基于 Hilbert-Huang 变换和听觉掩蔽的语音增强算法

宋倩倩，于凤芹
(江南大学通信与控制工程学院，江苏无锡 214122)

摘要：Hilbert-Huang 变换是一种新型的具有自适应性的时频分析方法，分析了 HHT 算法的原理，提出了一种基于 HHT 和听觉掩蔽的语音增强算法，首先将语音信号进行 EMD 分解得到各阶 IMF 分量，然后对高频 IMF 分量进行听觉掩蔽处理，最后将处理后的分量与剩余分量叠加得到重构信号。仿真结果表明所提出的算法降低了语音失真测度值，提高了语音信号的信噪比、清晰度及可懂度，并与听觉掩蔽算法和谱减法进行了比较，显示了该算法的优越性。

关键词：Hilbert-Huang 变换；经验模态分解；固有模态函数；听觉掩蔽；语音增强
中图分类号：TN912 **文献标识码**：A **文章编号**：1000-3630(2009)-03-0280-04
DOI 编码：10.3969/j.issn1000-3630.2009.03.018

Speech enhancement based on Hilbert-Huang tansform and human auditory masking

SONG Qian-qian, YU Feng-qin
(*College of Communications and Control Engineering*, *Jiangnan University*, *Wuxi* 214122, *Jiangsu*, *China*)

Abstract: HHT is a new and self-adaptable method for time-frequency analysis. The theory of HHT is studied and a speech enhancement method based on HHT and human auditory masking is brought forward. First the signal is decomposed into IMFs with the method of EMD, then the high-frequency IMFs is processed with the human auditory masking, finally the signal is reconstructed by adding the treated IMFs with the residual IMFs. Simulation experiments show that it can reduce the measured value of speech distortion and improve the SNR, speech articulation and intelligibility. The results display the superiority of this method over the human auditory masking and the spectral subtraction method.

Keywords: HHT; empirical mode decomposition; intrinsic mode function; human auditory masking; speech enhancement

1 引 言

语音作为一种非平稳信号在实际环境中会受到噪声的干扰，使得语音处理系统不能正常工作，因此抑制背景噪声、改善输出信噪比、提高语音通信质量这一要求使得语音增强具有重要的应用价值，且直接影响到语音识别、语音编码等后续研究工作。到目前为止的语音增强算法主要有基于语音生成模型、短时谱估计、小波分解、听觉掩蔽的算法等[1]，但上述方法的缺点是没有抓住语音信号非线性、非平稳这一本质特征。对于非平稳信号，时频分析是有效的分析方法，短时傅立叶变换(Short-Time Fourier Transform, STFT)、小波分析、Wigner-Ville 分布等不同程度上对非平稳非线性给予了恰当的描述[2]，但总体上来说还是基于 Fourier 变换的，因此受到 Fourier 变换的限制。为此，N.E.Huang 提出了一种新的时频分析方法——Hibert-Huang 变换(Hilbert-Huang Transform，HHT)[3]，HHT 是一种新的具有自适应的时频分析方法，它可根据信号的局部时变特征进行自适应的时频分解，其分析结果能够真实地描述信号的物理特性，本文将 HHT 这一新方法用于语音信号处理中，提出了一种基于 HHT 和听觉掩蔽的语音增强算法。本文算法是将含噪语音信号经过 EMD 分解后，对高频的 IMF 分量进行听觉掩蔽处理，然后将处理后的 IMF 分量与未经处理的低频 IMF 分量进行叠加，重构出增强后的信号。仿真实验表明，本文算法对各种信噪比下的语音信号均有较好的增强效果，并与听觉掩蔽算法和谱减法进行了比较，提高了语音信号的清晰度与可懂度，具有一定的优越性。

收稿日期：2008-07-17；**修回日期**: 2008-10-21
作者简介：宋倩倩(1986-)，女，山东人，硕士研究生，研究方向为非平稳信号时频分析、语音信号处理。
通讯作者：宋倩倩, E-mail: songqianqian1025@yahoo.cn

2　基于 HHT 和听觉掩蔽的语音增强算法

2.1　Hilbert-Huang 变换

Hilbert-Huang 变换包含两个过程：经验模态分解(Empirical Mode Decomposition，EMD)和 Hilbert 变换。其中最关键的是 EMD 过程，它是基于信号的局部特征时间尺度，能把复杂的信号分解为有限的固有模态函数(Intrinsic Mode Function，IMF)之和，IMF 必须满足以下两个条件：(1)整个信号长度上，一个 IMF 的极值点和过零点的数目必须相等或至多只相差一点。(2)任意时刻，由局部极大值点形成的上包络线和由局部极小值点形成的下包络线的平均值为零，即上下包络线相对于时间轴局部对称。对于给定的信号 $s(t)$，过 EMD 可以将信号分解为 n 个 IMF 和残余函数 $r_n(t)$ 之和：

