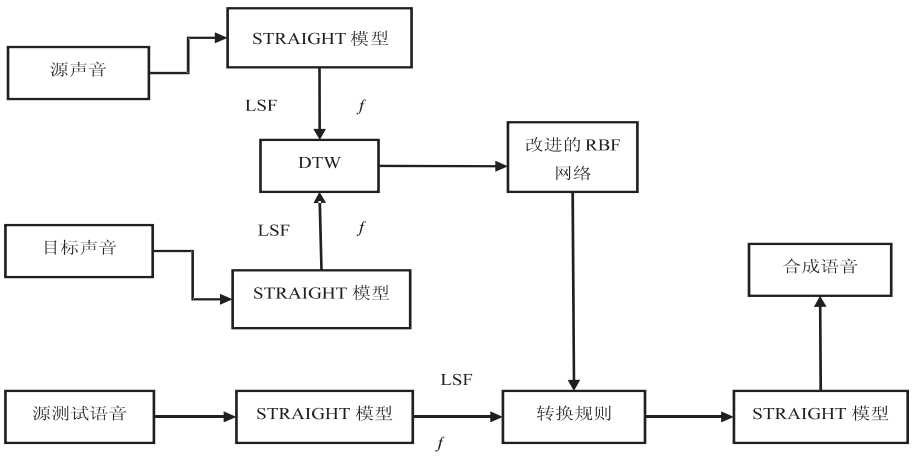

图 2 语音转换框图

3　实验结果分析

对基于传统 GMM,基于 K -均值聚类、基于 PSO 算法以及基于改进 PSO 算法的 RBF 神经网络进行语音转换实验。另外,由文献[7]可知基频在语音转换中起着重要的作用,将谱包络参数与基频联合通过径向基函数神经网络进行转换,转换后的基频含有更多目标人个性特征。实验中采用的数据库包含 2 个男子和 2 个女子的语音,每人的语音由 141 个单字和 6 句短语组成。采样率均为 16 kHz,并以 16 bit 量化。

3.1　主观评价

采用 ABX 和 MOS 对测试转换语音的效果进行主观评价。

ABX 法主要是对转换语音与目标语音的相似程度进行评价,A 和 B 分别代表源说话人声音和目标说话人声音,X 代表转换而来的声音。实验中,随机选择 10 位测评人对转换后的 30 个词语和 6 句短语进行评价,要求听众选择 A 或 B 哪个更接近 X,然后统计结果。表 1 给出了 3 种转换方法的"ABX"的测试结果。

表 1　ABX 测试结果　　%

转换类型	GMM		K 均值+RBF		改进 PSO+RBF		PSO+RBF	
	源	目标	源	目标	源	目标	源	目标
男转女	56.8	43.2	55.7	44.3	42.9	57.1	44	56
男转男	63.5	36.5	62.4	37.6	48.9	51.1	49.4	50.6
女转男	59.2	40.8	57.9	42.1	46.1	53.6	48.2	51.8
女转女	63.8	36.2	62.7	37.3	47.2	52.8	49.5	50.5

由表 1 结果可知,基于改进 PSO 算法的 RBF 网络所得到的转换语音相对于其他三种方法得到的语音更接近于目标语音,转换效果也较其他两种方法有显著提升。

平均主观意见分(Mean Opinion Score,MOS)将语音分为差、较差、尚可、好、极好这五个听觉质量等级,分别记为 1 ~ 5 分。实验中,同样随机选择 10 位测评人对转换后的 30 个词语和 6 句短语进行评价和打分,测试结果如表 2 所示

表 2　MOS 测试结果

转换类型	GMM	K 均值+RBF	改进 PSO+RBF	PSO+RBF
男转女	2.68	2.64	2.95	2.91
男转男	2.55	2.58	2.91	2.85
女转男	2.63	2.58	3.02	2.94
女转女	2.58	2.60	2.97	2.89

