中 文 信 息 学 报
第 18 卷 第 1 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 18 No. 1

文章编号:1003 - 0077(2004)01 - 0078 - 07

基于遗传径向基神经网络的声音转换①

左国玉[1,2],刘文举[1],阮晓钢[2]

(1. 中科院自动化所 模式识别国家重点实验室,北京 100080;
2. 北京工业大学 电子信息与控制工程学院,北京 100022)

摘要:声音转换技术可以将一个人的语音模式转换为与其特性不同的另一个人语音模式,使转换语音保持源说话人原有语音信息内容不变,而具有目标说话人的声音特点。本文研究了由遗传算法训练的 RBF 神经网络捕获说话人的语音频谱包络映射关系,以实现不同说话人之间声音特性的转换。实验对六个普通话单元音音素的转换语音质量分别作了客观和主观评估,结果表明用神经网络方法可以获得所期望的转换语音性能。实验结果还说明,与 K- 均值法相比,用遗传算法训练神经网络可以增强网络的全局寻优能力,使转换语音与目标语音的平均频谱失真距离减小约 10 %。

关键词:人工智能;自然语言处理;声音转换;RBF 神经网络;遗传算法;线谱频
中图分类号:TN912.3 **文献标识编码**:A

Voice Conversion by GA-based RBF Neural Network

ZUO Guo-yu[1,2], LIU Wen-ju[1], RUAN Xiao-gang[2]

(1. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing 100080, China;
2. School of Electronics Information and Control Engineering, Beijing University of Technology, Beijing 100022, China)

Abstract: Voice conversion technology makes the speech of one speaker sounds as though it were uttered by another speaker giving it a new identity while preserving the original content. This paper addresses a study on voice conversion using genetic algorithm (GA) to train the hidden layers of RBF neural network, which can help better capture the nonlinear mapping between different speakers. Both subjective evaluations and objective ones are conducted on the transformed speech quality with six mono-vowel phones in Mandarin speech. Experimental results show that desired performance of converted speech can be obtained when a neural network method is applied to voice conversion technique. The evaluations report that compared with K-means method, a genetic algorithm based RBF network has the ability of global optimization with a 10 % decrease in the spectral distance between the transformed speech and the target speech.

Key words: artificial intelligence; natural language processing; voice conversion; RBF neural network; genetic algorithm; line spectrum frequency

1 引言

① **收稿日期**:2003 - 08 - 11
基金项目:国家自然科学基金项目(60172055;60121302);中科院自动化所领域前沿项目(1M02J05)
作者简介:左国玉(1971 —),男,博士生,主要研究方向为声音转换、语音信号处理、模式识别和人工智能.

声音转换或声音个性化是一项改变说话人声音特性的技术,使一个人的声音听起来像是由另一个人说出来的[1]。这项语音技术应用前景非常广泛,包括文语合成系统的定制、电影广播剧角色的自动配音、多说话人语音语料的采集和传输、语音识别过程中的预处理技术等。

很多研究人员采用各种各样的转换技术以合成期望的目标说话人语音。Abe 等人提出了基于矢量量化(VQ)的码本映射技术[2]。为产生映射码本,用矢量量化算法将源说话人和目标说话人的声学特征空间进行划分,用动态时间规整算法(DTW)将源 - 目标特征矢量相关联,从而训练出一个源 - 目标说话人的映射码本。码本映射技术虽然计算简单,但是由于矢量量化作用造成的频谱的不连续性,转换语音质量还很低。有学者提出如线性多变量回归(LMR)等方法的局部函数转换技术[3]。说话人语音频谱空间由矢量量化划分成许多不同的子空间,每一个空间都训练一个转换函数(也称作局部函数),每个转换函数都表述了某一个声学空间源 - 目标说话人特征之间的关系,这样码本映射方案中的全局映射就被这些局部函数来近似。这种局部空间转换的方法可以产生无穷多目标特征量,但由于选择单个局部转换函数的离散性还存在,不连贯性仍然出现在输出语音中。一些学者通过概率方法,采用高斯混合模型(GMM)描述源 - 目标特征的联合概率分布,这样由给定源特征矢量寻找转换函数来预测目标语音特征就是一个回归问题[4,5]。GMM 技术比码本映射和 LMR 局部变换等方法有效性、鲁棒性也较好,其原因在于对频谱包络建立了一个连续性概率模型。GMM 联合概率方法理论上能使回归问题的混合成分(mixture)得到更合理的配置,但在进行 EM 运算时计算量较大,而且存在转换语音频谱过分光滑现象,影响了转换语音目标说话人特征的倾向性。在基于 HMM 的 TTS 系统中,音素 HMM 作为语音合成单元,初始说话人无关的音素 HMM 在训练阶段由观察矢量训练而成,使用目标说话人语音对说话人独立的音素 HMM 做 MLLR 自适应,从而使自适应后的音素模型具有目标说话人特征[6]。这种转换方法合成的语音质量还不是太好,但优点是只需少量数据做自适应就可以方便地合成不同目标说话人的声音。

