文章编号: 1006- 9348(2010) 12- 0376- 05

一种改进的基于 DCT变换的语音增强算法

李 潇,李 宏
(西北工业大学,陕西 西安 710129)

摘要: 研究语音识别问题,变换域分析是语音增强中最常用的方法,采用离散余弦变换(DCT)来描述带噪语音信号的频域特性,并构造拉普拉斯-高斯参数模型(Laplacian- Gaussian)来表示带噪语音信号的分布,可改善增强效果并减少误差。在参数估计上采用了最大似然估计法(ML),并结合谱相减估计法对拉普拉斯模型参数作了进一步优化,最后给出了检测语音信号存在的条件。仿真实验结果表明,算法在用于处理含有 F16、Pink、Babble及高斯白噪声的语音信号时,与其它基于 DCT变换的算法相比,取得了更好的增强效果。

关键词: 语音增强; 离散余弦变换; 拉普拉斯-高斯模型; 最大似然估计
中图分类号: TP391 文献标识码: A

An Improved Speech Enhancement Algorithm Based on DCT

LI Xiao, LI Hong
(Northwestern Polytechnical University, Xi'an Shanxi 710129, China)

ABSTRACT: Transform domain method is mostly used in speech enhancement. In this paper, we apply the DCT to develop an effective expression of the frequency character of the input signal, and build a noisy speech signal as the Laplacian- Gaussian model. The ML is used to estimate the model parameters, then the Laplacian model parameter is optimized by the spectral subtraction estimation. At last, the uncertainty of speech signal presence is estimated. The experimental results show that our method yields better performance than the conventional DCT- based algorithms in processing speech signal with the F16, Pink, Babble and Gaussian noise.

KEYWORDS: Speech enhancement; DCT; Laplacian- Gaussian; ML

1 引言

语音增强多用于语音编码以及语音识别的前处理阶段,其主要目的就是减少周围环境噪声的影响,提高语音的可理解性以及减少听者的听觉疲劳。语音增强有时域语音增强和变换域语音增强,本文主要考虑变换域语音增强。常见的变换域有 DFT(Discrete Fourier Transform)、DCT(Discrete Cosine Transform)、DST(Discrete sine Transform)以及 KLT(Karhunen Loeve Transform)等。DFT因具有较好的信号分离特性,在语音增强处理中得到广泛应用。而 DCT、DST以及 KLT变换能量压缩特性较好,最初用于图像的压缩处理之中;最近研究表明这些变换在语音增强应用中也能取得较好的效果[1,2],甚至比 DFT更具优势。KLT作为一种最优变换,能量压缩效果最好,但由于目前没有相应的快速运算算法而限制了它的应用,而 DCT作为一种效果接近于 KLT变换的次优变换,又有相应的快速运算算法,因而得到广泛应用。DCT在语音增强中,相对 DFT的优势有以下几点[2]:

1) DCT变换具备比 DFT更好的时频能量压缩性质。

2)在窗长相同的情况下, DCT变换具有更高的频谱分辨率(例如窗长为偶数 N, DCT变换将有 N个独立的的成分,而 DFT是一种复变换,仅有 N/2+ 1个独立的成分)。

3) DFT对语音信号的处理主要是从幅度考虑,在目前的相位估计中通常把带噪语音的相位作为增强语音的相位[3],这与实际有一定误差。而 DCT变化是一种实变换,变换的结果只有正值和负值两种,考虑到当噪声能量很小时,变换后得到的结果并不足以改变幅度的符号,当噪声能量相对较大时,经过增强后的信号幅度将得到很大程度的衰减,这样就进一步减少了失真。

在基于模型的语音增强估计上,实际语音及噪声信号的时域模型通常被视为符合高斯-高斯混合分布,然而实际情况是,语音成分更接近于超高斯分布(Laplacian或 Gamma),而且经 DCT或 KLT变换后的语音模型也能更好的适合超高斯分布,而噪声成分仍符合高斯分布[4,5,6],故本文采用

收稿日期: 2009- 11- 05 修回日期: 2010- 04- 08

Laplacian-Gaussian混合作为带噪语音信号的模型分布,为了进一步改善增强效果并减少误差,对文献[7]中语音信号的 Laplacian模型进行了修正,引用了 Cohen, Martin等人提出的 Laplacian参数模型[8,9],将对语音幅度均值的估计转化为方差的估计,而语音与噪声信号彼此独立,方差可通过谱相减法来进行估计。

