文章编号:1004-9037(2012)01-0051-06

基于 EEMD 域统计模型的话音激活检测算法

吴其前[1] 张雄伟[2]

(1. 解放军理工大学通信工程学院,南京,210007; 2. 解放军理工大学指挥自动化学院,南京,210007)

摘要:提出了一种基于 EEMD 域统计模型的话音激活检测算法。算法首先利用总体平均经验模态分解(Ensemble empirical mode decomposition, EEMD)对带噪语音进行分解,得到信号的本征模式函数(Intrinsic mode function, IMF)分量,选择与原信号的相关性最高的两个分量相加组成主分量;然后对主分量进行频域分解,引入统计模型,求出 EEMD 域特征参数;最后利用噪声与语音的 EEMD 域特征参数的不同来进行语音激活检测。实验结果表明,在不同信噪比情况下,本文算法性能优于目前常用的 VAD 算法,特别在噪声强度大时体现出明显的优势。

关键词:话音激活检测;经验模式分解;总体平均经验模式分解;EEMD 域统计模型

中图分类号:TN912.3 **文献标识码**:A

Voice Activity Detection Algorithm Based on Ensemble Empirical Mode Decomposition Domain Statistical Model

Wu Qiqian[1], *Zhang Xiongwei*[2]

(1. Institute of Communication Engineering, PLA University of Science & Technology, Nanjing, 210007, China;
2. Institute of Command Automation, PLA University of Science & Technology, Nanjing, 210007, China)

Abstract: Voice activity detection algorithm based on ensemble empirical mode decomposition domain statistical model is presented. The noisy speech is decomposed into intrinsic mode function (IMF) components by using ensemble empirical mode decomposition (EEMD) method. Two IMF components with the higher correlation with original speech are added to calculate the characteristic parameter of the statistical model. The decision of the speech/noise is made by comparing the characteristic parameter with its threshold. The proposed VAD algorithm is tested on speech signals under various noise conditions with several SNRs. Experimental results show that the proposed VAD algorithm outperforms some standard VAD algorithms, especially under a low SNR noisy condition.

Key words: voice activity detection (VAD); empirical mode decomposition (EMD); ensemble empirical mode decomposition (EEMD); EEMD domain statistical model

引 言

话音激活检测(Voice activity detection, VAD)是一种重要的语音信号处理技术,其有效性直接影响语音处理系统的性能,如语音编码中编码速率控制、语音识别系统的识别率等。常用 VAD 算法的本质是,在某种特征域或联合特征域中,研究语音和噪声内在特征,提取某种特征参数,利用特征参数之间的差异性来区分语音和噪声。传统时域、频域特征参数包括短时能量、过零率、LPC 参数[1]、时域联合参数[2]、频域特征熵[3]等。在低信噪比条件下,由于上述参数对语音和噪声的内在特征区分能力不强,VAD 算法判决性能很差。Sohn[4]等引入统计模型,分析语音和噪声的频域子带能量分布特性,建立基于特征熵的判决函数,提高了判决性能。考虑到 FFT 变换自身的特点,频域子带特征不能准确反映语音和噪声的内在特征,Shaojun 等提出了基于小波变换的 VAD 算法[5]。与 FFT 分析相比,小波变换能够更好地分离语音和噪声,但

