基于 HHT 和 OSF 的复杂环境语音端点检测

卢志茂*① 金 辉① 张春祥② 任明溪①
①(哈尔滨工程大学信息与通信工程学院 哈尔滨 150001)
②(哈尔滨理工大学软件学院 哈尔滨 150080)

摘 要：希尔伯特-黄变换是一种全数据驱动的自适应非平稳信号时频分析方法，但是在强噪声环境下语音信号的希尔伯特能量谱曲线波动较大，对语音端点检测造成很大的影响，该文提出了一种基于希尔伯特-黄变换和顺序统计滤波的检测方法。该方法将含噪语音信号进行经验模态分解，通过对固有模态函数进行自适应权重选取获得信号的希尔伯特能量谱，利用顺序统计滤波器对每帧的能量谱进行平滑处理作为语音/非语音的鉴别特征。实验结果表明，该方法适用于复杂噪声环境的端点检测，在低信噪比情况下仍然能够有效地检测出语音信号，降低信号误检率。

关键词：语音信号处理；端点检测；希尔伯特-黄变换；顺序统计滤波；经验模态分解
中图分类号：TN912.3 **文献标识码**： A **文章编号**：1009-5896(2012)01-0213-05
DOI: 10.3724/SP.J.1146.2011.00477

Voice Activity Detection in Complex Environment Based on Hilbert-Huang Transform and Order Statistics Filter

Lu Zhi-mao① Jin Hui① Zhang Chun-xiang② Ren Ming-xi①
①(*Information and Communication Engineering College*, *Harbin Engineering University*, *Harbin* 150001, *China*)
②(*School of Software*, *Harbin University of Science and Technology*, *Harbin* 150080, *China*)

Abstract: Hilbert-Huang Transform (HHT) is a fully data driven adaptive non-stationary signal time-frequency analysis method. But the Hilbert energy spectrum curve of speech signal is fluctuate in strong noise environment, it has a great influence to voice activity detection. So an effective voice activity detection algorithm is proposed based on HHT and Order Statistics Filter (OSF) in this paper. This method first decompose noise signal into intrinsic mode functions by empirical mode decomposition. Then the Hilbert energy spectrum is synthesized by adaptive weight selection of each intrinsic mode functions, through OSF to smooth the energy spectrum. Finally, the speech and noise divergence is judged by means of the smoothed energy spectrum. Experimental results show obviously that under complex noisy environment, this method is still able to effectively detect the speech signal, and reduce the error detection rate in low signal to noise ratio conditions.

Key words: Speech signal processing; Voice Activity Detection (VAD); Hilbert-Huang Transform (HHT); Order Statistics Filter (OSF); Empirical Mode Decomposition (EMD)

1 引言

在复杂的应用环境下从信号流中分辨出语音信号和非语音信号，是语音处理的一个基本问题。准确的语音端点检测不仅可以提高后续处理(如语音识别)的正确率和处理效率，还能够为后续处理提供段落分割的依据[1]。传统的基于短时能量和过零率的方法[2]只适用于弱噪声背景环境；后续有学者提出了基于谱熵[3]，隐马尔可夫模型[4]，小波变换技术[5]以及这些方法的改进算法等，但是在低信噪比以及复杂背景噪声环境下，其性能明显下降，也不能够满足后续处理的需求。

希尔伯特-黄变换(Hilbert-Huang Transform, HHT)是无需任何先验知识的时频分析方法，其分解依赖于信号本身，使数据的分解有真实的物理意义，并且具有更高的时频分辨率[6]。因此，该分析方法在分析非平稳非线性的语音信号方面有其独特的优势。目前很多专家学者致力于 HHT 端点检测方法的研究，也提出了一些好的改进算法，但随着信噪比的降低，噪声能量的加大带来了较高的误检率，语音的起止点也不够准确。顺序统计滤波(Order Statistics Filter, OSF)[7]方法能够有效利用语音的长时信息，克服语音波动较大，不利于阈值选择的

2011-05-19 收到，2011-09-05 改回
国家自然科学基金(60975042, 60903082)资助课题
*通信作者：卢志茂 lzm@hrbeu.edu.cn

缺点。本文提出一种基于 HHT 和 OSF 相结合的方法对语音信号进行检测，实验结果表明，本文算法在提高检测率的同时有效地降低了语音信号的误检率，取得了较好的效果。

2 Hilbert-Huang 变换(HHT)

1998 年美国 NASA 研究中心 Huang 等人[6]提出了一种新的时频分析方法 Hilbert-Huang 变换，该方法主要包含两大部分：经验模态分解(Empirical Mode Decomposition, EMD)和 Hilbert 谱分析(Hilbert Spectral Analysis, HSA)，其中经验模态分解是 HHT 的核心[8]。

