ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 30,No. 2,Mar. 2015,pp. 336－343
DOI:10.16337/j.1004-9037.2015.02.011

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于自适应粒子群优化径向基函数神经网络的语音转换

张玲华[1]　姚绍芹[2]　解伟超[2]

(1.南京邮电大学物联网学院,南京,210003;2.南京邮电大学通信与信息工程学院,南京,210003)

摘　要:语音转换是指在保持源说话人语义内容不变的前提下,通过改变源说话人的个性特征,使其听起来像目标说话人的语音。本文提出一种自适应粒子群优化算法训练径向基函数神经网络进行语音特征建模,以获取说话人谱包络的映射关系;此外,考虑到说话人谱包络参数与基频有着密切的联系,利用基于径向基函数神经网络的联合谱包络基频变换方法,将谱包络参数与基频联合进行建模和转换,使得转换后的基频含有更多的说话人个性特征。最后,运用主、客观方法对获得的转换语音进行性能测试。实验表明,与主流的基于高斯混合模型的语音转换相比,使用自适应粒子群优化的径向基函数神经网络方法能够获得更好的转换性能,且更加适用于男声到女声的转换。

关键词:语音转换;径向基函数神经网络;自适应粒子群优化;高斯混合模型;基频

中图分类号:TN912.3　　**文献标志码**:A

Voice Conversion Based on Adaptive Particle Swarm Optimization Radial Basis Function Neural Network

Zhang Linghua[1], Yao Shaoqin[2], Xie Weichao[2]

(1. College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China; 2. College of Telecommunication & Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003, China)

Abstract: Voice conversion is a technique for changing the personality characteristics of a source speaker's voice into the target speaker's, while preserving the original semantic information. An adaptive particle swarm optimization (PSO) based method is proposed to model voice features by training the radial basis function (RBF) neural network in order to capture the spectral envelope mapping function between speakers. In addition, the pitch transformation is captured by modeling pitch with the joint spectral feature parameters in RBF neural network, which makes the converted pitch contain more target details. Finally, the performance of the improved voice conversion system is tested by subjective and objective method respectively. Experimental results show that the performance of the proposed method is better than that of the Gaussian mixture model (GMM) based system, especially for the male to female conversion.

Key words: voice conversion; radial basis function neural network; adaptive particle swarm optimization; gaussian mixture model; pitch

---

基金项目:江苏省高校自然科学研究重大(13KJA510003)资助项目;江苏高校优势学科建设工程(PAPD)资助项目;江苏省普通高校研究生科研创新计划(CXLX12_0478)资助项目。
收稿日期:2015-01-15;**修订日期**:2015-01-28

引 言

语音转换技术[1]是指在保持源说话人语义信息不变的情况下,改变源说话人个性特征,使之具有目标说话人的个性特征。语音转换技术主要分为两个阶段:训练阶段和转换阶段。训练阶段的主要任务是提取源说话人和目标说话人的语音特征参数,并将提取的这些特征参数对齐,然后通过训练得到转换规则。转换阶段的主要任务是提取源说话人的语音特征参数,再根据训练阶段得到的转换规则对提取的语音特征参数进行转换得到转换的语音特征参数,最后根据转换的语音特征参数合成出语音[2]。