基金项目:江苏省自然科学基金(BK2009059)资助项目。

收稿日期:2011-02-22;**修订日期**:2011-04-12

小波基需要预先确定,且分解的尺度完全相同,不能根据信号内在特征进行自适应分解。

1998 年, N. E. Huang 提出一种新的自适应信号处理方法—经验模式分解(Empirical mode decomposition,EMD),该方法能够很好地处理非线性、非平稳信号[6]。EMD 分解可以理解为一种自适应滤波过程,根据信号内在波动特征,将信号中不同尺度的波动逐级分解开来,产生一组反映信号内在波动特征的数据,称为本征模式函数(Intrinsic mode function,IMF)。EMD 方法提出后被广泛应用于语音信号处理[7-9],并取得了一定效果,但是 EMD 方法处理包含间歇性分量的信号,如噪声或干扰时,经常会出现模式混合,导致信号的时频分布会出现严重的偏移[10]。模式混合是指在一个 IMF 中包含不同尺度的信号分量,或者多个不同 IMF 包含一个相似的尺度分量。为了克服这个问题,Huang 提出了一种改进的 EMD 方法——总体平均经验模式分解(Ensemble empirical mode decomposition,EEMD)。针对语音信号,与小波分解相比,EEMD 能够提供更好的尺度分解,IMF 分量能够准确体现语音的内在波动特征,便于在尺度空间上分离语音和噪声[11]。

考虑到语音和噪声不同的尺度空间分布特性,本文利用 EMMD 对带噪语音进行尺度分解,得到分别包含语音和噪声波动特征的 IMF 分量。首先分析 IMF 分量与原始语音的相关性,提取包含语音内在特征较多的两个 IMF 分量组合成主分量,以减小其他尺度的噪声干扰;然后计算主分量的 EEMD 域能量谱,引入统计模型,计算特征参数;最后结合判决门限给出话音激活检测结果。实验结果表明,本文算法具有较强的鲁棒性和较高的准确率。

1 经验模式分解原理

1.1 EMD 原理

EMD 是一个不断筛选出 IMF 分量的循环迭代过程,每个 IMF 分量满足以下两个条件:(1)整个序列上信号极值点个数和过零点个数相等或至多相差一个;(2)整个序列任意一点处,由其局部极大值拟合的上包络线和其局部极小值拟合的下包络线的平均值为零。具体流程如下:

(1)分别找出原始信号 $x(t)$ 中所有局部极大值和局部极小值,采用三次样条函数,拟合成该数据的上包络线、下包络线。

(2)求出上下包络线的均值信号 $m_1(t)$,将原始信号 $x(t)$ 减去该均值信号得到一个差值信号 $h_1(t)$,即

$$h_1(t) = x(t) - m_1(t) \tag{1}$$

利用式(2)计算 IMF 判定条件,如果满足条件则认为 $h_1(t)$ 为 IMF 分量,否则重复步骤(1),步骤(2),直到满足 IMF 判定条件为止。

$$\frac{\sum_{t=0}^{T} |h_{1,k-1}(t) - h_{1,k}(t)|^2}{\sum_{t=0}^{T} h_{1,k-1}^2(t)} \leqslant 0.3 \tag{2}$$

式中,k 为重复次数,T 为信号时长,$h_{1,0}(t)$ 定义为初始迭代信号。定义满足条件的差值信号为第一个 IMF 分量 $C_1(t)=h_{1,k}(t)$。

(3)将原始信号减去上述的 IMF 分量,得到去掉高频成分的差值信号 $r_1(t)$,即

$$r_1(t) = x(t) - C_1(t) \tag{3}$$

(4)将 $r_1(t)$ 代替原始信号 $x(t)$,重复步骤(1)~(3)得到第二个 IMF 分量 $C_2(t)$ 及新的差值信号 $r_2(t)$。

(5)不断重复上述操作,直到差值信号 $r_n(t)$ 为单调信号,不需要再分解为止,此时 $r_n(t)$ 代表原信号的趋势和均值。

通过上述所谓"筛选"过程,原始信号被分解为一系列 IMF 分量和残差项 $r_n(t)$,原始信号 $x(t)$ 可以表示为

$$x(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \tag{4}$$

1.2 EEMD 原理

EEMD 是一种借助于噪声的数据分析方法,其基本流程如下[11]:

(1)将原始信号加上强度很低的白噪声。

(2)将带噪信号进行 EMD 分解,得到一组 IMF 分量。

(3)重复执行上述步骤(1)~(2),每次添加不同的白噪声序列,共重复 N 次。

(4)为消除人工添加噪声的影响,将 N 次 EMD 分解得到的相应的 IMF 分量相加,取其平均值作为最终的 IMF 分量 $C_i^e(t)$,即

$$C_i^e(t) = \frac{1}{N}\sum_{k=1}^{N} C_{i,k}(t) \tag{5}$$

式中,$C_{i,k}(t)$ 为第 k 次 EMD 分解得到的第 i 个 IMF 分量。

EEMD 利用白噪声在尺度空间均匀分布的特点,将带噪信号不同尺度的分量投影到由白噪声所

建立的尺度空间，消除了模式混合现象。考虑到白噪声的不相关性，EEMD 利用多次分解求平均的方法消除人为添加的噪声。与 EMD 相比，EEMD 可以有效地消除模式混合，使得 IMF 分量的物理意义更加明显。图 1 示例了带噪语音的 EEMD 分解，语音采样率为 8 kHz，时长约为 0.57 s，N 为 100，噪声幅度为信号标准均方差的 0.1 倍。图 1(a)从上至下分别为原始加噪语音信号及 IMF1～IMF4 分量的波形图，图 1(b)从上至下分别为加噪语音语谱图及 IMF1～IMF4 分量的时频分布图，由图中可以看出 IMF1 分量的频域分布比较均匀，包含了噪声信号频谱和语音信号部分高频部分，而 IMF2～IMF4 分量的频域分布基本覆盖了原始语音的频谱分布，体现了语音信号不同尺度的内在波动特征。

2　基于 EEMD 的话音激活检测算法

2.1　EEMD 域统计模型特征参数

对语音信号进行 EEMD 分解后，考虑到语音高频、低频部分会被分解到不同尺度空间，分别计算第 i 个 IMF 分量与原信号的相关系数 $R_i(C_i^e(t), x(t))$，选择相关性最大的第 i 个 IMF 分量为第一特征分量，即

$$i = \mathrm{argmax}(R_i(C_i^e(t), x(t))) \tag{6}$$

在剩余分量中再次进行上述计算，选取相关性最大的 IMF 分量为第二特征分量，两个 IMF 分量相加得到 EEMD 域主分量。对 EEMD 域主分量进行短时傅里叶分析，得到其子带能量。

假设待测信号中由语音信号与背景噪声相加所得，且话音信号、噪声信号的第 j 个子带能量对数值

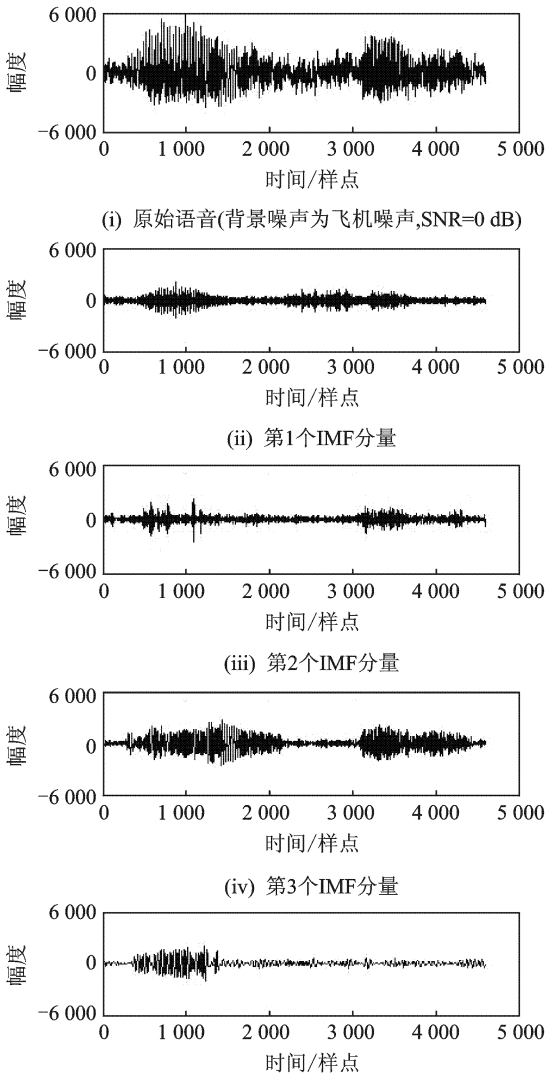

图 1　带噪语音的 EEMD 分解

满足均值为 $\mu_S(j)$，$\mu_N(j)$，方差分别为 $\lambda_S(j)$，$\lambda_N(j)$ 的高斯分布，其概率密度函数分别记为 $P_S{}^j(x)$，$P_N{}^j(x)$，定义子带特征参数为

