# 基于 SVR 的语音变换映射规则研究

崔丽珍[①] 宋 巍[②]
[①](内蒙古科技大学信息工程学院 包头 014010)
[②](解放军理工大学通信工程学院 南京 210007)

**摘 要**：该文介绍了语音变换与支持向量回归(SVR)的基本理论。提出了基于多输出支持向量回归的语音变换特征参数映射规则，并对该映射规则进行了仿真实验。对变换后语音所进行的主客观测试表明，该映射规则对比码书映射和高斯混合模型，能够在参数映射离散性和平滑性之间有效折中，提高语音可懂度。

**关键词**：语音变换；映射规则；支持向量回归
**中图分类号**：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2008)05-1144-04

# The Algorithm of Voice Transformation Based on SVR

Cui Li-zhen[①] Song Wei[②]
[①](*Information Engineering School Inner Mongolia University of Science & Techology*, *Baotou* 014010, *China*)
[②](*Institute of Communication Engineering PLA University of Science & Technology*, *Nanjing* 210007, *China*)

**Abstract**: In this paper, the fundamental theory of voice transformation and Support Vector Regression(SVR) are introduced. The algorithm of voice transformation technology based on multi-output SVR is put forward, and the corresponding simulation experiment is carried out. The subjective test and objective test show that compared with code book mapping and GMM, this mapping rule performs well between discreteness and smoothness of parameter.

**Key words**: Voice transformation; Mapping rule; Support Vector Regression

## 1 引言

语音中所包含的信息，可以分为两个部分：文本信息和讲话人信息。文本信息指的是语音中包含的可用文本再现的语言文字信息，讲话人信息是指语音中的可用来区分不同讲话人的信息。语音变换所要解决的问题就是要保持语音中的文本信息不变，变换其中的讲话人信息，使语音在听觉感知上趋向于其他人的声音。语音变换技术不仅在伪装通信、冒名通信及迷惑通信等方面有重要的应用价值，而且对语音信号处理的其它技术(例如语音合成、语音增强等方面)有很大的促进作用。

近年来，语音变换技术的研究已经成为语音信号处理研究中的一个前沿方向和热点。通常，语音变换系统由 3 个部分组成：语音特征参数、特征映射规则及语音库。其中，特征映射规则是语音变换系统的核心部分，它决定源说话人语音特征参数到目标说话人语音特征参数映射性能的好坏[1,2]。国内外已经有研究人员对其进行了探讨，提出了一些特征映射规则，主要有码书映射(Code Book Mapping，CBM)[3]、高斯混合模型(Gaussian Mixture Model, GMM)[4]等。码书映射离散性较大，容易造成变换后语音频谱不连续；GMM 连续性较好，有时也容易造成频谱过度光滑。为了改进上述问题，本文提出使用多输出的支持向量回归作为映射规则，对汉语语音的 15 个单韵母进行训练和变换，并同其它方法进行了对比研究，实验表明本算法能够较好地在两者间折中。

2007-02-05 收到，2007-07-30 改回

## 2 语音变换基本原理

### 2.1 语音的个性特征参数

要实现语音变换技术，首先要分析语音的个性特征即说话人特征。解决好究竟要变换什么的问题。表征语音的个性特征主要有 3 类：(1)音段特征：包括共振峰位置、共振峰带宽、频谱斜率、基音频率、能量等等。(2)超音段特征：包括音素的时长、音调等等。(3)语言学特征：包括习惯用语、方言、口音等等[5]。

从人的发音机理来分析，表征语音的个性特征又可分为：(1)线性特征：主要指说话人的声道特征。(2)非线性特征：主要指说话人的声门特征和激励特征等等。

现在报道的相关语音变换技术文献所采用的语音个性特征参数，主要有以下几类： (1)韵律模型[6, 7]；(2)线性预测(LPC)及其延伸模型[3, 8]；(3)正弦谐波模型[9]等。