语音转换中经常使用的模型主要有:矢量量化法(Vector quantization, VQ)[3]、模糊矢量量化法(Fuzzy vector quantization,FVQ)[4]、高斯混合模型(Gaussian mixture model, GMM)[5-7]、人工神经网络(Artificial neural network, ANN)[8-9]等。矢量量化法在语音转换的早期被广泛采用,然而其缺点在于基于硬聚类的原理以及离散映射导致转换的语音不连续。模糊矢量量化虽然避免了由于硬聚类带来的转换语音性能降低的问题,但是离散映射的现象依然存在。Stylianou 等[5]和 Kain 等[6]分别提出了基于 GMM 模型的语音转换方法,该模型是基于软聚类和连续映射的,其转换性能优于早先的语音转换方法,是目前语音转换中最常用的一种方法,但是存在的过平滑和过拟合现象成为影响转换语音质量的关键因素。此外,基于 ANN 模型[8]的语音转换可以获得与 GMM 模型相似的转换性能,同时能很好地表征不同说话人之间声道特性的非线性关系。在 ANN 模型中,径向基函数(Radial basis function, RBF)神经网络[10]具有计算量少、结构简单、参数逼近能力强及学习速度快等优点,因此,本文就 RBF 神经网络对语音转换进行研究。针对传统方法训练径向基函数神经网络时存在的收敛速度慢、易陷入局部最优、泛化性能不佳等问题,考虑引入自适应粒子群优化算法(Particle swarm optimization, PSO)来训练 RBF 神经网络以获取网络的连接权,并研究其在语音转换系统中的性能。此外,现有的基音频率变换算法中将基频与谱包络参数分开转换,导致基频特征中说话人的个性特征被部分丢失,为了保持更多的说话人个性特征,本文提出 RBF 神经网络的联合谱包络基频变换算法。该算法利用 RBF 神经网络在谱包络参数与基频之间建立联系,使得转换出的基频能够更好地跟踪目标基频的变化,并含有更多的目标说话人个性特征的细节。

1 特征参数提取

一个完整的语音转换系统主要解决以下两个问题:提取特征参数以及建立语音转换规则。首先,提取的特征参数必须能够较好地描述说话人的个性特征。在提取了源说话人和目标说话人的特征参数之后,需要对其进行训练,以建立两者之间的转换规则。

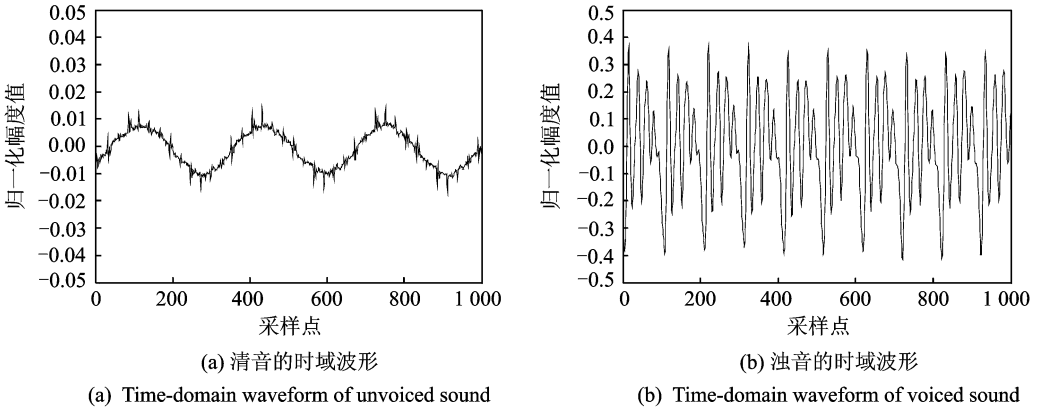
线谱频率(Linear spectrum frequency, LSF)[11-12]能够很好地表征共振峰的位置和带宽,具有良好的插值特征,并且特征参数的某一部分失真只会影响合成谱参数的一部分区域等特点,使得 LSF 成为描述音段信息的最优选择。此外,LSF 特征维数与音段的关系是至关重要的,有关研究表明[13],低阶(4~8 阶)的线性预测体现说话人的语言信息,而高阶(>12 阶)的线性预测分析不仅能够获得语言信息,还能得到特定说话人的信息,所以本文采用 16 维 LSF 特征参数。

在对语音信号进行特征提取时,针对语音的不同类别采取不同的处理。根据激励方式,可将语音划分为清音、浊音和爆破音,本文将爆破音作为浊音进行处理,图 1 所示为清音和浊音的时域波形。

从清音和浊音的时域波形图可看出,浊音帧是具有短时周期性的,且短时能量比较大,而清音帧则不具备短时周期性,且短时能量较小,一般可将其当作白噪声处理。因此,只对浊音帧提取 LSF 特征参数,用于训练和转换,清音帧保留不变,不进行提取 LSF 特征参数的处理。

LSF 特征提取的技术细节如下:

(1)使用 STRAIGHT 模型[14]对语音信号进行谱分析。



(a) 清音的时域波形 (b) 浊音的时域波形

(a) Time-domain waveform of unvoiced sound (b) Time-domain waveform of voiced sound

图 1 清浊音的时域波形图

Fig. 1 Time-domain waveform of voiced and unvoiced sound

(2)对浊音帧的谱进行快速傅里叶反变换得到自相关系数。

(3)利用 Levinson-Durbin 算法由自相关系数求取 LPC(Linear predictive coefficient)系数。

(4)采用 Chebyshev 多项式求根法将 LPC 系数转换成 LSF 参数。

2 本文算法描述

语音转换系统包括两个部分:训练部分和转换部分。本文提出的转换系统框图如图 2 所示,其中,STRAIGHT 模型用于提取语音的频谱特征参数和基频,以及合成转换语音;动态时间规整(Dynamic time warping, DTW)用于对源、目标特征参数的对齐;频谱的转换采用基于自适应 PSO 的 RBF 神经网络,基频的转换采用 RBF 神经网络[11]。

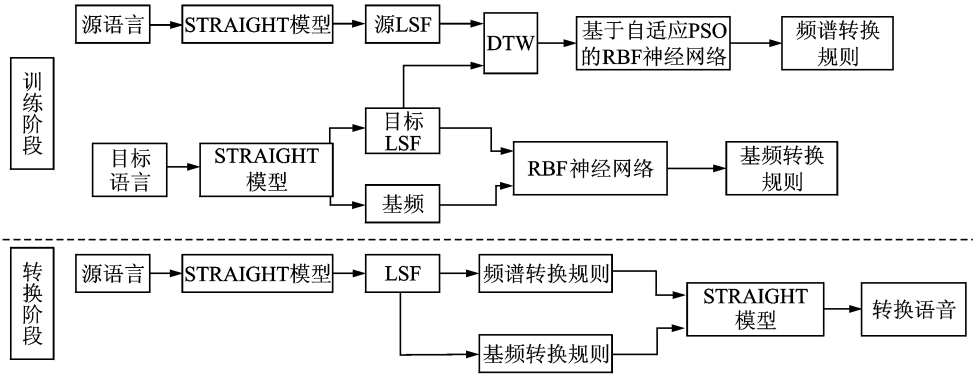


图 2 语音转换框图

Fig. 2 Block diagram of voice conversion

2.1 RBF 神经网络结构

本文采用的 RBF 神经网络是一种性能良好的 3 层前向神经网络,包括输入层、输出层和一个隐层,如图 3 所示。

在 RBF 神经网络中,隐层一般采用高斯函数的径向基函数,形式如下

$$\varphi_i(\boldsymbol{x}) = \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{c}_i\|^2}{2\sigma_i^2}\right) \qquad 1 \leqslant i \leqslant N \tag{1}$$

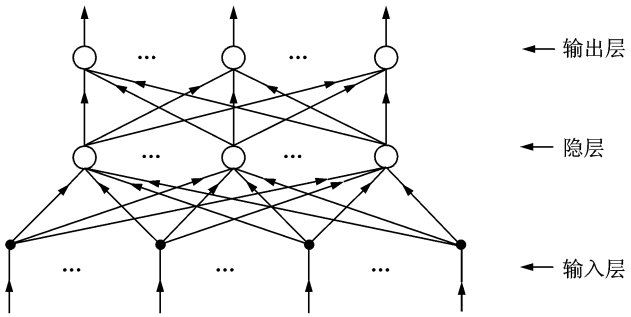


图 3 RBF 神经网络结构图

Fig. 3 Structure diagram of RBF neural network

式中:$\boldsymbol{c}_i$,σ_i 分别表示第 i 个高斯函数的中心和宽度;$\boldsymbol{x}$ 为输入向量;N 为隐层径向基函数的个数;$\| \boldsymbol{x} - \boldsymbol{c}_i \|^2$ 表示输入向量 $\boldsymbol{x}$ 与 $\boldsymbol{c}_i$ 之间欧几里德范数。