$$F_j = \frac{H(p_S^j \parallel p_N^j)H(p_N^j \parallel p_S^j)}{H(p_S^j \parallel p_N^j) + H(p_N^j \parallel p_S^j)} \tag{7}$$

该特征参数能较好地区分两个不同随机变量的概率密度[12]，则 EEMD 域特征参数定义为

$$F = \sum_{j=0}^{N-1} F_j = \sum_{j=0}^{N-1} \frac{H(p_S^j \parallel p_N^j)H(p_N^j \parallel p_S^j)}{H(p_S^j \parallel p_N^j) + H(p_N^j \parallel p_S^j)} \tag{8}$$

式中，N 为子带个数，$H(P_S{}^j \parallel P_N{}^j)$ 为 Kullback-Leibler 相关熵，计算公式为

$$H(p_S^j \parallel p_N^j) = \int p_S^j(x)\ln\left(\frac{p_S^j(x)}{p_N^j(x)}\right)\mathrm{d}x = \frac{1}{2}\left[\frac{\lambda_S(j)}{\lambda_N(j)} + \ln\frac{\lambda_N(j)}{\lambda_S(j)} + \frac{(\mu_S(j) - \mu_N(j))^2}{2\lambda_N^2(j)} - 1\right] \tag{9}$$

式中，$\frac{\lambda_S(j)}{\lambda_N(j)}$ 定义为先验信噪比 ε_j，可以利用下式通过决策反馈方法[13]进行估计。

$$\varepsilon_j(n) = \alpha G^2(\varepsilon_j(n-1), \gamma_j(n-1))\gamma_j(n-1) + (1-\alpha)P[\gamma_j(n) - 1] \tag{10}$$

式中，$\frac{|X_j|^2}{\lambda_N(j)}$ 定义为后验信噪比 γ_j，n 为帧号，X_j 为子带能量的对数值。α 为遗忘因子，$G(\varepsilon_j(n), \gamma_j(n))$ 定义为增益函数

$$G(\varepsilon_j, \gamma_j) = \Gamma(1.5) * \frac{\sqrt{V_j}}{\gamma_j} * \exp\left(-\frac{V_j}{2}\right) * \left[(1+V_j)I_0\left(\frac{V_j}{2}\right) + V_k I_1\left(\frac{V_j}{2}\right)\right] \tag{11}$$

式中，$\Gamma(1.5) = \frac{\sqrt{\pi}}{2}$，$V_j = \frac{\varepsilon_j}{1+\varepsilon_j}\gamma_j$，$I_0(\cdot)$，$I_1(\cdot)$ 分别为零阶和一阶修正贝塞尔函数(Modified bessel function)。$P[x]$函数计算公式为

$$P[x] = \begin{cases} x, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{12}$$

2.2 算法原理

本文算法选取与语音相关性高的 IMF 分量组合成主分量，计算其 EEMD 域特征参数，结合判决门限区分语音和噪声，原理框图如图 2 所示。首先对语音信号进行分帧处理，帧长为 30 ms，每帧重叠 20 ms。对分帧语音进行 EEMD 分解得到 IMF 分量，分别计算各 IMF 分量与原语音信号的相关系数，取最大值所对应的 IMF 分量作为主分量。然后对 IMF 主分量进行短时傅里叶分析，求出子带能量对数值，计算 EEMD 域特征参数。再将特征参数与判决门限进行比较，如果数值小于门限则判为噪声，否则判为语音。最终输出经过拖尾延迟保护后的判决结果。