由于线谱对(LSF)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。因此本文使用线谱对 LSF 作为变换用特征

参数。

### 2.2 语音变换系统结构

通常，语音变换的实现分为训练和变换两个阶段。在训练阶段，利用训练语音进行语音的个性特征建模，并寻找到一种特征转换规则。在变换阶段，利用训练得到的转换规则对源说话人的语音进行特征的变换，从而得到变换后的语音。语音变换系统结构如图 1 所示。

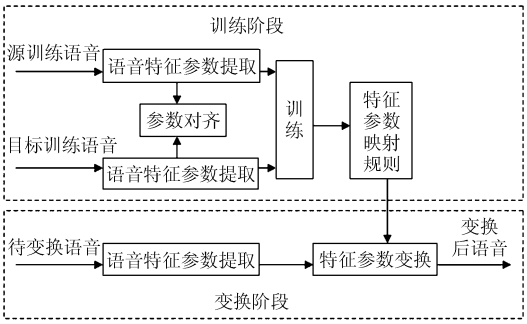


图 1 语音变换系统结构图

## 3 基于 SVR 的特征参数映射规则

### 3.1 SVR 基本原理

基于结构风险最小化原理的统计学习理论，使泛化误差的上限最小化，而经验风险最小化使相对于训练数据的误差最小化。SVR 最初是针对分类问题提出来的，应用到回归问题即为支持向量回归。

设训练集为

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_k,y_k)\}\subset R^n\times R \tag{1}$$

输入 $x_i$ 和输出 $y_i$ 之间存在某种函数映射关系，用映射 $\Phi$ 把输入 $x_i$ 所在的空间映射到一个高维特征空间，然后在高维特征空间中进行线性回归，设回归函数为

$$f(x)=\omega\cdot\Phi(x)+b \tag{2}$$

引入惩罚因子 $\varepsilon$ 和松弛因子 $\xi_i$， $\xi_i^*$

上述回归问题可归结为规划：

$$\mathrm{Min}\ \frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{k}(\xi_i+\xi_i^*) \tag{3}$$

$$\text{s.t.}\ \ f(x_i)-y_i\le\xi_i^*+\varepsilon \tag{4}$$

$$y_i-f_i(x)\le\xi_i+\varepsilon \tag{5}$$

$$\xi_i,\xi_i^*\ge 0,\ \ i=1,\cdots,k \tag{6}$$

利用拉格郎日函数和对偶原理，可得到其对偶问题为

$$\mathrm{Min}\ \begin{aligned}&\frac{1}{2}\sum_{i=1}^{k}\sum_{j=1}^{k}(\alpha_i-\alpha_i^*)^{\mathrm{T}}K(x_i,x_j)(\alpha_i-\alpha_i^*)\\&-\sum_{i=1}^{k}y_i(\alpha_i-\alpha_i^*)+\varepsilon\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)\end{aligned} \tag{7}$$

$$\text{s.t.}\ \ \sum_{i=1}^{k}(\alpha_i-\alpha_i^*)\ge 0,\ \ i=1,\cdots,k \tag{8}$$

$$0\le\alpha_i,\alpha_i^*\le C,\ \ i=1,\cdots,k \tag{9}$$

其中 $\alpha_i,\alpha_i^*$ 为拉格郎日乘子， T 表示为转置。

解此二次规划可得到 $\alpha_i,\alpha_i^*$ 以及回归函数

$$f(x)=\sum_{i=1}^{k}(\alpha_i-\alpha_i^*)K(x_i,x)+b \tag{10}$$

利用 KKT(Karush-Kuhn-Tucker)条件可以计算出常值偏差 $b$ 。

$K(x_i,x_j)=\left\langle\Phi(x_i),\Phi(x_j)\right\rangle$ 是核函数，本文采用类神经网络核函数：

$$K(x_i,x_j)=\tanh(\kappa x_i\cdot x_j-\delta) \tag{11}$$

### 3.2 SVR 映射规则

由 SVR 基本原理可知，SVR 是多输入单输出的。而语音变换中的特征参数是多维的，即特征参数由源说话人到目标说话人的映射是向量到向量之间的映射，使用 SVR 进行多对多的映射有两种方式：分组映射法和分段映射法。