由表 2 可知,通过改进 PSO 算法的 RBF 网络得到的转换语音 MOS 分都有不同程度的提高,说明转换语音的清晰度和自然度都有所提高,性能优于其他三种方案。

3.2　客观评价

实验采用以女生到男生声音的转换为例作为客观评价。为了更加直观地了解提出的改进 PSO 算法对 RBF 网络的优化情况,采用谱失真率作为衡量客观评价的标准,如式(10):

$$r_{sd} = \frac{1}{N}\sum_{i=1}^{N}\frac{\| x_{i.\,con} - x_{i.\,targ} \|}{\| x_{i,sour} - x_{i,targ} \|} \tag{10}$$

其中,$x_{i,con}$,$x_{i,targ}$ 和 $x_{i,sour}$ 分别为转换后的声音、源声音和目标声音的包络参数;N 为声音的帧数。r_{sd} 的值越小,网络的性能越好。

图 3 给出了频率失真图。

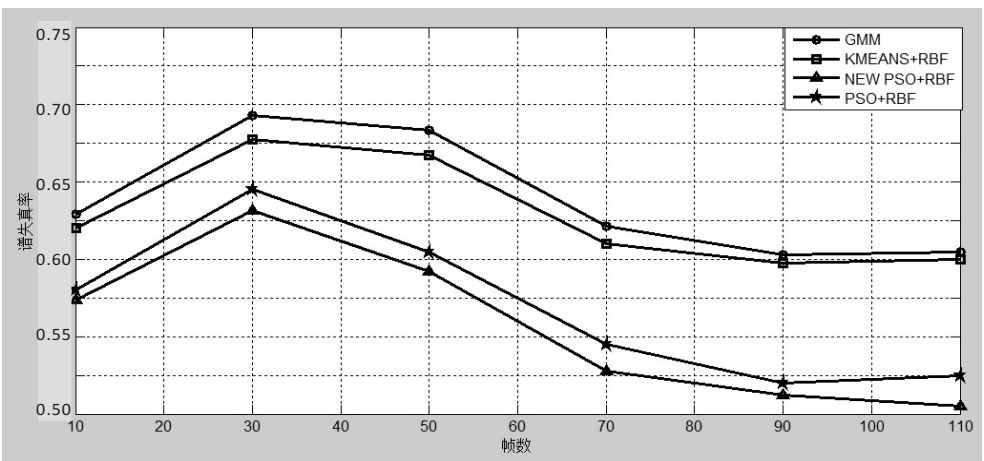

图 3　频率失真图

由图 3 可知,基于改进 PSO 算法优化 RBF 网络的语音转换的谱失真率最低,拥有更好的转换性能,在转换语音的质量上得带了进一步的提高。

为了进一步比较基于改进 PSO 算法优化的 RBF 网络转换的方法与其他方法之间谱包络的不同,将实验中得到的 LSF 系数通过一系列变换得到谱包络,如图 4 所示。