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

对式(1)中的每个固有模态分量 $c_i(t)$ 作 Hilbert 变换可得到 $\hat{c}_i(t)=\frac{1}{\pi}p.v\int_{-\infty}^{+\infty}\frac{c_i(\tau)}{t-\tau}\mathrm{d}\tau$，其中，$p.v$ 表示取积分主值，然后构造解析信号：

$$z_i(t)=c_i(t)+\mathrm{j}\hat{c}_i(t)=a_i(t)\mathrm{e}^{\mathrm{j}\varphi_i(t)} \tag{2}$$

其中

$$a_i(t)=\sqrt{c_i^2(t)+\hat{c}_i^2(t)}\qquad \varphi_i(t)=\arctan\frac{\hat{c}_i(t)}{c_i(t)}$$

进一步可求得瞬时频率为：

$$f_i(t)=\frac{1}{2\pi}\omega_i(t)=\frac{1}{2\pi}\times\frac{\mathrm{d}\varphi_i(t)}{\mathrm{d}t} \tag{3}$$

解析信号的极坐标形式(2)反映了 Hilber 变换的物理含义：它是通过一正弦曲线的频率和幅值调制获得信号局部的最佳逼近。式(2)中将瞬时频率定义为相位的导数而不需要整个波来定义局部频率，值得注意的是它与通常用 Hilbert 变换中将频率定义为相位的导数有相似之处，但是两者的数学意义和物理意义是完全不同的[3]。这样可得到:

$$s(t)=\mathrm{Re}\sum_{i=1}^{N}a_i(t)\mathrm{e}^{\mathrm{j}\varphi_i(t)}=\mathrm{Re}\sum_{i=1}^{N}a_i(t)\mathrm{e}^{\mathrm{j}\int\omega_i(t)\mathrm{d}t} \tag{4}$$

展开式(4)称为 Hilbert 谱，记作 $H(\omega,t)=\mathrm{Re}\sum_{i=1}^{N}a_i(t)\mathrm{e}^{\mathrm{j}\int\omega_i(t)\mathrm{d}t}$，因为函数 $r_n(t)$ 通常是一个常数或是一个单调分量，不能反映出信号的振荡模式，因此对不对其进行 Hilbert 变换处理。

2.2　听觉掩蔽算法

所谓听觉掩蔽现象是指当两个强度不同的声音作用于人耳时，一种声音成分由于另一种声音成分的存在而不被人所感知。听觉掩蔽分为时域掩蔽和频域掩蔽，虽然邻近声音的时域掩蔽已经被证明很有用处(特别是对于宽带音频编码)，由于很难对时域掩蔽进行定量分析，因此其在语音处理领域的应用还不够广泛。本文将利用掩蔽原理在频域中进行处理。一般来说，对于中等掩蔽强度，纯音最有效的掩蔽出现在它的频率附近，低频的纯音可以有效的掩蔽高频的纯音，而高频的纯音对低频的纯音的掩蔽作用较小，这是听觉掩蔽的基本原理。利用听觉掩蔽进行噪声抑制的算法可以描述为以下几步：(1)过平均非语音帧的能量得出背景噪声的功率谱估计,供谱减滤波器使用。(2)用 Johnston 提出的在各语音帧中计算掩蔽门限的方法[1]，从短时语音谱中计算每一帧的掩蔽门限曲线，记为 $T(pL,\omega)$。(3)减法语音增强的表达式为：

$$S(m,k)=\begin{cases}\left(1-\alpha\left[\dfrac{|N(m,k)|}{|Y(m,k)|}\right]^2\right)^{\frac{1}{2}}\times Y(m,k) & \left[\dfrac{|N(m,k)|}{|Y(m,k)|}\right]^2<\dfrac{1}{\alpha+\beta}\\ \left(\beta\left[\dfrac{|N(m,k)|}{|Y(m,k)|}\right]^2\right)^{\frac{1}{2}}\times Y(m,k) & \text{其它}\end{cases}$$

其中 $Y(m,k)$ 和 $N(m,k)$ 为带噪语音信号和噪声的频谱，α 和 β 分别为谱减阈值系数和谱减噪声系数。根据掩蔽曲线 $T(pL,\omega)$，对谱减滤波器的参数 α 和 β 进行调整。(4)利用第(3)步得到的谱减滤波器进行噪声抑制，然后合成出语音信号。