在参数估计上,MMSE是常用的参数估计算法,然而由于基于 Laplacian-Gaussian混合的 MMSE 估计是一种非线性估计,计算复杂,故可寻求其它方法,本文采用最大似然函数法(ML)进行参数估计,因 ML容易实现参数估计,而估计效果又与 MMSE近似相同[7]。

在语音的时域分布上,由于实际的语音信号会有许多间断的无声段,不可能在时域一直存在,因此对语音在时域存在的可能性概率分布条件进行了判定。

2 系统模型

2.1 DCT变换模型

对于一组 N 点序列信号 $x(n)$,其中 $0\leqslant n\leqslant N-1$, DCT变换为

$$X(k)=u(k)\sum_{n=0}^{N-1}x(n)\cos\left[\frac{\pi(2n+1)k}{2N}\right] \tag{1}$$

其中, $u(k)=\begin{cases}\sqrt{\frac{1}{N}} & \text{当 } k=0\\ \sqrt{\frac{2}{N}} & 1\leqslant k\leqslant N-1\end{cases}$

其逆变换为

$$x(k)=\sum_{k=0}^{N-1}u(k)X(k)\cos\left[\frac{\pi(2n+1)k}{2N}\right] \tag{2}$$

2.2 语音短时谱分析

语音信号严格意义上是一种非平稳信号,在直接处理时会比较困难,但发音在很短的一段时间内由于惯性的缘故声道特征近似不变,语音近似为平稳的,短时间隔一般为 5-30ms。以下分析均为语音信号的短时分析(short time, ST)。

设语音信号 $x(t)$及噪声信号 $v(t)$经 A/D转换采样后得到的混合信号是

$$y(n)=x(n)+v(n) \tag{3}$$

其中 $y(n), x(n), v(n)$分别为带噪语音信号、纯净语音信号及噪声信号。

则语音信号的 STDCT为:

$$Y(i,k)=X(i,k)+V(i,k) \tag{4}$$

其中 $Y(i,k), X(i,k), V(i,k)$分别为 STDCT后的带噪语音,纯净语音及噪声部分。 i为语音信号分帧后时间帧指数, k为每个时间帧内的频数,对于一个 N点 STDCT,则 $0\leqslant k\leqslant N-1$。在分帧考虑的情况下,一般可以简写为:

$$Y(k)=X(k)+V(k) \tag{5}$$

3 参数估计

3.1 噪声估计

噪声的 STDCT变换域模型服从 Gaussian分布,其概率密度函数(PDF)为

$$f[V(k)]=\frac{1}{\sqrt{2\pi\sigma_v^2(k)}}\exp[-V^2(k)/2\sigma_v^2(k)] \tag{6}$$

其中 $\sigma_v^2(k)$是噪声 $V(k)$的方差,一般噪声可视为一个零均值、独立的各态历经过程,通过采用 VAD(Voice Activation Detection)判决出语音的无声段即噪声段,则在此噪声段, $\sigma_v^2(k)$可采用下列方法近似估计:

$$\sigma_v^2(k)=\frac{1}{N}\sum_{k=0}^{N-1}|V(k)|^2 \tag{7}$$

为便于表达,以下 $\sigma_v^2(k)$仍用 $\sigma_v^2(k)$来代替。

3.2 语音信号模型

语音信号在 DCT变换后仍服从 Laplacian分布,对于 Laplacian模型的表达式,文献[7]采用了如下的表达方式:

$$f[X(k)]=\frac{1}{2a(k)}\exp[-X(k)/a(k)] \tag{8}$$

其中 $a(k)$是 Laplacian系数因子,考虑到语音信号可视为一个零均值的各态历经过程,通过 DCT的去相关变换后, $a(k)$可由最大似然法近似估计:

$$\hat{a}(k)=\frac{1}{N}\sum_{k=0}^{N}|X(k)| \tag{9}$$

可以看出, $a(k)$相当于 $X(k)$幅度的均值,由于无法获得纯净语音,因此一般情况下用带噪语音的幅度近似替代,这虽然简化了算法,但在某些区域会增大误差,而 $X(k)$幅值估计的准确程度,将会最终影响语音的增强效果[7]。为减少误差,引用了文献[8,9]中采用的 Laplacian模型:

$$f[X(k)]=\frac{1}{\sqrt{2\sigma_x^2(k)}}\exp(-\sqrt{2}|X(k)-\bar{X}|/\sqrt{\sigma_x^2(k)}) \tag{10}$$

