文章编号:1004-9037(2014)02-0293-05

基于小波包分解的含噪语音时频特性分析及端点检测

陈金龙 范影乐 倪红霞 武 薇

(杭州电子科技大学智能控制与机器人研究所,杭州,310018)

摘要:针对 Hilbert-Huang 变换方法在语音处理过程中存在模态混叠问题,本文提出了基于小波包分解的语音时频分析方法。首先对含噪语音进行小波包分解,对各分量分别进行经验模态分解,并运用相关系数阈值准则对固有模态函数进行筛选;然后建立语音信号的 Hilbert 谱和瞬时能量谱;最后将基于小波包分解的 Hilbert-Huang 变换瞬时能量谱方法应用于含噪语音的端点检测。实验结果表明:与传统广义维数以及谱熵算法相比,本文方法具有更好的准确性、稳定性和自适应性,能够有效描述语音信号非线性非平稳的时频特性。

关键词:语音端点检测;Hilbert-Huang 变换;时频分析;相关系数;阈值准则;小波包分解

中图分类号: **文献标识码**:A

Endpoint Detection of Noise-Corrupted Speech Time-Frequency Characteristics Based on Wavelet Packet Decomposition

Chen Jinlong, *Fan Yingle*, *Ni Hongxia*, *Wu Wei*

(Institute of Intelligent Control and Robot, Hangzhou Dianzi University, Hangzhou, 310018, China)

Abstract: To overcome the problem of mode mixing for Hilbert-Huang transform (HHT) in speech processing, a new method of time-frequency analysis based on wavelet packet decomposition (WPD) is proposed in this paper. Firstly, noise-corrupted speech is decomposed by using WPD, each component is carried out empirical mode decomposition (EMD) separately, and the intrinsic mode function (IMF) is selected by using correlation threshold criterion. Then, the Hilbert spectrum and instantaneous energy spectrum of speech signal are achieved. Finally, the method of instantaneous energy spectrum based on WPD is applied to noise-corrupted speech endpoint detection. Experimental results indicate that the proposed method is more accurate, robust and self-adaptive by comparison with the original generalized dimension (OGD) and the spectral entropy (SE) algorithms. The proposed method can effectively describe the time-frequency characteristics of the non-linear and non-stationary speech signal, and has provided a new idea for the research of speech signal.

Key words: speech endpoint detection; Hilbert-Huang transform; time-frequency analysis; correlation coefficient; threshold criterion; wavelet packet decomposition

引 言

语音在采集传输以及通信过程中不可避免的会引入各种噪声,噪声的存在将降低语音的清晰度和可懂度。因此含噪语音输出质量的改善程度,将直接影响到后续语音识别[1-2]、语音编码[3]等算法的准确性和复杂度。目前语音处理方法主要包括短时傅立叶变换、小波分析和 Wigner-Ville 分布等,上述方法考虑了语音信号在时频域上的特征表达,但他们仍基于语音信号具有短时线性平稳的假设,在语音的静态特征描述上具有较好的性能,但忽略了语音的非线性和非平稳特性。

1998 年,Huang NE.[4] 提出了一种适用于非线性、非平稳信号的 Hilbert-Huang 变换(Hilbert-huang transform,HHT)时频分析方法。其在语

基金项目:国家自然科学基金(60302027)资助项目;浙江省教育厅科研计划(Y201018050)资助项目。

收稿日期:2013-03-05;**修订日期**:2013-05-16

音信号的时频特性分析中得到了广泛的应用。例如文献[5]将 HHT 方法应用于语音信号的周期估计,有效地提高了基音识别的准确性与分辨率。KI. Molla 等人将 Hilbert 谱作为音频信号的时频描述,结果表明其与短时傅立叶变换相比具有显著的优势[6]。但在语音时频特性描述的上述应用中,HHT 也暴露了存在模态混叠以及低频覆盖等局限性[7]。针对上述问题,本文在 HHT 基础上,利用小波包对语音信号进行分解以及对固有模态函数的自适应筛选,能够有效的将频带进行细分,避免模态混叠,改善含噪语音的时频分辨率;引入相关系数阈值准则对固有模态函数(Intrinsic mode function,IMF)分量进行筛选,避免 Hilbert 谱中出现虚假频率。

1 基本原理

1.1 小波包分解

小波包分解(Wavelet package decomposition, WPD)具有良好的正交性、完备性、局部性,可将 WPD 视为函数空间中逐级正交剖分的扩展。WPD 在所有的频率范围内聚集的特性,使其具有更好的局部时频滤波特性,适合对语音进行经验模态分解(Empirical mode decomposition, EMD)前的宽带细化。