EMD 方法是将待分解信号分解成若干固有模态函数(Intrinsic Mode Function, IMF)之和的形式。每个 IMF 必须满足以下两个条件：(1)在整个信号长度上，极值点数目和过零点数目必须相等或者至多只相差一个；(2)在任意一点处，由局部极大值点构成的上包络线和由局部极小值点构成的下包络线的均值为零。

3 顺序统计滤波

对含噪语音信号进行希尔伯特-黄变换后，根据信号的时-频-幅值特性，绘制 3 维希尔伯特能量谱图如图 1 所示，从图中可以看出，高频部分的噪声能量得到了抑制，但是低频部分的噪声仍然有很大的起伏波动，噪声的波动不利于检测阈值的选取，影响语音端点检测的结果，通过分析局部噪声的分布情况可以看到局部噪声分布类似微弱冲击信号，而顺序统计滤波器是一种有效的信号平滑处理方法，因此我们对希尔伯特能量谱进行顺序统计滤波。

3.1 OSF 原理

20 世纪 80 年代，顺序统计滤波器的概念被提出来，主要用来增强信号的平滑度。设含有 n 个信号的序列 $X_1, X_2, \cdots, X_n$，取窗口长度为 L，对此序列进行 L 阶的顺序统计滤波，就是从输入序列中相继抽取出 L 个数，$X_{i-v},\cdots,X_{i-1},X_i,X_{i+1},\cdots,X_{i+v}$，其中 i 为窗口的中心位置，$v=(L-1)/2$，将这 L 个数按照其数值大小进行升序排列，即有 $X_{(1)} \le X_{(2)} \cdots \le X_{(L)}$，则顺序统计滤波器的输出就是 X_i 的线性组合。

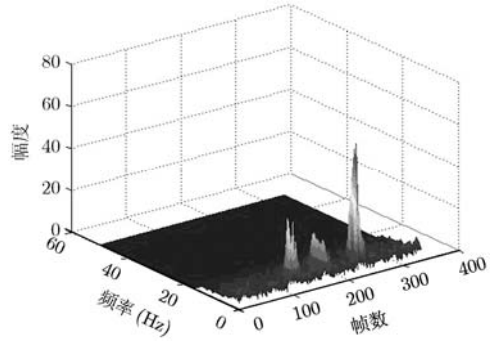

图 1 含噪语音信号 3 维希尔伯特能量谱

$$Y=\sum_{i=1}^{L} a_i X_{(i)} \tag{1}$$

系数 a_i 决定了滤波器的特性，根据不同的 a_i 可以得到不同类型的顺序统计滤波器[9]。

设 l 是当前分析的语音帧，$X_h[l]$ 是 $X[l-M],\cdots,X[l],\cdots,X[l+M]$ 中第 h 个最大值，经过滤波平滑处理后，第 l 帧的希尔伯特能量定义为

$$X[l]=(1-p)X_h[l]+pX_{h+1}[l] \tag{2}$$

其中 $h=\lfloor pL \rfloor (0<p<1, L=2M+1)$，$p$ 称为顺序统计滤波器的采样分位数，满足高斯分布。

3.2 OSF 的参数选取

已知 $X_1,X_2,\cdots,X_{2N+1}$ 是概率分布函数为 $f(x)$ 的均匀分布随机变量，根据顺序统计的渐进理论，当 M 足够大时，r 阶顺序统计滤波后 $X_{(r)}$ 的方差可以被近似为

$$\sigma^2_{X(r)}=\frac{p(1-p)}{f^2(F^{-1}(p))} \tag{3}$$

其中 $u_{(r)}$ 是 r 阶顺序统计滤波的均值，$f_{X(r)}(x)$ 是顺序统计量 $X_{(r)}$ 的概率分布函数，$F^{-1}(p)$ 是 p 的分位点函数。

为了更好地在带噪语音信号中进行端点检测，需要较高的分位数。但当取最大值 $X_{(2M+1)}$ 时，其方差较大，因此我们在高检测率和低误检率中折中选择 p=0.9。

4 结合 HHT 和 OSF 的端点检测

本文提出的方法主要针对复杂背景噪声环境中的低信噪比语音端点检测问题，利用 EMD 分解将信号分解成时间特征尺度由小到大的 IMF 分量，通过阈值去噪方法对每个 IMF 分量处理后进行 Hilbert 变换，得到信号的 Hilbert 能量谱，利用顺序统计滤波方法解决希尔伯特能量谱曲线的波动比较大，不利于阈值的选择的问题，通过判断噪声背景下的能量突变点，进行检测。