设对 LSF 系数进行聚类后，某类源说话人和目标说话人的训练特征参数集分别为

$$\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\} \tag{12}$$

$$\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_N\} \tag{13}$$

其中 $\boldsymbol{x}_i=\{x_{i1},x_{i2},\cdots,x_{ip}\}$， $\boldsymbol{y}_i=\{y_{i1},y_{i2},\cdots,y_{ip}\}$。

这里使用的是通过 DTW 对齐之后的参数，源讲话人和目标讲话人的语音参数帧都为 $N$ 帧， $p$ 为 LSF 系数的阶数。

SVR 的训练集为

$$\boldsymbol{T}=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),(\boldsymbol{x}_2,\boldsymbol{y}_2),\cdots,(\boldsymbol{x}_N,\boldsymbol{y}_N)\}\subset R^p\times R^p \tag{14}$$

(1)分组映射法　为特征参数建立 $p$ 个回归函数，分别为

$$f_m(x)=\omega_m\cdot\Phi(x)+b_m,\quad m=1,2,\cdots,p \tag{15}$$

第 $m$ 个回归的输出为 $Y$ 的第 $m$ 维特征参数。建立 $p$ 组训练集

$$\boldsymbol{T}_m=\{(\boldsymbol{x}_1,y_{1m}),(\boldsymbol{x}_2,y_{2m}),\cdots,(\boldsymbol{x}_N,y_{Nm})\}\subset R^p\times R,\quad m=1,2,\cdots,p \tag{16}$$

在训练阶段，使用 $p$ 组训练集分别对 $p$ 个支持向量机进行回归训练，得到 $p$ 组带核函数的回归函数：

$$f_m(x)=\sum_{i=1}^{k}(\alpha_{mi}-\alpha_{mi}^*)K(x_i,x)+b_m,\ \ m=1,2,\cdots,p \tag{17}$$

在变换阶段，将待变换的源说话人某帧的 LSF 系数聚类后，根据其类别依次送入该类的 $p$ 个支持向量机，得到一组新的 $p$ 维数据，即变换后的特征参数。

(2)分段映射法　为特征向量只建立一个回归函数

$$f(x)=\omega\cdot\Phi(x)+b \tag{18}$$

其训练集为

$$\boldsymbol{T}=\{(\boldsymbol{x}'_{nm},y_{nm})\,|\,1\le n\le N,1\le m\le p\} \tag{19}$$

其中 $\boldsymbol{x}'_{nm}=[\boldsymbol{x}_n,m]$。本来输入参数为 $P$ 维，增加一维输入参数 $m$，表示当前输入参数所对应的输出参数 $y_{nm}$ 的所在的维数，即在输入参数最后一维的不同数值段上实现对输出参数多维的映射。

## 4 变换算法实现

### 4.1 语音特征参数选择及提取

由于线谱对 LSF 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。因此本文使用线谱对 LSF 作为变换用特征参数，维数为 12。具体流程如下：

(1)对训练语音和变换语音进行预处理，包括去除直流分量、消噪等。

(2)语音端点检测及清浊音判决，方法为短时能量结合短时平均过零率算法[10]。

(3)对清音部分，不做处理，浊音部分进行基音周期检测，标记基音周期点，使用基于短时平均幅度差函数的基音周期检测算法(AMDF)。

(4)对浊音部分按基音周期标记点分帧，每隔 2 个基音标记点分为一帧，帧移为一个基音标记点。

(5)按帧提取 LPC 系数并将之转换为 LSF 系数，若为训练语音则建立残差码书。

### 4.2 训练特征参数对齐

由于对齐语音特征参数的目的是使语音信号在文本信息上对应一致，减小在对齐过程中使讲话人信息对对齐效果的影响。有文献指出，语音信号的倒谱系数进行语音识别时效果较好，本文使用倒谱系数作为对齐参数。