正交小波包分解如式(1)

$$\begin{cases} u_{2n}(t)=\sqrt{2}\sum_{k\in\mathbf{Z}}h(k)u_n(2t-k) \\ u_{2n+1}(t)=\sqrt{2}\sum_{k\in\mathbf{Z}}g(k)u_n(2t-k) \end{cases} \tag{1}$$

式中:$g(k)=(-1)^k h(1-k)$,$g(k)$和 $h(1-k)$是一对正交镜像滤波器。当 $n=0$ 时,$u_0(t)$和 $u_1(t)$分别为尺度函数 $\varphi(t)$和小波函数 $\psi(t)$。

按 j 级进行小波包的多分辨率分解时,最小的频率分辨率为:$\Delta f=\frac{1}{2^j}\times\frac{f_s}{2}$,因此第 j 级上小波包子带为 $[0,\Delta f]$,$[\Delta f,2\Delta f]\cdots[(2^j-1)\Delta f,2^j\Delta f]$。考虑到小波包分解层数与频率分辨率以及分析效率的关系,本文选择 3 层小波包分解,其中小波基为 Daubechies db3。

1.2 Hilbert-Huang 变换

Hilbert-Huang 变换包括两部分:EMD 和 Hilbert 谱分析。

1.2.1 经验模态分解

EMD 分解是把复杂的信号分解为有限个固有模态函数 IMF 分量之和,经过一系列分解后,信号 $x(t)$被分解成 n 个固有模态函数 $c_i(t)$和一个余项 $r_n(t)$,如式(2)所示

$$x(t)=\sum_{i=1}^{n}c_i(t)+r_n(t) \tag{2}$$

1.2.2 Hilbert 谱

分解后得到的 IMF 分量通过 Hilbert 变换,求得瞬时频率,得到 Hilbert 谱。对每个固有模态分量 $c_i(t)$作 Hilbert 变换

$$y_i(t)=\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{c_i(\tau)}{t-\tau}\mathrm{d}\tau \tag{3}$$

根据式(4)构造解析信号 $z_i(t)$

$$z_i(t)=c_i(t)+\mathrm{j}y_i(t)=a_i(t)\mathrm{e}^{\mathrm{j}\theta_i(t)} \tag{4}$$

式中:$a_i(t)$为解析信号幅值,$\theta_i(t)$为相角

$$a_i(t)=\sqrt{c_i^2(t)+y_i^2(t)}$$

$$\theta_i(t)=\arctan\left[\frac{y_i(t)}{c_i(t)}\right] \tag{5}$$

瞬时频率定义为

$$w_i(t)=\frac{\mathrm{d}\theta_i(t)}{\mathrm{d}t} \tag{6}$$

从而原始信号可以表示为

$$x(t)=\sum_{i=1}^{n}a_i(t)\mathrm{e}^{\mathrm{i}\int w_i(t)\mathrm{d}t} \tag{7}$$

式(7)表明信号的幅值和瞬时频率都是时间的函数,从而可以在时频平面中将幅值表示成时间和瞬时频率的函数 $H(w,t)$,即原始信号的 Hilbert 谱。$H(w,t)$对时间积分,就得到 Hilbert 边际谱(Marginal spectrum, MS)

$$H(w)=\int_0^T H(w,t)\mathrm{d}t \tag{8}$$

Hilbert 瞬时能量谱(Instantaneous energy spectrum, IES)为 $H(w,t)$对频率 w 的积分,其定义为

$$E(t)=\int_0^{w_0}H(w,t)\mathrm{d}w \tag{9}$$

1.2.3 相关系数阈值准则 IMF 分量筛选

由于 IMF 分量和剩余信号 $\boldsymbol{r}_n(t)$是原始信号的正交分量,因此相应的 IMF 与原信号具有很强的相关性。依次计算每个 IMF 与原信号的相关系数 $\boldsymbol{r}_i$ 作为判别相关性的依据,表达式为

$$\boldsymbol{r}_i=\frac{\sum_{i=1}^{N}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}})}{\sqrt{\sum_{i=1}^{N}(\boldsymbol{X}_i-\bar{\boldsymbol{X}})^2}\sqrt{\sum_{i=1}^{N}(\boldsymbol{Y}_i-\bar{\boldsymbol{Y}})^2}} \tag{10}$$