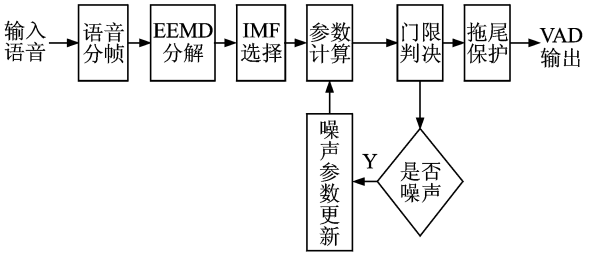

图 2 基于 EEMD 域统计模型的 VAD 算法原理

算法假定初始 10 帧为背景噪声，利用统计方法计算背景噪声子带能量对数值的均值、方差，计算后验信噪比 $\gamma_j(0)$，先验信噪比 $\varepsilon_k(0) = \alpha + (1-\alpha)P[\gamma_j(0) - 1]$。算法在噪声段进行参数更新，为了防止误判对参数更新的影响，只在特征参数远低于判决门限时才利用下式进行参数的更新。

$$\lambda_N^{\mathrm{new}}(j) = \beta \cdot \lambda_N^{\mathrm{old}}(j) + (1-\beta) \cdot \lambda_N^{\mathrm{cur}}(j) \tag{13}$$

式中，$\lambda_N^{\mathrm{new}}(j)$，$\lambda_N^{\mathrm{old}}(j)$ 分别为更新后、更新前背景噪声子带能量对数值的方差估计；β 为遗忘因子；$\lambda_N^{\mathrm{cur}}(j)$ 为当前帧背景噪声子带能量对数值的方差估计。同理更新背景噪声子带能量对数值的均值参数。

3 仿真与分析

本文算法采用多段 8 kHz 采样、16 bit 量化的标准测试语音作为干净语音，总时长约为 20 min，利用 Noisex-92 噪声库的数据作为背景噪声。为验证不同信噪比、不同背景噪声环境下的算法判决性能，根据不同的分段信噪比调整噪声幅度，与语音信号进行加性混合得到带噪语音。仿真过程中，α，β 分别设为 0.97，0.9。性能评估采用误判率和漏判率两个标准参数，误判率为将噪声判为语音的帧数占总帧数的比率，漏判率为将语音判为噪声的帧数占总帧数的比率。误判和漏判均定义为错判，错判率为误判率和漏判率之和。

图 3 分别示例了一段语音的 EEMD 域特征参数和白噪声(SNR=0 dB)条件下的 VAD 判决。图 3(a)为干净语音波形图(前 35 000 点)；图 3(b)为手工 VAD 判决示意图；图 3(c)为带噪语音波形图；图 3(d)为 EEMD 域特征参数；图 3(e)中本文算法 VAD 判决示意图。从图中可以看出，本文算

法与手工标注的判决结果基本一致，但在语音段开始部分还存在漏判现象。语音段开始部分大多为能量较低的清音，在低信噪比条件下，清音和背景噪声的内在振动模式比较接近，EEMD 很难分离清音与背景噪声，IMF 主分量包含部分背景噪声信息，导致特征参数区分度下降而引起漏判。

为了进一步分析算法性能，本文在白噪声、飞机噪声等典型背景噪声环境下，对算法进行了仿真，并与常用的标准语音编码算法中的 VAD 算法进行比较，如 G.729B，SMV，ARM2 等语音编码算法中的 VAD 算法。其中，SMV 中 VAD 算法的数据是在模式 2 下（速率为 4.5 kb/s）测得，VAD 选项和语音增强选项均选 A，AMR2 中 VAD 算法的数据是在速率为 12.2 kb/s 的模式下测得。表 1 给出不同测试条件下的算法误判率和漏判率比较。