其中 $\bar{X}$为幅度均值,考虑到语音信号是一个零均值过程,故可进一步表达为:

$$f[X(k)]=\frac{1}{\sqrt{2\sigma_x^2(k)}}\exp(-\sqrt{2}|X(k)|/\sqrt{\sigma_x^2(k)}) \tag{11}$$

$\sigma_x^2(k)$为 $X(k)$的方差, k为帧频数。

对于一 DCT域信号 $Y(k)=X(k)+V(k)$,噪声信号与语音信号是不相关并且零均值,则均方误差值可表达为:

$$\sigma_y^2(k)=\sigma_x^2(k)+\sigma_v^2(k) \tag{12}$$

由于均方值均为非负值,可采用幅度相减则有

$$\sqrt{\sigma_x^2(k)}=\sqrt{\sigma_y^2(k)-\sigma_v^2(k)} \tag{13}$$

考虑到实际一些样本点

$$d=\sigma_y^2(k)-\sigma_v^2(k)\leqslant 0$$

故作如下修正:

$$d=|\sigma_y^2(k)-\sigma_v^2(k)| \tag{14}$$

3.3 纯净语音信号的估计

对于 DCT域混合语音信号, 语音信号 $X(k)$服从零均值的 Laplacian分布, 噪声信号 V(k)服从 Gaussian分布, 而且语音信号与噪声信号相互独立, 由其各自的 PDF:

$$f[X(k)] = \frac{1}{\sqrt{2\sigma_x^2(k)}}\exp(-\sqrt{2}|X(k)|/\sqrt{\sigma_x^2(k)}) \quad (15)$$

$$f(V(k)) = \frac{1}{\sqrt{2\pi\sigma_v^2(k)}}\exp(-V^2(k)/2\sigma_v^2(k)) \quad (16)$$

由 $Y(k)=X(k)+V(k)$ 得

$$V(k) = Y(k) - X(k) \quad (17)$$

则关于 $Y(k)$, $X(k)$的联合 $PDFf(X, Y)$为:

$$f(X, Y) = \frac{1}{2\sqrt{\pi\sigma_x^2(k)\sigma_v^2(k)}}\exp\{-\frac{\sqrt{2}|X(k)|}{\sqrt{\sigma_x^2(k)}} - \frac{[Y(k)-X(k)]^2}{2\sigma_v^2(k)}\} \quad (18)$$

$X(k)$关于 $Y(k)$的条件概率分布密度为:

$$f(X|Y) = \frac{f(X, Y)}{f(Y)} = \frac{f(X, Y)}{\int_{-\infty}^{+\infty} f(X, Y)dX} \quad (19)$$

因此关于 $X(k)$的 ML估计为:

$$\hat{X} = \arg\max_x f(X|Y) = \arg\max_x f(X, Y)$$
$$= \arg\min_x\{\frac{\sqrt{2}|X(k)|}{\sqrt{\sigma_x^2(k)}} + \frac{[Y(k)-X(k)]^2}{2\sigma_v^2(k)}\} \quad (20)$$

通过对 $X(k)$求导后计算可得

$$X(k) = Y(k) \pm \frac{2\sqrt{2}\sigma_v^2(k)}{\sqrt{\sigma_x^2(k)}}$$

令 $d(k)=\frac{2\sqrt{2}\sigma_v^2(k)}{\sqrt{\sigma_x^2(k)}}$ 则忽略 k后,

$$\hat{X} = \begin{cases} Y-d & \text{当 } Y \geqslant d \\ Y+d & \text{当 } Y \leqslant -d \\ 0 & \text{其它} \end{cases} \quad (21)$$

4 语音存在域判决

以上语音信号的估计假定语音与噪声是同时存在的, 然而实际的语音信号存在着周期性的间隔, 因此也须对纯净语音在时间轴上的存在与否进行判决。假如用 H_0代表无声段, H_1代表语音与噪声共存段, 则最终关于语音信号 $X(k)$的估计应为:

$$X' = \hat{X}P(H_1|Y(k)) \quad (22)$$

其中 $P(H_1|Y(k))$为在给定 $Y(k)$情况下语音段的条件概率, 有全概公式及贝叶斯公式:

$$P(H_1|Y(k)) = \frac{P(H_1)P(Y(k)|H_1)}{P(H_0)P(Y(k)|H_0)+P(H_1)P(Y(k)|H_1)} \quad (23)$$