当前在语音识别和讲话人识别中参数对齐的算法主要有 DTW(动态时间规整)和 HMM(隐马尔可夫模型)两种方法。由于 HMM 的训练需要完善的语音库，并且训练过程十分复杂，而 DTW 对于小数据量的数据也有比较好的参数对齐效果，因此本文使用 DTW 作为对齐算法。

### 4.3 基音周期的建模与变换

在对 LSF 参数进行动态时间归整时，不光建立 LSF 参数训练集，同时也建立基音周期训练集。基音周期训练集的元素为 LSF 参数对应帧的基音周期，元素个数对应 LSF 参数训练集的元素个数。

为基音周期建立一个支持向量回归函数，对其基音周期训练集进行训练。回归函数输入为源说话人的基音周期，输出为目标说话人的基音周期。

### 4.4 参数变换及语音合成

按 3.2 节的步骤对源讲话人待变换语音进行分帧，并提取每帧的 LSF 系数。将各帧 LSF 系数输入训练好的 12 组支持向量回归函数，得到新的 LSF 系数，将之转换为 LPC 系数，在残差码书中寻找与之相对应的残差码字，合成为新一帧语音。将生成的各帧语音连同源待变换语音的非浊音段按时间顺序进行拼接，同一浊音段内语音采用 PSOLA 算法叠加，即得到变换后语音。

## 5 仿真实验及结果

实验所用语音库文本内容为中国汉语拼音方案中的 15 个单韵母(如表 1 所示)。训练语音和变换语音均为在实验室条件下录制，8000Hz 采样，16bit 存储的 wav 文件。

本文对采用 CBM，GMM 模型和 SVR 3 种映射规则产生的语音进行了主客观测试，并进行了对比研究，支持向量回归的核取径向基核函数。

**表 1 实验所用 15 个单韵母**

| | | | | |
|---|---|---|---|---|
| a | o | e | i | u |
| v | ui | ei | ao | ou |
| an | en | ang | eng | ong |

客观测试标准使用线性预测编码倒谱系数距离测度——LPC-CD。LPC 倒谱能够极好地逼近语音信号产生的声道传输模型。以线性预测模型参数(即 LPC 倒谱系数)距离 LPC- CD 作为客观失真测度是合理的。它定义为[11]

$$D_l = \sqrt{\left[C_x(l,0) - C_y(l,0)\right]^2 + 2\sum_{k=1}^{P-1}\left[C_x(l,k) - C_y(l,k)\right]^2} \quad (20)$$

式中：$C_x(l,:)$ 和 $C_y(l,:)$ 分别是第 $l$ 帧的变换语音和对照语音线性预测谱的倒谱系数；$P$ 是预测器的阶数，取 $P$=12；取各帧平均值作为比较参考。使用目标说话人的同文本信息的语音作为对照语音进行评测。

**表 2 LPC-CD 比较表**

| 方法 | CBM | GMM | SVR1 | SVR2 |
|---|---|---|---|---|
| LPC-CD | 0.052 | 0.067 | 0.059 | 0.065 |

表 2 中 SVR1 表示分组映射法，SVR2 表示分段映射法。从表 2 中可以看出，CBM 与目标语音的 LPC-CD 最小，主要是由于 CBM 直接利用目标语音的 LSF 码字，而对于参数转换而言，SVR 要略优于 GMM，SVR 分组映射法优于 SVR 分段映射法。

对变换后语音进行主观听觉测试，参加测试人员均为具有足够知识和理解能力的语音处理研究人员。对 15 个单韵母的语音变换进行了 ABX 测试，结果如表 3 所示，从表中可以看出，使用 SVR 映射规则的合成的可懂度有显著提高。在自然度测试中，将变换语音按照两两对比进行测试，测试人员一致认为使用 GMM 为变换规则的语音自然度好，SVR 略次之，CBM 最差。