VAD 判决时，误判和漏判是相互制约的两个指标，减小误判的同时必然导致漏判的增加，反之亦然。从表中数据可以看出，AMR2 中 VAD 算法

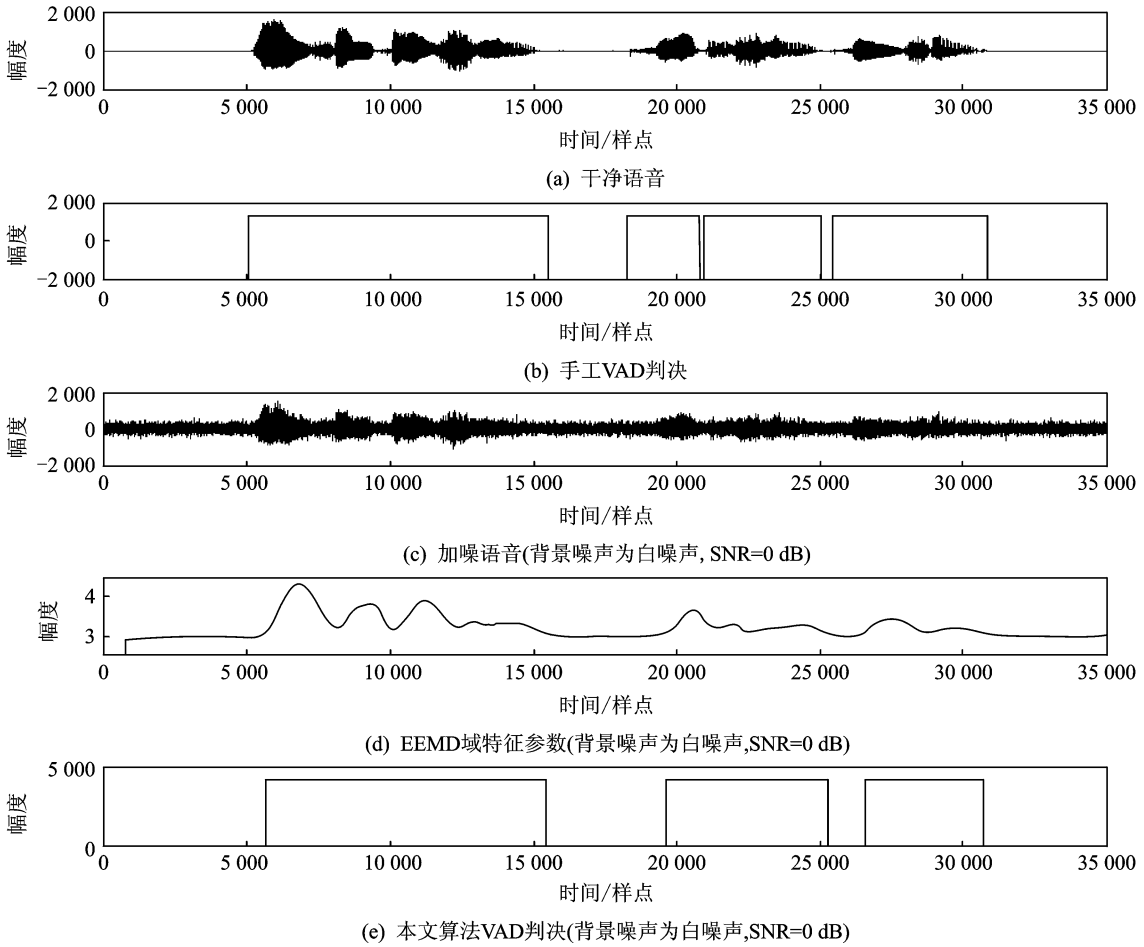

图 3　特征参数及 VAD 判决示意图

表 1　VAD 算法性能比较表

		G.729B		SMV		AMR2		本文算法	
		误判	漏判	误判	漏判	误判	漏判	误判	漏判
白噪声	10 dB	5.91	14.25	0	15.45	9.52	2.19	5.53	1.20
	5 dB	5.91	23.81	0.34	24.85	7.24	4.73	4.72	2.16
	0 dB	5.91	41.07	7.86	27.02	5.12	9.83	4.62	4.17
	−5 dB	5.91	58.65	8.75	55.71	0.39	28.54	7.86	7.34
飞机噪声	10 dB	5.50	7.56	0	12.78	8.43	0.57	2.65	3.28
	5 dB	5.50	12.27	0	21.92	7.02	2.01	2.69	3.59
	0 dB	5.12	25.94	3.01	30.53	6.19	7.92	3.35	7.17
	−5 dB	5.02	44.52	8.68	44.47	2.50	28.01	9.97	3.81