设输入的带噪的语音信号为 $x(t)=s(t)+n(t)$，其中 $s(t)$ 为纯净的语音信号，$n(t)$ 为噪声。算法具体步骤如下：

(1)通过预处理模块将带噪语音信号分割成相邻有重叠的信号帧；

(2)对每帧信号进行 EMD 分解，得到有限个固有模态函数；

(3)对 IMF 分量进行自适应权重选取；

$$\tilde{f}(x_i)=\begin{cases}0, & |x_i|<\xi \\ \mathrm{sgn}(x_i)\cdot(|x_i|-\xi), & |x_i|\ge\xi\end{cases} \tag{4}$$

其中 $\tilde{f}(x_i)$ 为经过处理后的IMF系数，ξ 代表软限幅函数的阈值，$\xi=\sigma\sqrt{2\lg(N)}$，σ 为噪声方差，N 为带噪的每个IMF的长度；

(4)对处理后的IMF逐个进行Hilbert变换，求解瞬时频率及幅值，并合成Hilbert能量谱；

(5)对合成的Hilbert能量谱进行顺序统计滤波，为了提高算法对端点检测的准确率，经由大量实验，本文选取 p=0.9, M=8；

(6)对语音信号进行背景噪声的估计。将最开始输入的前 N 帧作为无音片段，根据上述方法得到前 N 帧顺序统计滤波后的能量谱$X(l)$，用式(15)-式(17)进行语音检测的阈值T_S 设定：

$$E(X_l)=\frac{1}{N}\sum_{i=1}^{N}X_l(i) \tag{5}$$

$$D(X_l)=\frac{1}{N}\sum_{i=1}^{N}(X_l(i)-E(X_l))^2 \tag{6}$$

$$T_S=\alpha E(X_l)+\beta D(X_l) \tag{7}$$

式中 α, β 可以通过实验来灵活地调节，使得在提高检测率的同时尽量减少误减率。为使本文方法具有自适应性，不必根据噪声不同重新选取实验参数，经过大量实验统计，本文选取其在不同噪声环境下的均值：α= 5, β= 2 。

(7)通过阈值T_S 来判断语音段和非语音段，在纯净语音及带噪语音信号中标出，并计算准确率。

5　实验分析

在测试中，本文选用安静环境下实验室录制的一些短语和句子作为目标语音信号，采用标准噪声库 NOISEX-92 中的噪声作为复杂干扰噪声，采样率为 11.025 kHz，采样精度为 16 bit。为了验证算法的性能，实验对大量的语音数据进行检测统计，并与基于 HHT 的端点检测算法进行比较，采用检测率和误测率作为判别依据：

(1)检测率等于检测正确的语音信号帧/语音信号总帧数；

(2)误检率等于被误认为语音的噪声信号帧/非语音信号帧。

5.1 白噪声背景下检测结果

图 2 和图 3 中(a)为不加噪声的语音信号，(b)为添加高斯白噪声的语音信号，(c)为两种方法的希尔伯特能量谱，其中横线代表检测阈值。表 1 给出了白噪声情况下不同信噪比的检测结果，两种方法分别是：基于 HHT 的语音端点检测方法、本文方法，通过实验结果可以看出，本文方法对于低信噪比下的端点检测非常有效，经过顺序统计滤波后的能量分布更加明显，在检测率上有明显的提升，并且大大地降低了信号的误测率，有利于信号的后续处理。文献[10]中提出的方法在 0 dB 时能达到 89.7% 的正确率，但是随着信噪比的降低，该方法已经不能很好地检测出语音信号。本文方法在信噪比为−10 dB 的情况下检测率仍能够达到 85.05%，是一种有效的低信噪比下语音端点检测方法。

5.2 复杂噪声背景下检测结果

图 4 和图 5 中(a)为纯净语音信号，(b)为添加 f16 噪声后的检测结果，(c)为添加 pink 噪声后的检测结果，(d)为添加 factory1 噪声后的检测结果。

表 2 和表 3 分别给出不同信噪比情况下各种噪声的检测率、误检率。

由实验结果可知，对于复杂噪声环境下的端点检测，基于 HHT 的语音端点检测算法随着信噪比的降低，检测率逐渐下降，并且带来了较大误测率。而本文提出的方法在信噪比为 5 dB 和 0 dB 时是检