RBF 神经网络的实际输出为

$$\boldsymbol{y}_j = \sum_{i=1}^{N} w_{ij} \varphi_i(\boldsymbol{x}) = \sum_{i=1}^{N} w_{ij} \exp\left(-\frac{\| \boldsymbol{x} - \boldsymbol{c}_i \|^2}{2\sigma_i^2}\right) \qquad 1 \leqslant j \leqslant m \tag{2}$$

式中:w_{ij} 为隐层与输出层之间的连接权值;m 为输出矢量的维数。

RBF 神经网络的参数有 3 个,分别为径向基函数中心 c_i、径向基函数宽度 σ_i 以及隐层与输出层间的权值 w_{ij}。对于中心和带宽两个参数采用基本的基于最小均方误差准则的方法获取,而连接权参数本文采用自适应 PSO 算法训练 RBF 网络来确定。

2.2 自适应 PSO 算法

粒子群优化(PSO)算法[15-16]是一种基于群智能的随机全局优化计算方法。该算法以其建模简单、易于实现且收敛速度快等优点被广泛应用于函数优化、神经网络训练、模糊系统控制等领域。

假设有 N 个粒子组成一个粒子群,粒子 i 在 D 维空间中的位置表示为矢量 $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{iD})^{\mathrm{T}}$,飞行速度表示为矢量 $\boldsymbol{v}_i = (v_{i1}, v_{i2}, \cdots, v_{iD})^{\mathrm{T}}$,在每一次迭代中,粒子通过跟踪两个极值来更新自己,第 1 个极值是个体极值,第 2 个是全局极值,在找到这两个最优值后,粒子更新自己的速度和位置分别为

$$v_{id} = w \cdot v_{id} + c_1 \cdot r_{1d} \cdot (p_{id} - x_{id}) + c_2 \cdot r_{2d} \cdot (p_{gd} - x_{id}) \qquad i = 1,2,\cdots,N;\ d = 1,2,\cdots,D \tag{3}$$

$$x_{id} = x_{id} + v_{id} \qquad i = 1,2,\cdots,N;\ d = 1,2,\cdots,D \tag{4}$$

式中:c_1,c_2 为加速因子;r_{1d},r_{2d} 为介于 0 和 1 之间的随机数;p_{id},p_{gd} 分别为粒子本身到目前为止发现的最优位置,即个体极值和整个粒子群中所有粒子发现的最优位置,即全局极值;w 为惯性因子,w 较大时,具有较强的全局搜索能力,反之,当 w 较小时,具有较强的局部搜索能力。为了平衡全局和局部的搜索能力,在每次迭代中,本文采用动态 w

$$w = w_{\max} - \frac{w_{\max} - w_{\min}}{\mathrm{iter}_{\max}} \cdot \mathrm{iter} \tag{5}$$

式中:$w_{\max}$,$w_{\min}$ 分别表示初始惯性权值与迭代至最大次数时的惯性权值,$\mathrm{iter}_{\max}$ 是最大迭代次数,iter 是当前迭代次数。最后,当达到最大迭代次数或是微粒群迄今为止搜索到的最优位置满足最小适应阈值时,PSO 算法结束。

2.3 基于自适应 PSO 算法的 RBF 神经网络权值参数流程

本文将 RBF 神经网络的权值 w_{ij} 看做是 PSO 算法的粒子,利用自适应 PSO 算法训练 RBF 神经网

络的过程如下：

(1) 初始化粒子群，主要包括设定粒子群大小，初始每个粒子的位置、速度、惯性因子的上、下限值，最大迭代次数等。

(2) 根据适应度函数计算得到每个粒子的个体最佳适应度值和全局最佳适应度值，本文采用的粒子适应度函数为

$$f=\sum_{i}^{D}\sum_{j}^{n}(y_{ij}-y'_{ij})^{2} \tag{6}$$

式中：D 为输出节点的个数，即样本的维数；n 为训练样本的个数；y_{ij} 为样本的期望输出；y'_{ij} 为网络的实际输出。

(3) 开始进行迭代计算，通过式(3,4)来调整每个粒子的速度和位置。

(4) 对调整过后的粒子重新进行适应度的计算，若粒子当前的最佳适应度值优于粒子的个体最佳适应度值，则用当前的适应度值去替代个体最佳适应度值，然后将新得到的个体最佳适应度值与全局最佳适应度值进行比较，若优于后者，则用新得到的个体最佳适应度值去替代全局最佳适应度值。