的判决错误则以误判为主，而本文算法以错判率最低为目标，根据先验信噪比估计值自适应选用不同的判决门限。从表中数据可以看出，在不同信噪比、不同背景噪声条件下，本文算法均取得了较好的判决效果，性能优于 AMR2 VAD 算法。尤其在低信噪比条件下，EEMD 能够根据语音信号自身的波动特征进行自适应分解，对噪声的分离效果优于傅里叶变换和小波分解等其他方法，EEMD 域特征参数能够较好地区别语音和噪声，错判率明显小于各种标准 VAD 算法。本文算法中 EEMD 的运算量较大，如何在保证分解性能的同时降低运算量需要进一步研究。

4 结束语

由于 EEMD 消除了模式混合问题，本文算法利用 EEMD 分解对带噪语音信号进行分解，得到的 IMF 分量能够更为准确地描述语音信号自身波动特征。利用最大相关特性选择 IMF 主分量，引入统计模型特征参数定义，提出了一种基于 EEMD 域统计模型的 VAD 算法。不同背景噪声条件下的仿真结果表明，本文算法性能优于 AMR2 VAD 算法。

参考文献：

[1] Rabiner L, Juang B H. Fundamentals of speech recognition[M]. New Jersey: Prentice-Hall PTR, 1993.

[2] Benyassine A, Shlomot E, Su H Y, et al. ITU-T G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data application[J]. IEEE Commun Mag, 1997,35(9):64-73.

[3] Shen J L, Hung J W, Lee L S. Robust entropy-based endpoint detection for speech recognition in noisy environments[C]//ICSP 1998. Sydney, Australia: IEEE Press, 1998:232-235.

[4] Sohn J, Kim N S, Sung W. A statistical model-based voice activity detection[J]. IEEE Signal Processing Letters, 1999,6(1):1-3.

[5] Shaojun J, Hitato G, Fuliang Y. A new algorithm for voice activity detection based on wavelet transform[C]//Proceeding of International Symposium on Intelligent Multimedia, Video and Speech Processing. Hong Kong: [s.n.], 2004:222-225.

[6] Huang N E, Shen Z, Long S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-statitionary time series analysis [J]. Proceeding of the Royal Society A, 1998,454 (1971):903-995.

[7] Zou Xiaojie, Li Xueyao, Zhang Rubo. Speech enhancement based on Hilbert-Huang transform theory [C]//Proceeding of the First International Multi-Symposiums on Computer and Computation Sciences. Hangzhou, China: IEEE Computer Society Press, 2006.

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

[9] 潘欣裕，赵鹤鸣，陈雪勤，等. 基于 EMD 拟合特征的耳语话音端点检测[J]. 电子与信息学报，2008,30(2):362-366.
Pan Xinyu, Zhao Heming, Chen Xueqin, et al. Endpoint detection of whispers based on the fitting characteristic of EMD[J]. Journal of Electronics & Information Technology, 2008,30(2):362-366.

[10] Wu Z H, Huang N E. A study of the characteristics of white noise using the empirical mode decomposition method[J]. Proceeding of the Royal Society A, 2004,460(2046):413-418.

[11] Wu Z H, Huang N E. Ensemble empirical mode decomposition: a noise assisted data analysis method [J]. Advances in Adaptive Data Analysis, 2009,1: 41.

[12] Johnson D H, Sinanovic. Symmetrizing the Kullback-Leibler distance[R]. Technical Report, Rice University, 2001.

[13] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator[J]. IEEE Trans on Acoustics, Speech, and Signal Processing, 1985,33(2):443-445.

作者简介：吴其前(1981-)，男，博士研究生，研究方向：低速率语音编码，E-mail：wuqiqian_wqq@yahoo.com.cn；张雄伟(1965-)，男，教授，博士生导师，研究方向：数字通信、多媒体信息处理。