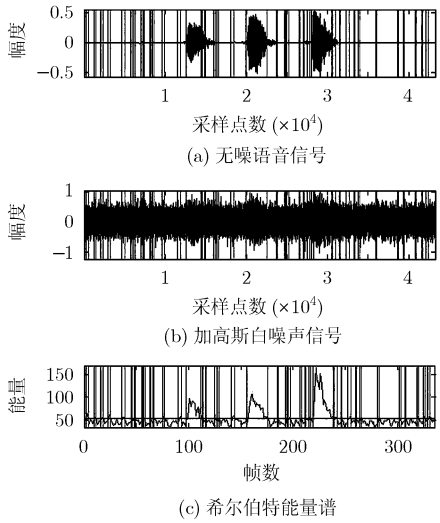

图 2 基于 HHT 的端点检测方法(SNR=−10 dB)

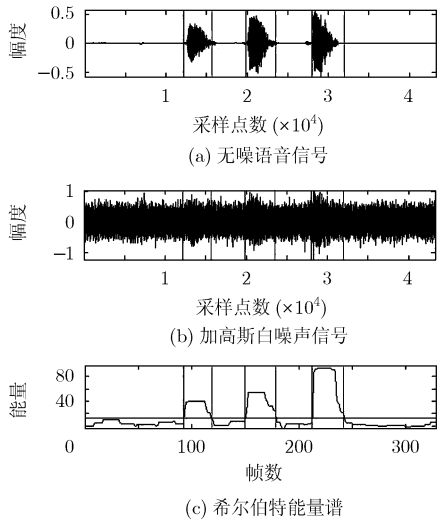

图 3 本文改进算法(SNR=−10 dB)

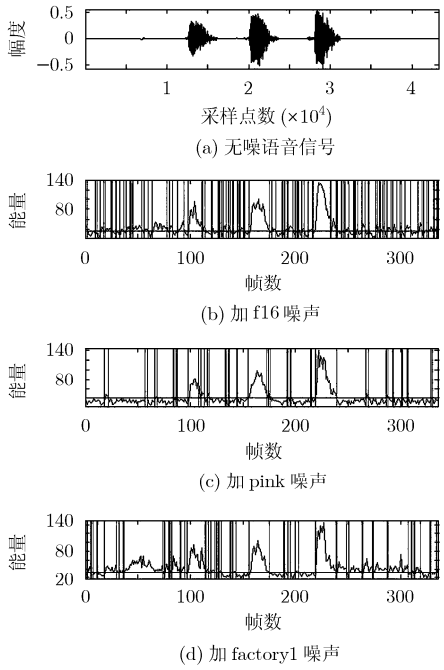

图 4 基于 HHT 方法不同噪声的端点检测结果(SNR=−5 dB)

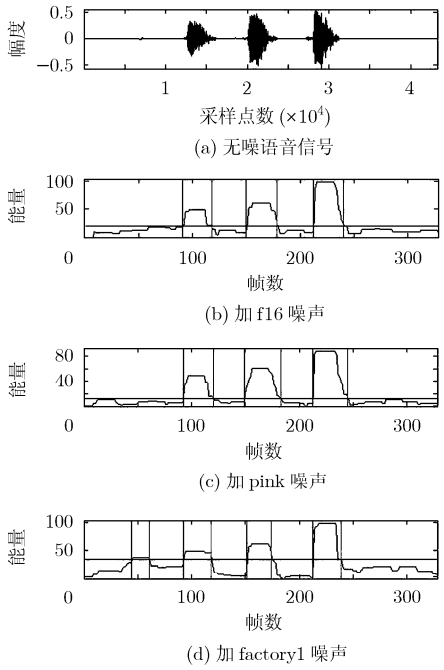

图 5 本文算法对不同噪声的端点检测结果(SNR=−5 dB)

表 1 白噪声下基于 HHT 方法和本文算法的结果比较

SNR(dB)	检测率(%)		误检率(%)	
	HHT 方法	本文方法	HHT 方法	本文方法
5	92.81	95.71	2.07	0.63
0	89.03	92.84	3.11	0.84
−5	85.26	89.36	3.53	0.95
−10	79.37	85.05	4.29	1.36

表 2 基于 HHT 方法和本文算法的检测率比较(%)

SNR (dB)	f16 噪声		pink 噪声		factory 噪声	
	HHT 方法	本文方法	HHT 方法	本文方法	HHT 方法	本文方法
5	91.57	94.11	93.25	95.83	92.64	94.72
0	86.92	90.18	90.14	92.66	87.21	91.29
−5	81.34	85.23	87.90	88.13	82.85	87.60
−10	72.46	74.39	82.77	83.97	76.50	79.21