(5) 判定是否满足终止条件，此处的终止条件是达到最大迭代次数或是得到的全局最佳适应度值小于设定的阈值，若满足，则转向第(6)步，否则，转向第(3)步。

(6) 将通过自适应 PSO 算法得到的全局最优值作为 RBF 神经网络的权值。

3 基频转换

在过去大量的语音转换系统中，主要只考虑了频谱特征的转换，但是，韵律特征，尤其是基频 F0 的转换对于高质量的语音转换系统也是至关重要的。考虑到谱包络特征与基频的相关性[17]以及特征之间的非线性，本文将采用基于 RBF 神经网络的联合谱包络特征参数的基频转换[11]。具体的实现过程为：

(1) 对进行训练的目标语音进行分帧处理并进行清浊音判别，然后对浊音帧利用 STRAIGHT 模型逐帧提取出语谱参数和基音频率。

(2) 对语谱参数进行降维及转换，首先对每一帧的语谱参数进行快速傅里叶逆变换(Inverse fast Fourier transform，IFFT)得到自相关系数，再由自相关系数利用 Levinson-Durbin 算法得到自回归(Autoregressive，AR)参数，该 AR 参数就是该帧语音的 LPC 系数，最后由 LPC 系数转换得到 LSF 参数，本次实验中 LSF 参数的维数定为 16。

(3) 将提取出的每一帧浊音的 LSF 和基频分别作为 RBF 神经网络的输入和输出，同时采用高斯函数作为 RBF 神经网络的径向基函数，以建立基频和频谱特征参数间的映射关系。

(4) 利用第(3)步中的映射关系对转换阶段获得的频谱特征参数 LSF 进行转换，得到目标基频。

4 实验分析

本文主要采样主观和客观实验对提出的语音转换算法的性能进行评估。主观评价主要采用 MOS (Mean opinion score)分和 ABX 实验进行，客观测试则考虑频谱失真。

实验参数设置如下：粒子群大小为 30，加速因子 $c_1=c_2=2$，惯性因子的上下限值分别为 0.9 和 0.4，最大迭代次数为 2 000。

4.1 数据库

本文所采用的数据库是在我校的消声实验室中录制的纯净语音[18]，采样频率为 16 kHz，并以16 bit 量化，选择其中的两名男子和两名女子的发音，其每个人以标准普通话朗读 120 句，并且每个人朗读的

内容是相同的,其中,60 句用作训练集,60 句用作测试集。

4.2 主观评价

主观评价中采用了 MOS 分和 ABX 两种实验。

在 MOS 分的测试中,本文采用 10 个经验丰富的测评人对转换的语音进行 5 个等级的打分,这 5 个等级为:差、比较差、一般、好和很好,分别对应 1～5 分。测评结果如表 1 所示。

表 1 MOS 分测试结果

Table 1 Result of MOS test

转换类型	女-男	男-女	男-男	女-女
GMM	2.4	2.5	2.6	2.7
RBF	2.4	2.5	2.6	2.8
RBF+PSO	2.6	2.9	3.0	3.1
RBF+PSO+F0	2.8	3.0	3.1	3.2

从表 1 可以看出,GMM 模型与传统的 RBF 模型的转换性能相当;同性间由于个性差异比较小,所以转换的性能要高于异性间的转换性能。本文提出的利用自适应 PSO 算法优化 RBF 神经网络进行频谱转换的性能要明显好于传统的基于最小均方误差准则确定 RBF 神经网络参数的方法。

ABX 实验是另外一种主观评价转换语音质量的方法。它主要是用来衡量转换语音与目标语音相似度的,其中 A 和 B 分别表示源说话人和目标说话人的语音,X 指的是转换出来的语音。在该实验中,通过 10 位经验丰富的测评人对转换出的 20 句进行评价,从 A 和 B 中选出与 X 相近的语音。测试结果如表 2 所示。

表 2 ABX 测试结果(正确率)