2.3　本文算法

本文针对高斯白噪声提出了一种基于 HHT 和听觉掩蔽的语音增强算法。实验中采用高斯白噪声污染语音信号，因为白噪声是概率均值为零、方差为常数的随机信号，其尺度一般比较小，采用 EMD 分解信号得到的 IMF 分量具有时间特征尺度由小到大即频率由高到低的特性。因此当用 EMD 分解受白噪声污染的语音信号时，白噪声主要集中在首先分解出的尺度小的 IMF 分量中，小尺度 IMF 分量即为高频的 IMF 分量。由于宽带噪声的频谱分布与语音频谱重叠，如果对高频的 IMF 分量直接采用低通尺度滤波则会丢失有用信息，因此本文算法对高频的 IMF 分量采用听觉掩蔽进行处理，对高频的 IMF 分量从时域转换到频域，然后计算其所对应的听觉掩蔽阈值，利用阈值调整谱减阈值系数 α 和谱减噪声系数 β，求得增强后的语音谱，进行傅立叶反变换即从频域转换到时域，即可得到增强后的高频 IMF 分量。最后将高频 IMF 分量与剩余 IMF 分量叠加重构增强后的语音信号。利用本文算法对含噪信号 $s(t)$ 实现语音增强的具体步骤如下：

(1) 确定信号所有的局部极值点，然后用三次样条线将所有的局部极大值点和局部极小值点分别连接起来形成上包络线和下包络线，上、下包络线应包含所有的数据点。

(2) 下包络线的平均值记为 $m_1(t)$，求出 $s(t)-m_1(t)=h_1(t)$，如果 $h_1(t)$ 满足 IMF 的条件则它就是第一个 IMF 分量。如果 $h_1(t)$ 不满足 IMF 的条件，则把 $h_1(t)$ 作为原始数据，重复上述步骤，得到上下包络线的平均值 $m_{11}(t)$，再判断 $h_{11}(t)=h_1(t)-m_{11}(t)$ 是否满足 IMF 的条件，如果不满足，则重复循环 k 次，得到 $h_{1(k-1)}(t)-m_{1k}(t)=h_{1k}(t)$，使得 $h_{1k}(t)$ 满足 IMF 的条件，记 $c_1(t)=h_{1k}(t)$，则 $c_1(t)$ 为信号 $s(t)$ 的第一个满足 IMF 条件的分量。

(3) 从 $s(t)$ 中分离出来，得到 $r_1(t)=s(t)-c_1(t)$，将 $r_1(t)$ 作为原始数据重复步骤(1)~(2)，第 2 个满足 IMF 条件的分量 $c_2(t)$，重复循环 n 次，得到信号 $s(t)$ 的 n 个满足 IMF 条件的分量，当 $r_n(t)$ 成为一个常数或是单调函数而不能再从中提取出满足 IMF 条件的分量时，循环结束。信号被分解为 n 个 IMF 和残余函数 $r_n(t)$ 之和 $s(t)=\sum_{i=1}^{n}c_i(t)+r_n(t)$，为了保证固有模态分量能够反映物理意义上实际的幅度和频率调制，本算法中采用以下的停止准则：

$$\sqrt{\sum_{i=1}^{N}\left|\left(h_{1(k-1)}-h_{1k}\right)\right|_i^2}<\varepsilon, k=1,2\cdots$$

其中，N 离散信号序列的总长度，k 为重复次数，可取 0.2~.3 之间的一个值。

(4) 频分量进行听觉掩蔽处理。利用听觉掩蔽阈值公式计算阈值 $T(m,k)$, $T(m,k)=\max(T'(m,k), T_a(m,k))$, $T'(m,k)$ 为每帧的听觉掩蔽阈值，可由 Johnston 提出的在各语音帧中计算掩蔽阈值的方法求得[1]，$T_a(m,k)$ 为绝对掩蔽阈值，计算公式为：

$$T_a(m,k)=3.64f^{-0.8}-6.5\exp(f-3.3)^2+10^{-3}f^4$$

然后根据谱减系数的计算公式用听觉掩蔽阈值调整谱减系数 α 和 β。谱减系数的计算公式为：

$$\frac{T_{\max}-T(m,k)}{\alpha-\alpha_{\min}}=\frac{T(m,k)-T_{\min}}{\alpha_{\max}-\alpha}$$

$$\frac{T_{\max}-T(m,k)}{\beta-\beta_{\min}}=\frac{T(m,k)-T_{\min}}{\beta_{\max}-\beta}$$