由图 4 可知,基于改进粒子群优化 RBF 神经网络的语音转换得到的谱包络更接近目标声音的包络,显著提高了捕捉共振峰的能力,所以能更好反映人说话

声音的特性。

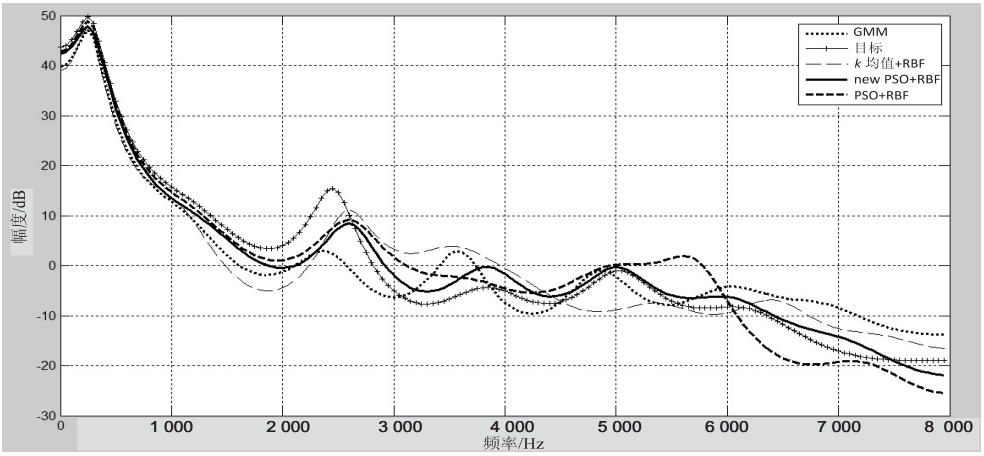

图 4 包络的比较

4 结束语

为了更加准确地建立语音转换的映射关系,改善语音转换的效果,提出了一种改进粒子群算法,以优化径向基函数神经网络性能,从而使得到的转换语音更接近目标声音。通过对四组实验的主客观评价结果进行比较可知,该方法能够更加准确地映射源声音与目标声音的关系,使得转换后的声音具有更多目标人声音的个性特征。

参考文献:

[1] 张玲华,姚绍芹,解伟超. 基于自适应粒子群优化径向基函数神经网络的语音转换[J]. 数据采集与处理,2015,30(2):336-343.

[2] 李 波,王成友,蔡宣平,等. 语音转换及相关技术综述[J]. 通信学报,2004,25(5):109-118.

[3] 简志华,杨 震. 语声转换技术发展及展望[J]. 南京邮电大学学报:自然科学版,2007,27(6):88-94.

[4] 解伟超,张玲华. 基于自组织聚类和改进粒子群算法的语音转换方法[J]. 声学学报,2014,39(1):130-136.

[5] Valbret H,Moulines E,Tubach J P. Voice transformation using PSOLA technique[C]//International conference on acoustics,speech,and signal processing. [s. l.]:IEEE,1992:145-148.

[6] 郭 通,兰巨龙,李玉峰,等. 基于量子自适应粒子群优化径向基函数神经网络的网络流量预测[J]. 电子与信息学报,2013,35(9):2220-2226.

[7] Chen Xiantong, Zhang Linghua. An improved ANN method based on clustering optimization for voice conversion[C]//International conference on audio,language & image processing. [s. l.]:IEEE,2014:464-469.

[8] Man Chuntao,Wang Kun,Zhang Liyong. A new training algorithm for RBF neural network based on PSO and simulation study[C]//Proceedings of IEEE international conference on computer science and information engineering. [s. l.]:IEEE,2009:641-645.

[9] Xie Fenglong,Yao Qian,Soong F K,et al. Pitch transformation in neural network based voice conversion[C]//Chinese spoken language processing. [s. l.]:IEEE,2014:197-200.

[10] 何隆玲. 基于改进 PSO-RBF 神经网络的高分辨率雷达目标检测研究[D]. 南宁:广西大学,2013.

[11] Andrews P S. An investigation into mutation operators for particle swarm optimization[C]//IEEE congress on evolutionary computation. [s. l.]:IEEE,2006:1044-1051.

[12] Bratton D,Kennedy J. Defining a standard for particle swarm optimization[C]//IEEE international conference on swarm intelligence symposium. [s. l.]:IEEE,2007:120-127.

[13] Qiao Y,Minematsu N. Mixture of probabilistic linear regressions:a unified view of GMM-based mapping techniques[C]//Proceedings of IEEE international conference on acoustics, speech and signal processing. Taipei, Taiwan: IEEE, 2009:3913-3916.

[14] Toda T,Saruwatari H,Shikano K. Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum[C]//IEEE international conference on acoustics, speech, and signal processing. [s. l.]:IEEE,2001:841-844.

[15] Desai S,Black A,Yegnanarayana B,et al. Spectral mapping using artificial neural Networks for voice conversion[J]. IEEE Transactions on Audio, Speech and Language Processing, 2010,18(5):954-964.