式中:$i=1,\cdots,n$,$\boldsymbol{X}_i$ 为 IMF 分量序列,$\boldsymbol{Y}_i$ 为重构信号序列,N 为采样点数,$\bar{\boldsymbol{X}}$ 为 $\boldsymbol{X}_i$ 序列的均值,$\bar{\boldsymbol{Y}}$ 为 $\boldsymbol{Y}_i$ 序列的均值。

对于 n 个 IMF 的相关系数 $r_i(i=1,\cdots,n)$,剔

除阈值为

$$\lambda=\frac{\max(r_i)}{\eta}(i=1,\cdots,n) \tag{11}$$

式中：η 为比例因子。计算每个 IMF 的相关系数，筛选准则如下：若大于 λ，则保留该 IMF，否则剔除该 IMF 加入到剩余分量中。通过该方法，可以有效去除 IMF 中相关性较差的分量，避免希尔伯特谱中出现虚假频率分量[8]。

2　实验结果

本文实验数据为自采集数据库中的孤立词语：对象为 50 名来自全国各大区的大学生，每人读 5 次，每次读 26 个英文字母各一遍，采样频率为 8 kHz，16 bit 量化，wav 格式。背景噪声数据来源于 NOISEX92 标准噪声数据库[9]，选择其中 3 种噪声，分别为飞机噪声(F16)、工厂噪声(Factory1)和办公室噪声(Babble)。对含噪语音均采用数字滤波器 $H(z)=1-\mu z^{-1}$($\mu=0.937\ 5$)进行预加重处理，用于消除低频交流电工频等干扰。

2.1　相关系数阈值准则的有效性

为了说明相关系数阈值准则的有效性，对含噪语音(工厂噪声，下同)用 db3 小波基进行 3 层分解，对分解的各个信号进行重构，得到重构信号，记为 $\mathrm{WPD}_i(i=1,2\cdots,8)$，对重构信号进行 EMD 分解，计算对应的 IMF 分量以及相关系数，结果如表 1 所示，其中 $\eta=50$。

从表 1 可以看出，EMD 分解具有自适应性，表现为较高的相关系数一般集中于前几个 EMD 分解出来的 IMF 分量中。因此根据式(11)筛选出来的 IMF 分量在所有的 IMF 分量中占主导作用，也进一步说明相关系数阈值准则的有效性。通过相关系数阈值准则筛选有效的 IMF 分量，剔除相关系数较差的 IMF 分量，避免在 Hilbert 谱中出现虚假频率分量。

表 1　各 IMF 与对应 WPD 分量的相关系数对比表

Table 1　Correlation coefficient comparsion of each component of IMF and corresponding WPD