**表 3 ABX 测试结果比较表**

| 方法 | CBM | GMM | SVR |
|---|---|---|---|
| ABX 值 | 81% | 90% | 93% |

## 6 结束语

本文提出了基于 SVR 的语音变换特征映射规则，解决了语音变换特征映射中参数映射离散性和过度平滑之间的折中问题。在同 CBM 和 GMM 两种映射规则比较的仿真实验表明，SVR 解决了 CBM 离散性的问题，同时 GMM 中由于参数过度平滑造成文本信息丢失的现象在 SVR 中得到了一定的控制。

## 参 考 文 献


[1] Kain A. High resolution voice transformation [D]. Illinois: Rockford College, 2001.

[2] 李波，王成友，蔡宣平等. 语音变换及相关技术综述[J]. 通信学报, 2004, 25(5): 109–118.
Li Bo, Wang Cheng-you, and Cai Xuan-ping. A survey of voice conversion and its relevant technology. *Journal of China Institute of Communications*, 2004, 25(5): 109–118.

[3] Arslan L M and Talkin D. Voice conversion by codebook mapping of line spectral frequencies and excitation spectrum[A]. Proceedings of the EUROSPEECH[C]. Rhodes, Greece, 1997: 1347–1350.

[4] Kain A and Macon M W. Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction [J]. IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, 2001, 2: 813–816.

[5] 左国玉, 刘文举, 阮晓刚. 声音转换技术的研究与进展[J].电子学报, 2004, 32(7): 1165–1172.
Zuo Guo-yu, Liu Wen-ju, and Ruan Xiao-gang. Voice conversion technology and its development[J]. *Acta Electronica Sinica*, 2004, 32(7): 1165–1172.

[6] 王浩, 刘春林, 岳振军. 一种基于基音周期调整的简单语音变换技术[J]. 军事通信技术, 2004, 25(4): 1–5.
Wang Hao, Liu Chun-lin, and Yue Zhen-jun. A simple voice technology based on Pitch adjustment transform[J]. *Military Communications Technology*, 2004, 25(4): 1–5.

[7] Inanoglu Z. Transforming pitch in a voice conversion framework[D]. Cambridge: College University of Cabridge, 2003.

[8] Chen Y and Chu M, *et al.*. Voice conversion with smoothed GMM and MAP a daptation [A]. Proc Eurospeech[C]. Geneve, Switzerla-nd: ISCA Sept. 2003: 2413–2416.

[9] 岳振军，王浩，张雄伟. 基于正弦谐波模型和 BP 神经网络的语音变换算法及实现[J]. 信号处理, 2005, 21(4): 208–211.
Yue Zhen-jun, Wang Hao, and Zhang Xiong-wei. Based on the harmonic sinusoidal model and BP neural network algorithm speech and implementation[J]. *Signal Processing*, 2005, 21(4): 208–211.

[10] 张雄伟, 陈亮, 杨吉斌. 现代语音处理技术及应用[M]. 北京: 机械工业出版社, 2003: 19–34.
Zhang Xiong-wei, Chen Liang, and Yang Ji-bin. Modern Speech Processing Technology and Application[M]. Machinery Industry Press, 2003: 19–34.

[11] 黄惠明, 王瑛, 赵思伟等. 语音系统客观音质评价研究[J].电子学报, 2000, 28(4): 112–114.
Huang Hui-ming, Wang Ying, and Zhao Si-wei. Study of objective quality evaluation for the speech systems [J]. *Acta Electronica Sinica*, 2000, 28(4): 112–114.



崔丽珍：　女，1968 年生，副教授，研究方向为 DSP 应用技术、通信技术.

宋　巍：　男，1982 年生，研究生，研究方向为语音信号处理.