Table 2 Result of ABX test %

转换类型	男-男	男-女	女-女	女-男
GMM	35.5	42.7	36.4	41.3
RBF	35.4	42.1	36.2	41.5
RBF+PSO	46.7	55.3	48.4	51.2
RBF+PSO+F0	49.6	56.7	52.1	54.9

从表 2 中可以看出,本文提出的方法明显优于基本的 RBF 神经网络,且在男声转成女声的过程中尤为突出;另外,异性间转换的语音比同性间转换的语音更加接近目标语音。

4.3 客观评价

本文的客观评价主要通过频谱失真来完成

$$\mathrm{SD}=\frac{1}{n}\sum_{i=1}^{n}\frac{\|\boldsymbol{y}_i-\boldsymbol{y}'_i\|}{\|\boldsymbol{y}_i-\boldsymbol{x}_i\|} \tag{7}$$

式中:n 为语音的帧数;$\boldsymbol{x}_i$,$\boldsymbol{y}_i$ 和 $\boldsymbol{y}'_i$ 分别为源 LSF 矢量、目标 LSF 矢量和转换的 LSF 矢量。图 4,5 分别给出了女声到男声转换和男声到女声转换的频谱失真图。

从图 4,5 可以看出,由 GMM 方法与 RBF 神经网络方法进行语音转换得到的谱失真大体相当。但是,与传统的基于 GMM 的方法相比,本文提出的基于自适应粒子群优化 RBF 神经网络的语音转换方法能够明显改善转换语音的质量;如果同时考虑基音频率,则可使转换语音的质量得到进一步提高,特别是对女声到男声的转换,在 40 帧之后改进效果较为明显。

5 结束语

GMM 是主流的语音转换方法,但是鉴于 ANN 的转换性能与 GMM 的转换性能相当[8],而 ANN 方

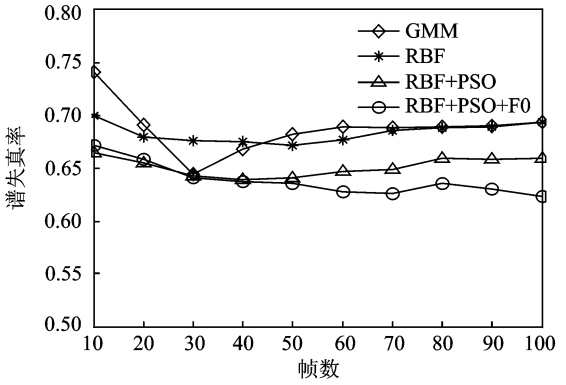


图 4 语音的频谱失真图(女声-男声)

Fig. 4 Spectral distortion (F-M)

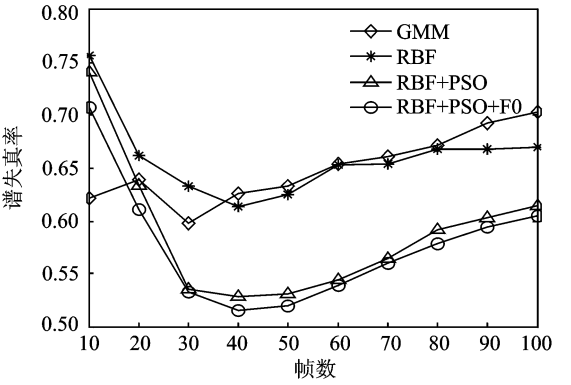


图 5 语音的频谱失真图(男声-女声)

Fig. 5 Spectral distortion (M-F)

法又能反映出特征参数间的非线性关系,因此本文采用 RBF 神经网络作为转换模型。在获取 RBF 神经网络的模型参数过程中,本文提出利用自适应 PSO 算法来确定 RBF 神经网络的权值参数,以获取说话人谱包络的映射关系;此外,考虑到基频的转换对于高性能语音转换系统的重要性,文中采用了基于 RBF 神经网络的联合谱包络基频变换方法,将谱包络参数与基频联合进行建模和转换,使得转换后的基频含有更多的说话人个性特征。主客观测试表明,相对主流的 GMM 模型和传统的神经网络,本文提出的基于 PSO 算法的 RBF 神经网络以及考虑基频的语音转换系统具有更加优越的性能,特别是对男声到女声的转换效果明显。

参考文献:

[1] Lee K S. A unit selection approach for voice transformation[J]. Speech Communication, 2014, 60: 30-43.