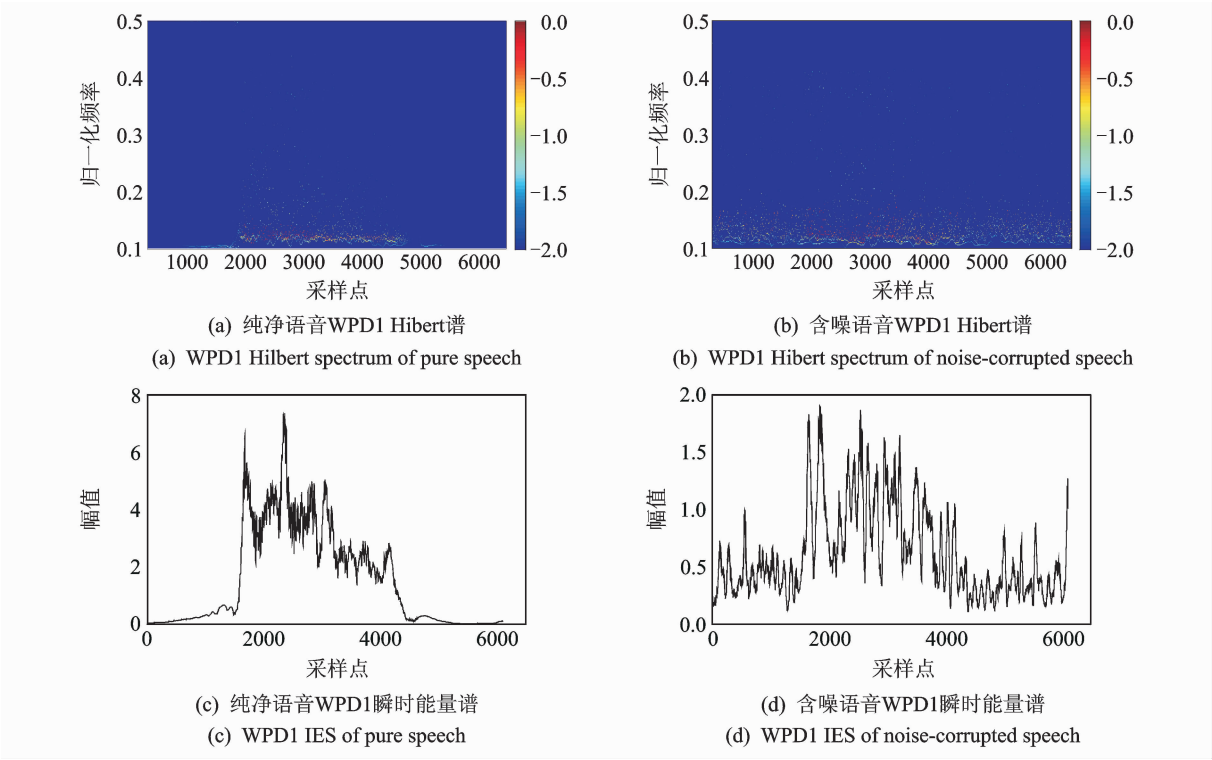
	WPD1	WPD2	WPD3	WPD4	WPD5	WPD6	WPD7	WPD8
IMF1	0.212 1	0.684 4	0.911 6	0.900 6	0.993 1	0.943 1	0.955 1	0.944 5
IMF2	0.451 1	0.563 8	0.173 7	0.326 7	0.141 5	0.135 5	0.360 8	0.230 4
IMF3	0.684 0	0.141 2	0.033 4	0.020 7	0.009 4	0.049 5	0.016 1	0.000 9
IMF4	0.282 4	0.039 1	0.009 9	0.000 8	0.004 8	0.012 0	0.001 0	0.000 3
IMF5	0.049 3	0.000 5	0.002 1	0.000 2	0.003 5	0.001 6	0.000 1	0
IMF6	0.028 1	0.000 1	0.000 1	0.000 2	0.000 9	0	0.000 2	0
IMF7	0.013 4	0.000 2	0	0.000 1	0.000 1	0	0.000 2	0
IMF8	0.001 0	0.001 0	0	0.000 2	0.000 1	0	0.000 2	0
阈值	0.013 7	0.013 7	0.018 2	0.018 0	0.019 9	0.018 9	0.019 1	0.018 9

2.2　纯净语音和含噪语音的时频分析

基于小波包分解的 HHT 变换方法，采用相关系数阈值准则筛选 IMF 分量，分别对纯净语音和含噪语音进行 Hilbert 谱分析，如图 1 所示。图 1(a,b)分别为纯净语音和含噪语音 WPD1 的 Hilbert 谱，可以发现纯净语音在时间轴上 2 000～5 000采样点之间有低频能量分布，而含噪语音在整个时间轴采样点上都存在低频能量分布。图 1(c,d)显示了纯净语音和含噪语音 WPD1 的瞬时能量谱，可以发现它们的瞬时能量谱差异较大；纯净语音的瞬时能量谱主要集中于语音区域；而含噪语音的瞬时能量谱在整个时间轴采样点上都有分布，但语音区域段的瞬时能量谱占主导地位，而噪声段瞬时能量谱相对语音区域较弱。因此语音和噪声的瞬时能量谱特征具有较好的区分度，后文将此特征作为语音端点检测的依据。