其中 $P(H_1)$, $P(H_0)$分别为语音段及非语音段的概率, 取 $P(H_0)=P(H_1)=0.5$, $P(Y(k)|H_1)$, $P(Y(k)|H_0)$为在这两种条件下的概率分布。

则

$$P(Y(k)|H_0) = \frac{1}{\sqrt{2\pi\sigma_v^2(k)}}\exp[-\frac{(Y(k))^2}{2\pi\sigma_v^2(k)}] \quad (24)$$

$$P(Y(k)|H_1) = \int_{-\infty}^{+\infty} f_x[Y(k)-V(k)]f_v(V(k))dV(k)$$
$$= \frac{\exp[2\sigma_v^2(k)/\sigma_x^2(k)]}{2\sqrt{\sigma_x^2(k)}}\{\exp[-\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}}] + \exp[\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}}] + \exp[-\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}}]\cdot erf[\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}} - \frac{\sqrt{2\sigma_v^2(k)}}{\sqrt{\sigma_x^2(k)}}] - \exp[\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}}]\cdot erf[\frac{\sqrt{2}Y(k)}{\sqrt{\sigma_x^2(k)}} + \frac{\sqrt{2\sigma_v^2(k)}}{\sqrt{\sigma_x^2(k)}}]\} \quad (25)$$

5 仿真实验结果分析

噪声信号来源于 NOISEX-92数据库, 采样频率为 16kHz, 采用 256点汉明窗分帧处理, 采用 1/4的帧移。采用的数据库噪声信号为 Gaussian白噪声, F16噪声, Pink噪声及 Babble噪声, 各种噪声的增强效果图如图 1-4(其中纵轴为语音信号单位幅度, 横轴代表采样点数)。

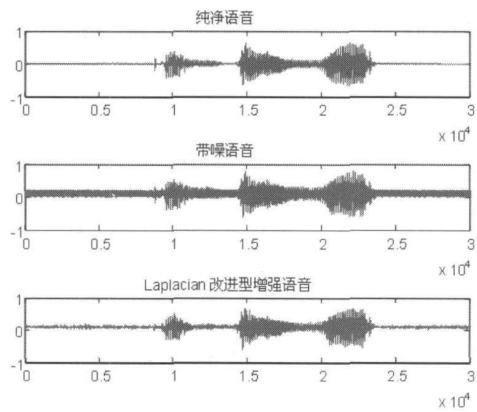

图 1 SNR = -2.47dB 高斯白噪声信号增强效果图

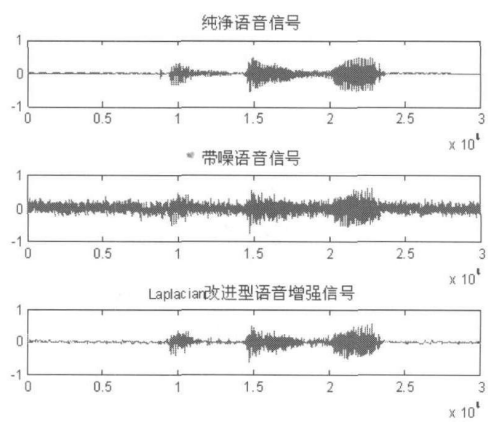

图 2 SNR = -2.05dB F16 噪声信号增强效果图

图 1中高斯白噪声成分得到一定程度的抑制, 但也出现了失真现象; 在处理图 2以及下面的图 3、图 4中的 F16、Pink 及 Babble噪声成分时, 增强效果相对比较明显, 不仅使噪声成分得到较大程度的抑制, 而且增强后的语音能够较好的符合原音信号。

