基于 EMD 拟合特征的耳语音端点检测

潘欣裕 赵鹤鸣 陈雪勤 徐 敏
(苏州大学电子信息学院 苏州 215021)

摘 要：耳语音作为人类发音的一种特殊形式，与正常语音相比具有信噪比低、元音的周期特征不明显等特性，因而耳语音处理比正常语音更为困难。耳语音处理研究的第 1 个关键步骤就是语音的端点检测，该文利用希尔伯特-黄变换(Hilbert-Huang Transform, HHT)中的经验模态分解(Empirical Mode Decomposition, EMD)，首次提出了一种基于 EMD 拟合特征的耳语音端点检测新方法。利用 EMD 得到的内禀模态函数(Intrinsic Mode Function, IMF)能量，以其归一化拟合参数为耳语音端点检测的特征，可以准确地划分出耳语音端点。实验表明，该方法在耳语音端点检测中取得了很好的效果，在 1200 个信噪比为 2~10dB 的测试样本中，检测准确率为 98.25%。

关键词：希尔伯特-黄变换；经验模态分解；内禀模态函数；归一化拟合特征
中图分类号：TN912.3 **文献标识码**：A **文章编号**：1009-5896(2007)02-0362-05

Endpoint Detection of Whispers Based on the Fitting Characteristic of EMD

Pan Xin-yu Zhao He-ming Chen Xue-qin Xu Min
(*School of Electronics and Information Engineering*, *Soochow University*, *Suzhou* 215021, *China*)

Abstract: Whispered speech is the especial form of people's pronunciation. There is lower Signal-to-Noise Ratio (SNR) in whispers and unobvious pitch waveform compared with the normal speech, so it is more difficult to process the whispered speech. The endpoint detection of whispers is the first pivotal step of whispered speech signal processing. This paper uses the Empirical Mode Decomposition (EMD) of Hilbert-Huang Transform (HHT) to solve the problem, and firstly proposes a novel algorithm of endpoint detection of whispered speech based on the fitting characteristic of EMD. Normalize the energy of Intrinsic Mode Function (IMF) obtained by EMD, and use the fitting parameters of the energy as the characteristic and then the endpoint of whispers can be easily divided. The results of experiments show that it is very useful in endpoint detection of whispers, and the accurate rate is 98.25% in 1200 samples (SNR=2~10dB).

Key words: Hilbert-Huang Transform (HHT); Empirical Mode Decomposition (EMD); Intrinsic Mode Function (IMF); Fitting characteristic of normalized energy

1 引言

语音信号的端点检测是语音预处理技术的关键步骤，其准确性直接影响到整个语音处理及识别系统的性能。语音的端点检测技术，伴随着计算机数字语音处理技术的发展而越来越成熟，涌现出了许多有效的算法。目前用于语音端点检测的算法主要有：传统的基于短时能量和短时过零率的双门限法[1]；根据频域参数的信息熵值法[2]；倒谱距离测量法[3]和残差方差与 LPC 谱能比例距离测度法[4]；基于语音混沌和湍流特性的分形维数测度法[5]；基于模型的人工神经网络方法[6]和隐 Markov 模型(HMM)方法[7]等等。这些语音检测手段有些基于数字信号处理理论，有些是基于数学模型，也有些基于语音本身的声学特性和人的发声机理，并应用于不同的场合中，取得了很好的效果。

耳语音作为人类语言发音的一种特殊形式，也越来越受到国内外学者的重视[8, 9]。同时利用数字信号处理的手段研究耳语音，对于在信号处理领域研究人类的发音规则与人耳的听觉感知特性，有着重大的意义。作为耳语音研究的第 1 个步骤，也是关键的一个步骤就是耳语音的端点检测。耳语音由于信噪比非常低，且由于其发音方式所致而无基音频率，因而正常语音端点检测的方法将无法适合。已有学者利用耳语音频域熵值法进行耳语音的分割[10]，取得了一定的成效。而本文根据语音信号的非平稳非线性特性，利用一种自适应的时频分析方法——希尔伯特-黄变换(Hilbert Huang Transform, HHT)[11]，其中的经验模态分解(Empirical Mode Decomposition, EMD)，将语音信号分解成若干个内禀模态函数(Intrinsic Mode Function, IMF)，再利用这几个 IMF 能量归一化后的分段拟合特征作为耳语音端点检测的参数。

2006-07-10 收到，2007-01-12 改回
国家自然科学基金(60572076)资助课题

HHT 应用于语音增强[12]、基音检测[13]已有许多成功的案例，实践证明这是一种有效的非平稳信号分析理论。由于 EMD 是根据信号本身固有的特征自适应的分解，并且分解前不对语音作短时分帧处理，所以也不需基于语音的短时平稳假设，更符合了语音信号的特性，并取得到了良好的实验结果，实验选择了南京大学的耳语音库中的 1200 个语音进行测试[14]，正确率为 98.25%。

2　EMD 应用于耳语音端点检测

2.1 耳语音的特点

人体的发音器官主要由肺、气管、喉、声带、咽、口和鼻组成。汉语耳语音的清辅音与汉语正常语音的发音方式没有很大的区别，而对于浊辅音或者元音则有较大的区别。正常发音时，肺部气流通过声带靠拢形成的窄缝声门，形成准周期的声源激励；耳语音发音时，声门一直保持半开的状态，气流通过时声带不振动，故声源为噪声激励。由于耳语音声韵母激励源皆为噪声，语音气流的湍流特性的改变造成耳语音声道特性的变化，并由共振峰表现出来，与正常音相比耳语音的第一共振峰偏移 1.3–1.6 倍，第二共振峰偏移 1–1.2 倍[15]，且带宽均有不同程度的拓宽，频谱更为平坦。