[2] Xu Ning, Tang Yibing, Bao Jingyi, et al. Voice conversion based on Gaussian process by coherent and asymmetric training with limited training data[J]. Speech Communication, 2014, 58: 124-138.

[3] Abe M, Nakamura S, Shikano K, et al. Voice conversion through vector quantization[C]// Proc IEEE International Conference on Acoustics, Speech and Signal Processing. New Jersey: IEEE Press, 1988: 655-658.

[4] Shikano K, Nakamura S, Abe M. Speaker adaptation and voice conversion by codebook mapping[C]// IEEE International Symposium on Circuits and Systems. New Jersey: IEEE Press, 1991: 594-597.

[5] Stylianou Y, Cappé O, Moulines E. Continuous probabilistic transform for voice conversion[J]. IEEE Transactions on Speech and Audio Processing, 1998, 6(2): 131-142.

[6] Kain A, Macon M W. Spectral voice conversion for text-to-speech synthesis[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. New Jersey: IEEE Press, 1998: 285-288.

[7] 姚绍芹,张玲华. 基于 GMM 和 ANN 混合模型的语音转换方法[J]. 数据采集与处理,2014,29(2):227-231。
Yao Shaoqin, Zhang Linghua. Voice conversion based on mixed GMM-ANN model[J]. Journal of Data Acquisition and Processing, 2014, 29(2): 227-231.

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

[9] Narendranath M, Murthy H A, Rajendran S, et al. Transformation of formants for voice conversion using artificial neural networks[J]. Speech Communication, 1995, 16: 207-216.

[10] Niros A D, Tsekouras G E. A novel training algorithm for RBF neural network using a hybrid fuzzy clustering approach[J]. Fuzzy Sets and Systems, 2012, 193: 62-84.

[11] 解伟超,张玲华. 语音转换中声道谱参数和基频变换算法的研究[D]. 南京:南京邮电大学,2013.
Xie Weichao, Zhang Linghua. The research on vocal tract spectrum and pitch frequency transformation in voice conversion [D]. Nanjing: Nanjing University of Posts and Telecommunications, 2013.

[12] Ye H, Young S. Quality-enhanced voice morphing using maximum likelihood transformations[J]. IEEE Transactions on Audio,Speech and Language Processing, 2006, 14(4): 1301-1312.

[13] Misra H S, Yegnanarayana B. Speaker-specific mapping for text-independent speaker recognition[J]. Speech Communication, 2003, 39:301-310.

[14] Kawahara H, Masuda-Katsuse I, Cheveigné A de. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction:Possible role of a repetitive structure in sounds[J]. Speech Communication, 1999, 27: 187-207.

[15] 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.
Guo Tong, Lan Julong, Li Yufeng, et al. Network traffic prediction with radial basis function neural network based on quantum adaptive particle swarm optimization[J]. Journal of Electronics & Information Technology, 2013, 35(9): 2220-2226.

[16] Kennedy J, Eberhart R C. Particle swarm optimization[C]// IEEE International Conference on Neural Networks. New Jersey: IEEE Press, 1995: 1942-1948.

[17] Shao X, Milner B. Pitch prediction from MFCC vectors for speech reconstruction[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. New Jersey: IEEE Press, 2004: 97-100.

[18] 徐宁,杨震,张玲华. 基于状态空间模型的子频带语音转换算法[J]. 电子学报,2010,38(3):646-653.
Xu Ning, Yang Zhen, Zhang Linghua. Sub-band voice morphing algorithm based on state-space model[J]. Acta Electronica Sinica, 2010, 38(3): 646-653.

作者简介:**张玲华**(1964-),女,教授,博士生导师,研究方向:语音信号处理与语音通信、智能信号处理和无线通信中的信号处理,E-mail:zhanglh@njupt.edu.cn;**姚绍芹**(1988-),女,硕士研究生,研究方向:语音信号处理;**解伟超**(1987-),男,硕士研究生,研究方向:语音信号处理。