表 3 基于 HHT 方法和本文算法的误检率比较(%)

SNR(dB)	f16 噪声		pink 噪声		factory 噪声	
	HHT 方法	本文方法	HHT 方法	本文方法	HHT 方法	本文方法
5	1.93	0.84	1.47	0.79	2.17	1.95
0	2.57	1.38	1.92	1.14	3.46	2.84
−5	3.08	1.72	2.48	1.69	5.01	3.06
−10	4.44	2.30	2.71	1.90	6.58	4.49

测率均能达到 90%以上，效果非常明显，−5 dB 时也有较高的准确率。随着信噪比的降低，噪声能量逐渐加大，检测效果逐渐下降，但是在相对 HHT 方法还是有一定的提高；在误检率方面 f16 噪声和 pink 噪声的误检率都有明显的降低，factory1 噪声对于−10 dB 的情况下误检率有待进一步提高，从总体结果上看，本文提出的方法对于低信噪比复杂环境下的端点检测结果令人满意。

6 结论

本文提出了一种基于 HHT 和 OSF 的复杂噪声环境语音端点检测方法，该方法以希尔伯特能量谱作为语音和非语音的鉴别特征，利用顺序统计滤波器平滑各帧的希尔伯特能量谱，使得语音和非语音的过渡带窄且陡峭，通过自适应阈值判别语音信号。实验结果表明，该方法适用于复杂的噪声环境的端点检测，在低信噪比情况下仍然能够有效地检测出语音信号，并且对信号误检有很好的抑制作用，在各种噪声环境下均具有良好的鲁棒性。

参 考 文 献

[1] Ghosh P K, Tsiartas A, and Narayanan S. Robust voice activity detection using long-term signal variability [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2011, 19(3): 600–613.

[2] 韩志艳, 王旭, 王健. 基于短时能零积和鉴别信息的语音端点检测[J]. 东北大学学报(自然科学版), 2009, 30(12): 1690–1693.

Han Zhi-yan, Wang Xu, and Wang Jian. Speech endpoint detection algorithm based on short time energy zero product and discrimination information [J]. *Journal of Northeastern University* (*Natural Science*), 2009, 30(12): 1690–1693.

[3] 刘华平, 李昕, 郑宇, 等. 一种改进的自适应子带谱熵语音端点检测方法[J]. 系统仿真学报, 2008, 20(5): 1366–1371.
Liu Hua-ping, Li Xin, Zheng Yu, *et al*.. Speech endpoint detection based on improved adaptive band-partitioning spectral entropy [J]. *Journal of System Simulation*, 2008, 20(5): 1366–1371.

[4] Othman H and Abounasr T. A semi-continuous state transition probability HMM-based voice activity detection [C]. IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal, Quebec, Canada, May 17–21, 2004, 5: V821–V824.

[5] Mohadese Eshaqhi and Karami Mollaei M R. Voice activity detection based on using wavelet packet [J]. *Digital Signal Processing*, 2010, 20(4): 1102–1115.

[6] 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]. *Proceedings of the Royal Society A*, 1998, 454: 903–995.

[7] Ramirez J, Segura J C, Benitez C, *et al*.. An effective subband OSF-Based VAD with noise reduction for robust speech recognition [J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2005, 13(6): 1119–1129.

[8] 曲从善, 路廷镇, 谭影. 一种改进型经验模态分解及其在信号消噪中的应用[J]. 自动化学报, 2010, 36(1): 67–73.
Qu Cong-shan, Lu Ting-zhen, and Tan Ying. A modified empirical mode decomposition method with applications to signal de-noising[J]. *Acta Autmatica Sinica*, 2010, 36(1): 67–73.

[9] Celebi M E. Alternative distance/similarity measures for reduced ordering based nonlinear vector filters[C]. 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing, Dallas, TX, March 14–19, 2010: 1266–1269.

[10] 钱颜旻, 刘加. 基于交叉熵顺序统计滤波的语音端点检测算法[J]. 清华大学学报(自然科学版), 2009, 49(10): 87–90.
Qian Yan-min and Liu Jia. Cross-entropy OSF-based voice activity detection algorithm [J]. *Journal of Tsinghua University* (*Science and Technology*), 2009, 49(10): 87–90.

卢志茂：　男，1972 年生，博士，教授，研究方向为模式识别、机器视觉听觉.

金　辉：　女，1986 年生，硕士生，研究方向为智能信息与语音信号处理.

张春祥：　男，1972 年生，博士，副教授，研究方向为模式识别与人工智能.

任明溪：　男，1987 年生，硕士生，研究方向为智能信息与语音信号处理.