由于码本映射等技术的离散性和转换语音目标特征倾向的差异性,人工神经网络方法也被应用于声音转换技术。在这种方法中,即使训练语音数据量较少但只要选取合适,也能较好地学习一个连续特征映射函数,这种泛化特性有助于降低数据储备要求而能较好完成说话人特性之间的变换。因此有学者借助于由 BP 算法训练的人工神经网络实现共振峰频率变换[7]。但由于 BP 网络的隐层函数采用了 Sigmoid 函数,使得其对数据的分辨能力不高,学习算法的收敛速度都很慢,存在局部极小等缺点。而径向基函数(RBF)神经网络[8]是一种性能良好的前向神经网络,计算量少,学习速度较 BP 算法为快,在参数逼近和分类能力上均优于 BP 网络。但是径向基网络隐层的训练一般由 K - 均值聚类算法完成[9],而这类聚类算法对初值的选择比较敏感,因而可能收敛于局部最佳值,而不具备全局优化特性。遗传算法具有极强的搜索能力,可以在全局范围内进行寻优。本文将遗传算法用于 RBF 神经网络的训练过程,对普通话中的[a1][o1][e1][i2][u][yv]等六个单元音音素[10]作了说话人声音转换的实验性研究。

2 遗传 RBF 神经网络

2.1 RBF 神经网络

RBF 神经网络由 3 层组成,其结构如图 1 所示,输入层节点传入到隐层节点,权值固定为 1。隐含层由一组相同的径向基函数组成,网络的输出层是隐含层输出的简单线性表示。

一般径向基函数采取高斯核函数表示形式

$$R_i(x) = \exp\left\{-\frac{\| x - c_i \|^2}{2\sigma_i^2}\right\}, i = 1,2,\cdots,m \tag{1}$$

其中,x 是 n 维输入矢量,c_i 是第 i 个径向基函数的中心,与 x 具有相同维数的矢量,σ_i 代表一个半径范围的量,确定了对称式的隐层节点响应的宽度。m 就是隐层感知单元的个数。$\| x - c_i \|^2$ 表示矢量 $x - c_i$ 的欧几里德范数。从高斯核函数的形式可以看到,在 c_i 处,隐层节点的响应值达到最大,随着 $\| x - c_i \|$ 增大,$R_i(x)$ 迅速衰减为零。对于给定的输入 $x \in R^n$,只有一小部分靠近 c_i 的输入值被激活。因此 RBF 网络具有局部逼近能力。图 1 表明,输入层到隐层实现(1)式表示的非线性变换 $x \rightarrow R_i(x)$,而隐层到输出层是线性变换关系,表示如下:

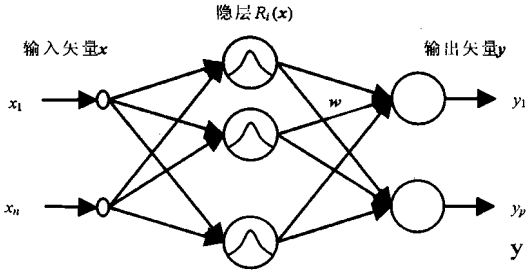

图 1 RBF 神经网络的拓扑结构

$$y_k = \sum_{i=1}^{m} w_{ik} R_i(x), k = 1,2,\cdots,p \tag{2}$$

其中,w 为输出层与隐层之间的联结权值,p 为输出矢量的维数。由上可见,RBF 神经网络是通过非线性基函数的线性组合实现非线性映射。RBF 网络中需要学习确定的参数有两类:一类是基函数中心 c_i 和宽度 σ_i,一般可采用 K-均值算法或 FCM 聚类算法来训练 c_i,基函数宽度的选取往往根据聚类的结果来确定。另一类为输出层与隐层之间的联结权值。在确定 c_i 和 σ_i 之后,权值的学习可采用通常的梯度下降法[9]或采用矩阵求伪逆的正则化方法[11]。基函数的中心和宽度的选取对 RBF 网络的性能产生巨大的影响。因此本文采用遗传算法优化径向基中心和宽度的选取。

2.2 用遗传算法训练 RBF 神经网络隐层

遗传算法是一类借助于生物界自然选择和自然遗传机制的随机化搜索算法。它是一种群体性操作,操作对象是群体中的所有个体,通过选择、交叉和变异等操作产生新一代群体。遗传聚类算法与 K-均值算法不同,由于采用群体搜索策略,使得初始聚类中心的选取对聚类结果影响不大,能达到全局最优解,因此使用遗传算法代替 K-均值等聚类算法训练网络隐层,能提高 RBF 网络的性能[12]。将遗传算法引入训练过程后,RBF 网络的训练算法总结如下:

(1)以聚类类别号作为基因值构造染色体和初始种群。

(2)参照 K-均值聚类算法,适应度函数取为总失真的倒数:

$$f = 1/J_v = \left[\sum_{k=1}^{K}\sum_{x_i \in \Omega_j} \| x_i - c_j \|^2\right]^{-1} \tag{3}$$

其中,c_i 为某聚类中心。

(3)进行选择、交叉和变异等遗传操作,直到算法收敛。

(4)将遗传聚类算法所得的每一类都作为隐层节点,分别求取径向基的中心和宽度:

$$c_j = \frac{1}{n_j}\sum_{x_i \in \Omega_j} x_i, \sigma_j^2 = \frac{1}{n_j}\sum_{x_i \in \Omega_j} \| x_i - c_j \|^2 \tag{4}$$

式中,n_j 是训练中聚于第 j 类的训练矢量个数。

(5)用梯度下降法训练网络隐含层与输出层之间的联结权值。

3 转换算法描述

一般地,声音转换技术可以分为训练和转换两个阶段进行。训练时,计算两个说话人语音样本的声学参数,估计由源说话人到目标说话人声学空间的映射规则(变换函数),用于捕获源语音和目标语音的模型参数之间的变换关系;在转换阶段,利用已训练好的转换函数对新输入源语音的声学特征进行变换,合成所期望目标说话人音色的语音。

3.1 语音频谱表示

声音转换用的频谱特征表示为线谱频(LSF)。因为线谱频与共振峰频率密切相关,而与共振峰频率参数相比,其参数可以鲁棒地估计得到,很容易由 LPC 参数多项式求出。由于人耳对声音低频段的分辨率更高,为使线谱频之间的数值距离更好地反映人耳的感官距离,这里将每一帧语音的线谱频由线性尺度映射到 Bark 尺度上,表示如下:

$$b(f) = 6.0\log\left\{\frac{f}{1200} + \sqrt{\left[\frac{f}{1200}\right]^2 + 1}\right\} \tag{5}$$

LPC 残差可由每一帧与 LSF 参数相对应的 LPC 参数经反滤波计算求得。

3.2 转换方法

RBF 神经网络声音转换方法如图 2 所示。在训练阶段,用前一部分所述的遗传算法训练 RBF 网络隐层的中心和宽度,用梯度下降法训练 RBF 网络输出层的联结权值。在训练集的源-目标说话人发出相同音素的每一个 LSF 特征矢量对中,源说话人频谱矢量作为输入,目标说话人频谱矢量作为输出目标。在转换阶段,测试集中源说话人语音被解析成每一帧为加 Hanning 窗半帧步长的 LSF 频谱包络和 LPC 残差信号。已训练好的 RBF 网络将输入 LSF 频谱做非线性映射产生转换频谱,与解析获得的残差信号结合,进行叠加操作,最后用 TD-PSOLA 技术使源语音平均 F0 和目标语音平均 F0 相匹配,其对应关系可表示如下:

$$f_0^c = \frac{\sigma_t}{\sigma_s} \cdot (f_0^s - \mu_s) + \mu_t \tag{6}$$

其中,μ 和 σ 分别为假设基频为高斯分布下的均值和标准差,s, t 和 c 分别表示源说话人、目标说话人语音和转换语音。

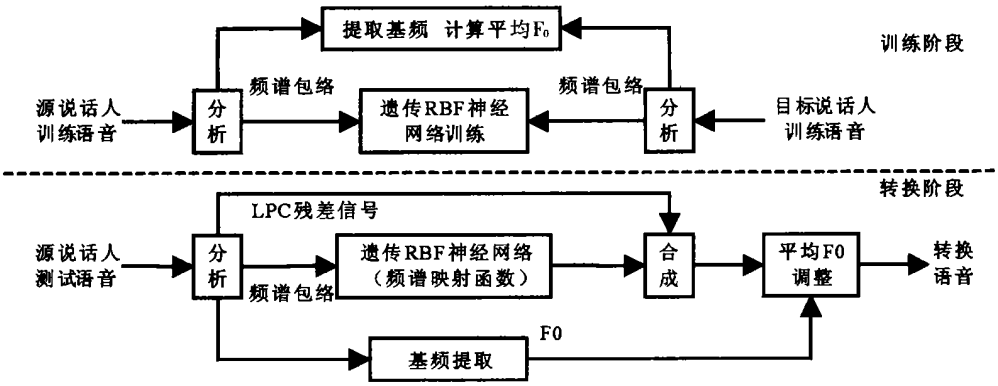

图 2 基于 RBF 神经网络的声音转换

4 实验研究

4.1 实验数据

这项研究选取普通话中的[a1][o1][e1][i2][u][yv]等六个单元音音素[10]作为考察对象。源说话人语音数据收集于连续语音中这些元音的稳态发音区域,目标说话人说出同样的语句,手工抽取与源说话人相对应的稳态语音区域。从源 - 目标说话人的对应稳态区域分别提取22阶LPC频谱矢量,计算出LSF频谱矢量,分别作为RBF网络的输入 - 输出矢量对。共提取约600个阴平声调的矢量对用于RBF网络的训练。从863数据库中选取四个人(男女声各两个)的声音作为元音音素采集的语料。

4.2 实验结果与分析

本文实验实现了男声到男声、女声到女声和女声到男声等多种声音变换过程,并分别对转换语音质量进行了客观和主观测试。说话人频谱距离的客观度量可用Itakura距离[13]表示:

$$d_I = \log\left\{\int_{-\pi}^{\pi} \frac{|A|^2}{|A_p|^2}\frac{dw}{2\pi}\right\} = \log\frac{a^t R_p a}{\sigma_p^2} \qquad (7)$$

式中,A为线性预测系数a所表示的逆滤波器,p表示比较时的参照语音(这里指目标说话人语音),R是自相关系数。

表1 转换语音频谱距离的客观测试结果

说话人组合	转换前	转换后
女声1—男声2	1.8579	0.2069
女声2—男声2	1.7702	0.1885
男声1—男声2	1.0973	0.1244
女声1—女声2	0.7963	0.1277

为了考察多说话人到某个特定说话人的转换语音性能,除了同性声音转换以外,实验还分别做了由女声1、女声2向男声2的转换。表1显示各种声音变换组合时源说话人语音分别在转换前和转换后与目标说话人语音的平均Itakura距离。可以看到,在女声到男声变换时,源说话人语音在转换后与目标语音的频谱距离与在转换前相比显著减小,而同性声音变换时,频谱距离减小幅度要小得多。相对于女声1,女声2的语音在转换后与男声2的频谱距离要小。

表2 转换语音的ABX性能测试(正确响应率)