其中 $T_{\max}$ 和 $T_{\min}$ 是每一语音帧的听觉掩蔽阈值的最大值和最小值。听觉掩蔽阈值大则说明在此频段中人耳对其他相近的频率段的语音信号和噪声信号的抗干扰能力比较强，此时采用较小的谱减系数，如果掩蔽阈值比较小则采用较大的谱减系数。实验表明在兼顾信噪比和听觉质量的前提下，选取 $\alpha_{\max}=6$ 和 $\alpha_{\min}=1$，$\beta_{\max}=0.02$ 和 $\beta_{\min}=0$。

(5) 于各 IMF 分量具有正交性和完备性，因此将经过听觉掩蔽处理后的 IMF 分量和剩余的 IMF 分量进行叠加得到增强后的语音信号。

3 仿真实验

实验中选取了 80 条包括短语和语句的语音样本，以内容为女声“经济生活”为例，采样频率为 11025Hz，将这段纯语音信号叠加高斯白噪声 $N(0,\sigma^2)$，改变方差 σ^2 的值，构成不同信噪比下的语音信号。图 1 为信噪比为 1dB 下的含噪语音信号，图 2 为对含噪语音信号进行 EMD 分解所得到的各阶 IMF 分量。从图 2 可以看出 IMF1~MF4 包含了信号的高频成分，噪声基本包含在这些分量中，本文算法对较高频的 IMF 分量采用听觉掩蔽进行处理，然后将处理后的 IMF 分量与低频 IMF 分量进行叠加，重构增强后的语音信号。

为了表示重构信号和原始信号之间的相似度，定义重构误差为：$E=\frac{1}{N}(\sum_{i=0}^{N}(s(i)-\hat{s}(i))^2)^{1/2}$，其中 N 为信号长度，$\hat{s}(i)$ 为重构信号，根据以上公式得到的重构误差 E =7.3205e-4，这就说明重构信号和原始信号极为相似，信号失真很小。

实验中采用信噪比和 Itakura-Saito 语音失真测度(IS)两个客观测试指标以及平均意见得分(Mean Opinion Score, MOS)主观方式。信噪比和语音失真测度的主要区别是信噪比不能反映人们对语音信号的听觉质量而语音失真测度在一定程度上可以反应人们对语音信号的主观感受，其系数越小，说明语音的品质越好。

考虑到噪声样本的随机性，表 1 是数据为 20 次不同噪声下的均值。表 1 中 A 为本文算法；B 为听觉掩蔽算法；C 为传统的谱减法。从表 1 中可以看出，在不同输入信噪比下，本文算法都比听觉掩蔽算法和传统的谱减法更好地提高了输出信噪比。从语音失真测度值上看，基于本文算法的语音增强可以使有用的语音能充分保留下来，保证语音的质量，