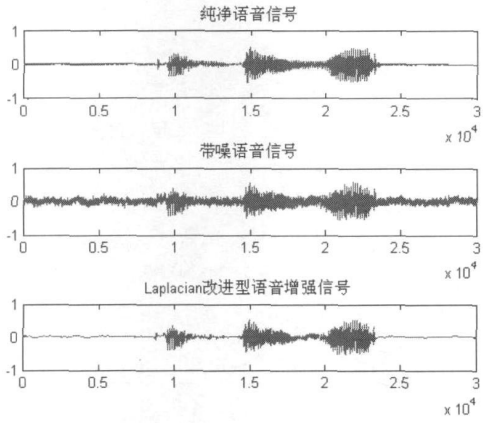

图 3　SNR = 8.3dB Pink 噪声信号增强效果图

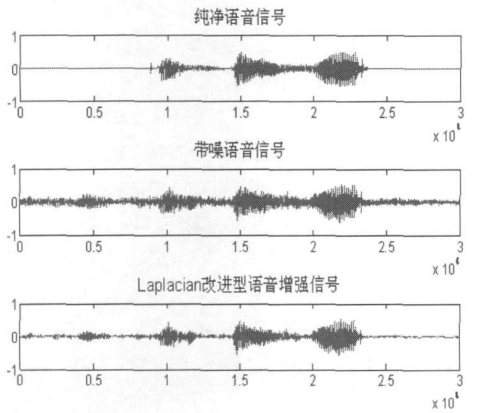

图 4　SNR = -1.83dB Babble 噪声信号增强效果图

为了进一步体现该算法具有的优点, 现拿此改进型与 Gaussian- Gaussian混合模型, Laplacian- Gaussian模型就输出信噪比这一方面作一对比, 输出信噪比采用总体信噪比来衡量:

$$SNR = 10\log_{10}[\sum_{n=1}^{N}x^2(n)/\sum_{n=1}^{N}(y(n)-x(n)^2] \quad (26)$$

所得结果如表 1- 表 4所示 (为便于比较, 各种算法分别简写为 G- G, L- G及本文所采用的 L改进型)。

表 1　**Gaussian**处理后输出信噪比比较

输入 SNR(dB)	G- G	L- G	L改进型
- 10.5	18.56	13.4	14.82
- 2.5	20.88	17.23	18.25
11.3	26.55	24.9	26.55

表 2　**F16**处理后输出信噪比比较

输入 SNR(dB)	G- G	L- G	L改进型
- 2	10.11	9.08	12.87
6	17.22	15.36	18.92
20	29.47	27.46	30.05

表 3　**Pink**处理后输出信噪比比较

输入 SNR(dB)	G- G	L- G	L改进型
- 1.8	11.02	9.4	13.7
8.3	26.79	24.65	28.15
16.5	26.79	28.15	24.65

表 4　**Babble**处理后输出信噪比比较

输入 SNR(dB)	G- G	L- G	L改进型
- 7.5	7.25	5.55	9.22
- 1.8	11.94	9.3	13
6.3	18.46	15.36	18.52

由以上各表可以看出, 在处理高斯白噪声时, 语音信号的信噪比改善相对稍差, 但在处理其他几类噪声信号时, 尤其在低信噪比下处理 F16及 Babble噪声时, Laplacian改进型语音的增强效果相对其他两类算法优势特别明显。

6　结论

本文基于对 DCT能取得比 DFT更好的增强效果的理解, 采用了 DCT变换, 在 DCT变换域纯净语音采用了 Laplacian模型, 而噪声仍采用 Gaussian分布, 在对信号的估计中采用了 ML估计算法, 在信号估计准确性基本不变的前提下简化了算法, 为了减少估计误差, 采用了改进的 Laplacian模型, 通过对模型因子中的幅度均值采用方差来代替, 再结合谱相减法进行估计, 在一定条件下提高了准确度。仿真实验结果表明, 该算法不仅可以处理常见的高斯白噪声, 而且在处理其他噪声如 F16、Pink及 Babble等噪声上也体现出很好的增强效果, 在现实中具有一定的应用价值。

参考文献:

[1] Chang Joon- Hyuk. Warped Discrete Cosine Transform- Based Noisy Speech Enhancement [J]. IEEE Transactionson Circuits and Systems- II: Express Briefs, September 2005, 52(9): 535- 539.

[2] Sooning Yann, Koh Soo Ngee, Yeo Chai Kiat. Noisy speech enhancement using discrete cosine transform[J]. Speech Communication, June 1998, 24(3): 249- 257.