说话人组合	K- 均值算法	遗传算法
男声1—男声2	77 %	79 %
女声1—女声2	76 %	76 %

实验用ABX方法对表1所示的转换语音做了主观评价。激励X表示转换语音,激励A和B或是源说话人语音或是目标说话人语音,每一个三元组就是这样三句语音的组合,对A和B中的哪一个与X的声音最相似做出选择。实验表明,女声向男声转换时,虽然转换语音与目标语音的听觉感知仍有差别,但是测试者采用ABX法均做出了正确的选择。而同性声音之间转换时,表2中数据表明ABX测试结果的满意度并不很高,只有近80 %的转换语音被认为与目标语音的说话人特性更接近。但听觉测试表明,与女声1和女声2相比,男声1的转换语音听起来更像是男声2的语音。比较表1和表2,转换语音的主观测试结果与客观评估结果存在差异,客观频谱距离的差别并不能说明一定存在相应的感官性能差别。由表2还可以看出,不同聚类算法产生的语音主观测试结果差别还不是很明显,从另一角度反映了客观度量和主观评价的差异性。

实验比较了K- 均值算法和遗传算法用于训练神经网络时对转换语音频谱所产生的影响。表3是分别用这两种算法生成的转换语音与目标语音之间的Itakura距离。采用遗传算法的声音变换,频谱失真距离平均比K- 均值算法减小达10 %左右。数值结果说明遗传聚类

算法所形成的聚类中心优于 K- 均值算法的聚类中心,由遗传算法训练网络产生的转换语音在频谱上与目标语音更为接近。而且,与表 1 中转换前频谱距离相比较,女声 1、女声 2 和男声 1 的语音经过两种算法转换后与同一目标语音男声 2 的频谱距离变化的趋势基本上是一致的。

表 3　两种聚类算法的语音频谱距离

说话人组合	K- 均值算法	遗传算法
女声 1→男声 2	0.2276	0.2069
女声 2→男声 2	0.2192	0.1885
男声 1→男声 2	0.1340	0.1244
女声 1→女声 2	0.1351	0.1277

图 3 画出了女声到男声变换时各元音音素频谱包络的一次实例。粗实线表示目标说话人语音频谱,细实线和虚线分别代表源说话人语音和转换语音。从这些音素的频谱曲线可以比较直观地看到,转换语音在很大程度上更接近于目标语音。