2.3　小波包分解在时频分析中的优势

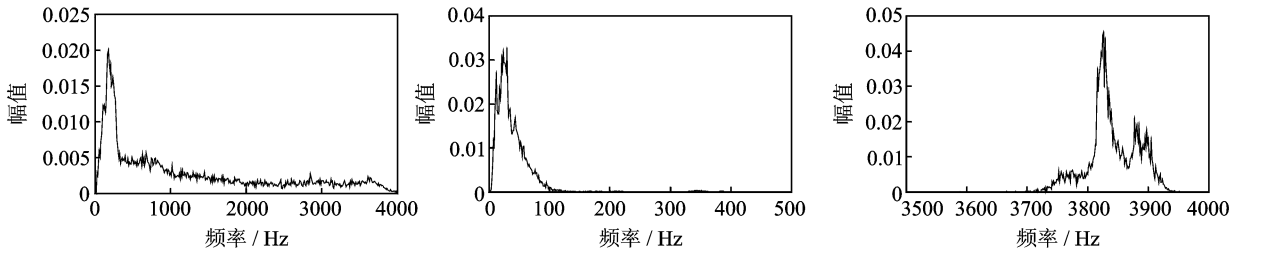
为便于比较，本文对含噪语音分别按如下两种方法进行处理：(1)HHT 变换；(2)小波包分解后的 HHT 变换，其结果如图 2 所示。图 2(a-c)分别为含噪语音 HHT 边际谱、含噪语音 WPD1 的边际谱以及含噪语音 WPD8 的边际谱，未引入小波包分解的边际谱(图 2(a))的频带范围分布较广，在整个频带范围都有分布，而引入小波包分解的 WPD1 和 WPD8 的边际谱(图 2(b,c))分布的频带范围较窄，分别集中于低频和高频段分布。由实验结果可知：小波包分解在含噪语音 Hilbert 谱分析中具有显著的优势，将频带范围细分，避免模态混叠，使其满足 HHT 模态的单一组分要求，由于小波包分解具有正交性与自适应性，从而提高 EMD 的分解能力，改善时频分辨率。



(a) 纯净语音WPD1 Hibert谱
(a) WPD1 Hilbert spectrum of pure speech
(b) 含噪语音WPD1 Hibert谱
(b) WPD1 Hibert spectrum of noise-corrupted speech
(c) 纯净语音WPD1瞬时能量谱
(c) WPD1 IES of pure speech
(d) 含噪语音WPD1瞬时能量谱
(d) WPD1 IES of noise-corrupted speech

图 1 纯净语音和含噪语音时频谱

Fig. 1 Time-frequency spectrum of pure speech and noise-corrupted speech



(a) 含噪语音HHT边际谱
(a) HHT MS of noise-corrupted speech
(b) 含噪语音WPD1 HHT边际谱
(b) WPD1 HHT MS of noise-corrupted speech
(c) 含噪语音WPD8 HHT边际谱
(c) WPD8 HHT MS of noise-corrupted speech

图 2 含噪语音 HHT 边际谱和含噪语音 WPD HHT 边际谱

Fig. 2 HHT and WPD HHT marginal spectrum of noise-corrupted speech

2.4 基于瞬时能量谱特征的含噪语音端点检测

2.4.1 实验步骤

为了验证小波包分解 HHT 方法在分析含噪语音时频特征方面的有效性，本文提出了基于小波包分解的 HHT 变换瞬时能量谱方法，用于含噪语音的端点检测，详细步骤如下：

(1) 对含噪语音进行预加重处理，选用 db3 小波基进行 3 层分解，将分解的信号重构记为 WPD_i ($i=1,2\cdots,8$)。

(2) 对重构的 WPD1 进行 EMD 分解并运用相关系数阈值准则筛选获得有效的 IMF 分量。

(3) 对有效的 IMF 分量进行 Hilbert 变换并进行分帧处理。

(4) 计算相应的瞬时能量谱 $E(t)$，将前 5 帧瞬时能量谱均值作为噪声能量谱 E_{noise}。

(5) 采用起-止双门限阈值法进行端点检测，若 $E(t)<aE_{\text{noise}}$，则继续检测，如果 $E(t)\geqslant aE_{\text{noise}}$，则记录为语音开始点，直到 $E(t)<bE_{\text{noise}}$，则记录为语音结束点；其中 a 和 b 分别为比例因子。

如果语音结束点与语音开始点之差小于长度阈值 c，则认为检测得到的语音起点和终点均为干扰点，将它们舍弃；然后对后续瞬时能量谱序列继续重复步骤(5)进行语音端点检测，直到检测到有效的语音端点或语音序列结束为止。

2.4.2　实验结果与分析

设帧长为 240，帧移为 80，参数 $a=1.5$，$b=1$，$c=5$。在端点检测时，如果自动检测的前后端点与手工标定的端点差别在±5 帧以内，则视为正确[10]。