[3] Y Ephraim, D Mala. Speech enhancement using a minimum mean- square error short- time spectral amplitude estimator[C]. IEEE Trans Acoust Speech Signal Process, ASSP- 32, 1984. 1109-

1121.
[4] S Gazor, W Zhang. Speech probability distribution[J]. IEEE Signal Processing Lett, Jul 2003, 10(7): 204-207.
[5] JH Chang, N S Kim. Speech enhancement using warped discrete cosine transform[C]. in Proc. IEEE Speech Coding Workshop, Tsukuba, Japan, Oct 2002.
[6] C Beithaupt, R Martin. MMSE estima- tion of magnitude-squared DFT coefficients with superGaussian priors[C]. in Proc IEEE Int Conf Acoustics, Speech, and Signal Processing, ICASSP'03, Apr 2003, 1: 896-899.
[7] Gazor Saeed, Zhang Wei. Speech enhancement employing laplacian gaussian mixture[C]. IEEE Transactions on Speech and Audio Processing, Sep. 2005, 13(5): 896-904.
[8] Cohen Israel. Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation[J]. Speech Communication, November 2005, 47(3): 336-350.
[9] Martin, Rainer. Speech enhancement based on minimum mean-square error estimation and supergaussian priors[C]. IEEE Transactions on Speech and Audio Processing, September 2005, 13(5): 845-856.

[作者简介]

李　潇(1981-),男(汉族),河南省周口市人,硕士研究生,主要研究领域为信号与信息处理,信息对抗技术,语音信号处理等。
李　宏(1958-),男(汉族),陕西省人,教授,硕士研究生导师,主要研究领域为统计与自适应信号处理、盲信号处理、电路优化等方向。

(上接第166页)

从预报结果来看,神经网络集合预报模型的预报平均绝对误差明显小于逐步回归相应的结果,显示了其优良性。然而从预报过程来看,在神经网络模型输入数据的处理方法上,核主成分的选取除了考虑其与预报量的相关关系、累积方差贡献,还应考虑其它哪些因素还值得作进一步的深入研究。

参考文献:

[1] 王诗文. 国家气象中心台风数值模式的改进及其应用试验[J]. 应用气象学报, 1999, 10(3): 347-353.
[2] 杨平章. 作用于台风系统的动力—热力因子分析[J]. 气象科学, 2000, 20(3): 348-353.
[3] 金龙,等. 南海西行台风强度的一种客观预报新方法[J]. 应用基础与工程科学学报, 2006, 12(14): 63-69.
[4] 金龙,等. 基于遗传算法的神经网络短期气候预测模型[J]. 高原气象, 2005, 24(6): 981-987.
[5] 吴建生,金龙,汪灵枝. 遗传算法进化设计BP神经网络气象预报建模研究[J]. 热带气象学报, 2006, 22(4): 411-416.
[6] 冯利华,骆高远. 基于模型叠加方法的登陆台风强度预报[J]. 海洋学报, 2001, 23(1): 127-132.
[7] L D Davis. Handbook of Genetic Algoithms[M]. Van Nostrand Reinhold, 1991.
[8] 周明,孙树栋. 遗传算法原理及应用[M]. 北京: 国防工业出版社, 1999. 18-59.
[9] Long Jin, Cai Yao, Xiao-Yan Huang. A Nonlinear Artificial Intelligence Ensemble Prediction Model for Typhoon Intensity[J]. Monthly Weather Review, 2008, 136: 4541-4554.
[10] 金龙. 神经网络气象预报建模理论方法与应用[M]. 北京: 气象出版社, 2004.
[11] 金龙,罗莹,李永华. 长期天气的人工神经网络混合预报模型研究[J]. 系统工程学报, 2003, 118(4): 331-336.
[12] 夏国恩,金炜东,张葛祥. 非线性主成分分析新方法[J]. 统计与决策, 2006, (3): 10-11.
[13] 刘遵雄,况志军,刘觉夫. 核主成分回归方法在电力负荷中期预测中的应用[J]. 计算机工程, 2006, 32(1): 31-33.
[14] 徐义田,等. 核主成分分析(KPCA)在企业经济效益评价中的应用[J]. 数学的实践与认识, 2006, 36(1): 35-38.
[15] 杨道军,等. 核主成分分析法在生态经济可持续发展评价中应用[J]. 环境科学与技术, 2007, 30(12): 91-93.
[16] B Scholkopf, A J Smola, K R Muller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem[J]. Neural Computation, 1998, (10): 1299-1319.

[作者简介]

肖　慧(1984-),女(汉族),广西桂林人,硕士研究生,主要从事神经网络、数理统计研究。
刘苏东(1979-),男(汉族),山东单县人,助工,主要从事水库移民研究。
黄小燕(1978-),女(汉族),广西崇左人,工程师,硕士,主要从事天气预报技术方法研究。
金　龙(1952-),男(汉族),上海人,研究员,主要从事人工智能的气象预报技术研究。