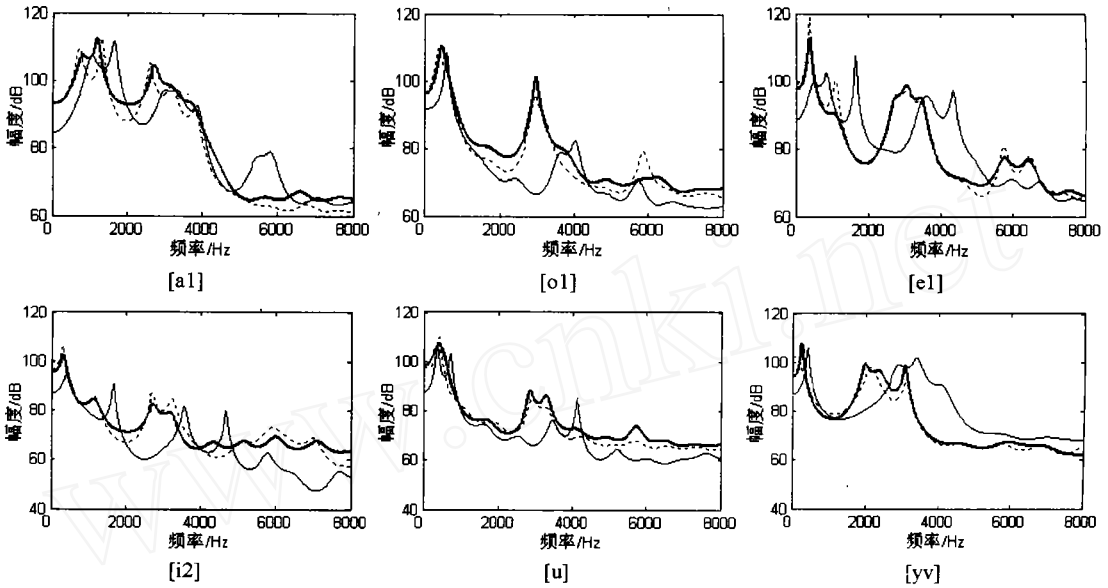

图 3　转换语音与源 - 目标语音的 LPC 频谱包络

5　结束语

本文探讨了神经网络方法在声音转换技术中的应用,转换语音质量的主观和客观测试结果均表明,径向基神经网络能够捕获不同说话人特性之间的非线性映射关系。试验从数值上说明将遗传聚类算法用于 RBF 网络隐层的训练过程,可在全局上进行搜索寻找最优解,减小了转换语音与目标语音的频谱距离。听辨实验表明,转换语音与目标语音的频谱度量并不一定与主观评估结果保持一致。在源 - 滤波器语音模型中,由 LPC 分析获得残差信号仍然含有一定的说话人特征信息,因此,要进一步增加转换语音声音特征的目标说话人倾向性,还需考虑残差信号对转换语音质量的作用。此外,在连续语音中,还要考虑声音转换技术中除语音频谱之外基频曲线、能量曲线和时长等超音段特征的影响。

参　考　文　献:

[1]　E. Moulines and Y. Sagisaka. Voice conversion: state of the art and perspectives [J]. Speech Communication, Elsevier, Feb. 1995, 16(2): 125 - 126.

[2] M. Abe, et al. Voice Conversion through Vector Quantization [A]. Proc. ICASSP [C], New York, USA, 1988(1): 655 - 658.
[3] H. Valbret, et al. Voice transformation using PSOLA technique [J]. Speech Communication, 1992, 11 (2 - 3): 175 - 187.
[4] Y. Stylianou, et al. Continuous Probabilistic Transform for Voice Conversion [J]. IEEE Transactions on Speech and Audio Processing, March 1998, 6(2): 131 - 142.
[5] A. Kain and M Macon. Spectral Voice Conversion for Text-to-Speech Synthesis [A]. Proc. ICASSP [C], Seattle, USA, May 1998(1): 285 - 288.
[6] T. Masuko, et al. Voice characteristics conversion for HMM-based speech synthesis system [J]. Proc. ICASSP [C], Munich, Germany, 1997: 1611 - 1614.
[7] M. Narendranath, H. A. Murthy, S. Rajendran and B. Yegnanarayana, Transformation of formants for voice conversion using artificial neural networks [J]. Speech Communication, 1995, 16(2): 207 - 216.
[8] T. Watanabe, et al. Transformation of Spectral Envelope for Voice Conversion Based on Radial Basis Function Networks [A]. Proc. ICSLP [C], Denver, USA, Sept. 2002: 285 - 288.
[9] J. Moody and C. J. Darken. Fast learning in networks of locally-tuned processing units [J]. Neural Comput., 1989 (1): 281 - 294.
[10] 祖漪清. 汉语连续语音数据库的语料设计[J]. 声学学报, 1999, 24(3): 236 - 247.
[11] S. Chen, et al. Orthogonal least squares learning algorithm for radial basis function networks [J]. IEEE Trans Neural Networks, 1991(2): 302 - 309.
[12] 岳喜才, 管桦, 叶大田. 说话人识别使用遗传 RBF 网络[J]. 应用声学, 19(2): 35 - 38, 2000.
[13] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition [M]. Prentice-Hall, Inc., Upper Saddle River, New Jersey, 1993.

[书讯]

1.《Advances in Computation of Oriental Languages》

该论文集中 2003 年 8 月 3—6 日在沈阳召开的"20 届东方语言的计算机处理国际会议"论文集,刊登了 81 篇高水平学术论文,由清华大学出版社正式出版。论文的主要内容是词法、语法和语义;工具和资源;机器翻译;基于内容的信息检索;字符集、文档图象分析和 OCR 后处理及语音识别和 TTS 等。

书价:150 元

2.《语言计算与基于内容的文本处理》

该论文集中 2003 年 8 月 9—11 日在哈尔滨召开的"全国第七届计算语言学联合学术会议(JSCL—2003)"论文集,由清华大学出版社正式出版。论文集中刊登了 90 篇论文及 4 篇大会特邀报告。本次会议是对我国计算语言学研究领域最新成果的检阅和水平的显示,开得也非常成功,必将对促进我国计算语言学的研究和应用产生积极的影响。

书价:100 元

以上两种论文集存数不多,欲购从速。

联系方式:中国中文信息学会办公室　电话:010 - 62562916　Email:cips@admin.iscas.ac.cn