为了说明本文方法的可行性，对不同类型以及不同强度的含噪语音引入传统广义维数（Original generalized dimension，OGD）以及谱熵（Spectral entropy，SE）的端点检测方法，如表 2 所示。可以发现，当信噪比为 20 db 时，小波包分解的 HHT 瞬时能量谱算法的准确率要略低于传统广义维数和谱熵算法，但是当信噪比降到 10 db 以下时，本文端点检测算法的准确率较其他两种算法具有显著的优势，尤其当信噪比为 0 db 时，谱熵算法的准确率已经下降到 50%左右，传统广义维数在 70%左右，而本文的算法仍旧保持在 90%左右（F16 时只有 74%，但是仍高于其他两种方法）。传统广义维数与谱熵算法在高信噪比的情况下，语音端点检测的效果较理想，但是对于信噪比较低的情况下，端点检测效果不是很理想，而本文的算法相对于信噪比的变化，端点检测效果较为稳定，具有较好的检测能力、自适应性及较强的鲁棒性。

表 2　本文方法与传统方法的语音端点检测准确率对比表(%)

Table 2　Correct rate comparison of speech endpoint detection with different methods (%)

SNR /dB	Babble			Factory1			F16		
	WPD-HHT	OGD	SE	WPD-HHT	OGD	SE	WPD-HHT	OGD	SE
20	88.0	97.6	94.8	90.0	96.2	97.3	92.0	95.2	94.9
10	88.0	86.5	71.4	90.0	86.1	82.6	86.0	76.9	72.0
5	88.0	78.9	51.7	92.0	76.9	67.2	80.0	75.0	51.6
0	94.0	74.5	49.6	92.0	76.0	58.5	74.0	67.3	47.8

3　结束语

本文提出对含噪语音信号进行小波包分解，以改善 Hilbert-Huang 变换方法的模态混叠问题，提高时频分辨率；另外提出相关系数阈值准则对 IMF 分量进行筛选，将避免 Hilbert 谱中出现的虚假频率。通过含噪语音的端点检测应用，验证了本文语音时频分析方法的有效性。本文方法将为后续语音复原、语音识别以及语音编码的研究提供一个新的思路。

参考文献：

[1] Kim K, Kim M Y. Robust speaker recognition against background noise in an enhanced multi-condition domain[J]. IEEE Transactions on Consumer Electronics, 2010, 56(3): 1684-1688.

[2] 余华，黄程韦，金赟，等. 基于粒子群优化神经网络的语音情感识别[J]. 数据采集与处理，2011，26(1)：57-62.
Yu Hua, Huang Chengwei, Jin Yun, et al. Speech emotion recognition based on particle swarm optimizer neural network[J]. Journal of Data Acquisition and Processing, 2011, 26(1): 57-62.

[3] Backstrom T, Magi C. Effect of white-noise correction on linear predictive coding[J]. IEEE Signal Processing Letters, 2007, 14(2): 148-151.

[4] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proc. R. Soc. Lond. A, 1998, 454: 903-995.

[5] Huang H, Pan J Q. Speech pitch determination based on Hilbert-Huang transform[J]. Signal Processing, 2006, 86(4): 792-803.

[6] Molla K I, Shaikh M, Hirose K. Time-frequency representation of audio signals using Hilbert spectrum with effective frequency scaling[C]// Proceeding of 11th International Conference on Computer and Information Technology (ICCIT). Khulna: IEEE, 2008: 335-340.

[7] Peng Z K, Tse P W, Chu F L. An improved Hilbert-Huang transform and its application in vibration signal analysis[J]. Journal of Sound and Vibration, 2005, 186(2): 187-205.

[8] Yuan L, Yang B H, Ma S W, et al. Combination of wavelet packet transform and Hilbert-Huang transform for recognition of continuous EEG in BCIs [C]// Proceeding of the 2nd IEEE International Conference Computer Science and Information Technology. Beijing, China: IEEE, 2009: 594-599.

[9] Varga A. Assessment for automatic speech recognition: Ⅱ. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems [J]. Speech Communication, 1993, 12(3): 247-251.

[10] 武薇，范影乐，庞全. 基于广义维数距离的语音端点检测方法[J]. 电子与信息学报，2007，29(2)：465-468.
Wu Wei, Fan Yingle, Pang Quan. A speech endpoint detection method based on the feature distance of generalized dimension[J]. Journal of Electronics & Information Technology, 2007, 29(2): 465-468.

作者简介：陈金龙（1988-），男，硕士研究生，研究方向：生物医学信号处理，E-mail：chenjinlong126@163.com；范影乐（1975-），男，博士，教授，研究方向：模式识别、智能信息处理；倪红霞（1969-），女，高级实验师，研究方向：语音信号处理、生物信号测量与处理；武薇（1979-），女，博士，讲师，研究方向：模式识别与信号检测。