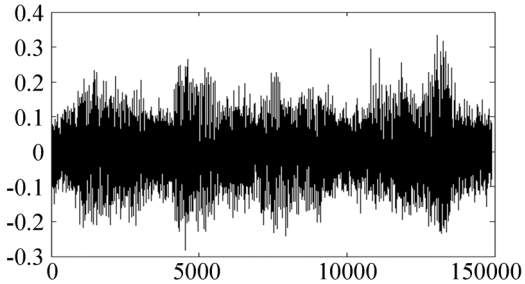

图 1 含噪语音信号

Fig.1 Noisy speech signal

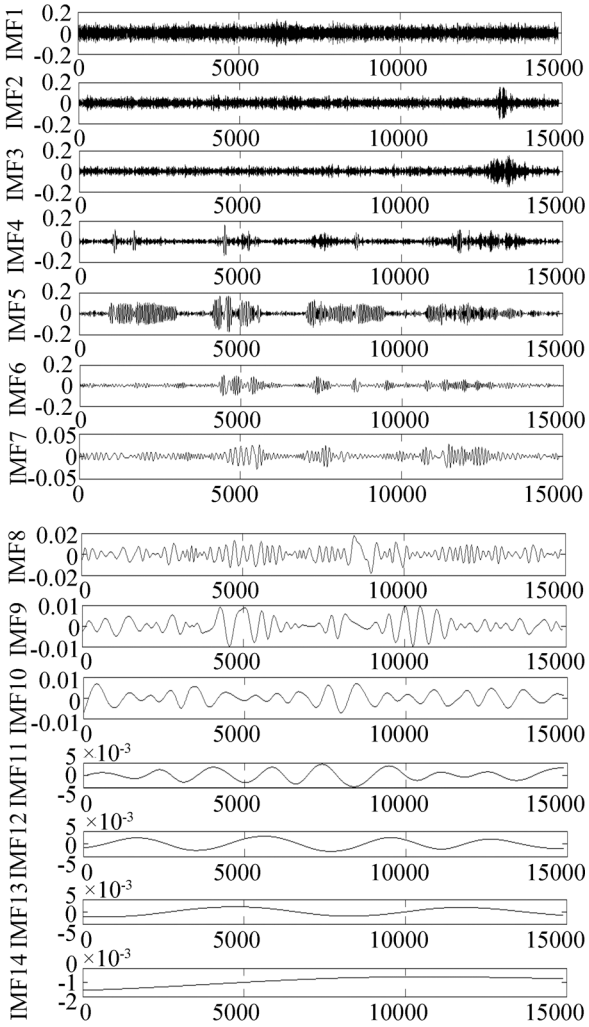

图 2　含噪语音信号的各阶 IMF 分量
Fig.2　The IMFs of noisy speech signal

表 1　算法的比较
Table 1　The comparision of algorithms

输入信噪比/dB	输出信噪比/dB			语音失真测度/IS		
	A	B	C	A	B	C
-10.02	3.11	0.76	0.12	2.21	3.65	4.34
-5.06	5.51	3.89	2.91	1.86	2.65	3.98
0.05	8.73	6.21	4.74	1.26	2.17	3.02
5.04	11.01	9.36	8.63	1.14	1.92	2.21
10.09	14.59	13.52	12.90	0.93	1.54	1.99

使语音失真度达到更低水平。

为了确认客观性评价，实验中采取了非正式的测听方法，5 位同学参加主观测试，根据被测声音的残留噪声、语音清晰度和可懂度等情况给出综合评分，将每人所给出的 MOS 评分求平均值作为其最后的得分，如表 2 所示。

对比主观测试结果可以看出，本文算法的性能优于听觉掩蔽算法和谱减法，经过主观试听可以感觉到音乐噪声和背景噪声明显减小，本文算法显示出优越性。

表 2　主观 MOS 评分比较
Table 2　The comparision of subjective MOS

输入信噪比	MOS 评分		
	本文算法	听觉掩蔽算法	传统谱减法
-10.02	2.1	1.5	1.2
-5.06	2.9	2.3	1.8
0.05	3.4	2.8	2.4
5.04	3.9	3.3	2.9
10.09	4.4	4.0	3.8

4　结 语

本文分析了 Hilbert-Huang 变换这一处理非线性、非平稳信号的新方法，并将它应用于典型的非平稳信号－语音信号中，提出了基于 HHT 和听觉掩蔽的语音增强算法。仿真结果证明了本文算法的可行性和有效性，为语音信号处理的后续研究工作提供了有效手段。

参　考　文　献

[1] Thomas F. Quatieri. 离散时间语音信号处理-原理与应用[M]. 北京: 电子工业出版社, 2004: 526-555.
Thomas F. Quatieri. discrete-time speech signal processing principles and practice[M]. Beijing:Publishing House of Electronics Industry, 2004: 526-555.

[2] 张贤达, 保铮. 非平稳信号分析与处理[M]. 北京: 国防工业出版社, 1998: 12-49.
ZHANG Xianda, BAO Zheng. The analysis and processing of non-stationary signal[M]. Beijing:Publishing House of National Defense Industry, 1998: 12-49.

[3] Norden E. Huang, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstatin-ary time series and analysis[J]. Proceedings of the Royal Society of London Series, 1998, 454: 903-995.

[4] 蔡汉添, 袁波涛. 一种基于听觉掩蔽模型的语音增强算法[J]. 通信学报, 2002, **23**(8): 93-98.
CAI Hantian, YUAN Botao. A speech enhancement algorithm based on masking properties of human auditory system[J]. Journal of China Institute of Communications, 2002, **23**(8): 93-98.

[5] Stephen D. Howard J. Chizeck. Some properties of an empirical mode type signal decomposition algorithm[J]. IEEE International Conference on March 31 2008, 2008, **8**(9): 3625-3628.

[6] ZOU Xiaojie, LI Xueyao. Speech enhancement based on Hilbert-Huang transform theory[J]. Computer and Com- putional Sciences, 2006, **4**(1): 208-213.

[7] REN Zhong, ZHANG Haiyong. Research on properties of Hilbert spectrum[C]. The Eighth Internatonal Conference on Electronic Measurement and Insrruments, 2007, **7**(1): 322-325.