虽然耳语音较正常语音失去了一些明显的特性，但是人们还是可以识别出语音，感知出其声高，故而耳语音仍然携带了语音作为人类最重要交流方式的特点。根据声音由振动产生的原理，必然可以通过对已知语音数据的分析，得出人产生耳语音时的不同发音状态，从而进一步了解人的发音机理。鉴于此，利用已成功应用于振动模态分解的 EMD 方法，对耳语音信号进行自适应的分离研究，是一种值得探索的手段。

2.2 HHT 和 EMD 的原理及实现

HHT 是由 Huang 等人于 1998 年提出的一种新的时频分析手段[11]，HHT 主要分为两个步骤，一是 EMD，其利用自适应的极点提取拟合方法，筛选出每个 IMF；二是将每个 IMF 与其 Hilbert 变换构成一个解析复函数，并由此导出瞬时增益和瞬时频率，进一步得到信号的时频分布关系，称为信号的 Hilbert 谱分析。

由 EMD 分解得到的 IMF 必须满足两个条件：(1)对于整个信号段内，其极值点和过零点的数目相同或至多相差 1；(2)在信号中的任意一个采样点，由其局部极大点拟合插值构成的上包络线与其局部极小点拟合插值构成的下包络线的均值为 0。

EMD 的目的主要有两个：一是将原信号中的主要叠加波进行分离；二是使波的包络更加对称。因此，EMD 的具体实现可以归纳为以下步骤：

(1)先分别找出整体信号的局部极大值和极小值，利用三次样条插值，分别形成信号的上包络和下包络。

(2)求出上、下包络的均值序列，将原信号减去该均值序列，得到该信号的差值序列。

(3)计算上、下包络均值与差值序列的均方能量比，如果该比例低于 0.3，则该差值序列就是本次循环获得的 IMF，并且跳转至步骤(4)；如果该比例高于 0.3，用该差值序列替代原信号，并跳转至步骤(1)。

(4)将原信号减去 IMF，得到该级 IMF 对应的剩余项。

(5)对步骤(4)得到的剩余项序列进行局部极值点检测，若检测得到的极大值与极小值个数之和小于 2，则 EMD 结束；否则，用剩余项替代原信号，进行下一级的 IMF 提取，跳转至步骤(1)。

信号 $x(t)$最终被分解成若干有限个 $\mathrm{IMF}_i(t)$ (i=i~n)以及剩余项 $r(t)$，根据上述的 EMD 的步骤，可以推断 EMD 是一可逆过程，由此信号可以重构为

$$x(t)=\sum_{i=1}^{n}\mathrm{IMF}_i(t)+r(t) \tag{1}$$

根据 EMD 算法可知，$r(t)$是常数或是单调函数，可以省略。对于 Hilbert 谱的运算由于在耳语音端点检测中并未涉及，故不再赘述，可参见文献[11]。

2.3 EMD 拟合特征应用于耳语音端点检测算法

耳语音作为人类语言发音的特殊形式，信号及发音过程的非线性、非平稳特性更加突出，常用的正常语音分析方法往往不再有效。所以基于人的发音机理以及 EMD 分解得到的 IMF 的自适应调频调幅性质，利用 EMD 进行耳语音信号分析，更符合耳语音信号非线性、非平稳的特征；并且由于 EMD 过程是对整体耳语音信号进行分解，而不需要分帧处理，避免了语音短时平稳的假设，更符合耳语音处理的实际。

EMD 提取信号的过程，实际上是依次分离信号的局部最高频分量的过程，所以每个 IMF 的细节信息也将越来越简单。对于无声段或称平稳宽带噪声段，信号低频能量相对较大，随着频率升高能量逐步下降而后趋于平稳，总体频域能量是较为均匀的，所以 EMD 提取的 IMF 能量也较为均匀。对于语音段，由耳语音的发音机理，知其声韵母均为清音特性。虽无周期脉冲激励，但耳语音的元音仍然存在共振峰，只是峰值向高频偏移，一般都存在两个或两个以上的频域能量聚集区，从而导致其能量在频域中非均匀分布，因此 EMD 提取的各个 IMF 分量能量差异较大。所以利用 IMF 能量增长速度来区分元音与噪声是一种有效的方法。

通过对大量耳语音样本语谱图的观察统计，耳语音的辅音发音与正常语音区别不大[16]，不送气清塞音在语谱图上模式仍然为直冲条形态；送气清塞音模式为直冲条加乱纹；清擦音模式为由弱到强的无规则噪音乱纹；浊擦音模式为共振峰加乱纹；塞擦音模式为冲直条加噪音乱纹的混合形态，只是送气的塞擦音乱纹分布较宽。由此可见，耳语音的辅音保留有正常辅音的基本特征，故而在频谱上可以顺利区分出辅音段和噪音段[10]，导致 EMD 分解提取得到的各阶 IMF 局

部能量非均匀增长。这也解释了 EMD 拟和特征同样适用于辅音段检测的原因。

由于耳语音皆为清音，从信号分析的角度看耳语音的辅音和元音反而更为相似，区别更小；信号在时域波形上没有声门脉冲的干扰，突变点大大缩减，反而更适合于 EMD 的自适应提取，这也为利用 IMF 的归一化能量拟合特征用于耳语端点检测提供了有力的物理依据。

由于在低频段耳语音信号与噪声信号的区别不如高频段那样明显，所以在进行拟合特征提取时，只选择包含高频分量较多的几个 IMF。本文提出算法的主要流程：

(1)使用 2.2 节中所述的 EMD 方法，将耳语音信号整体分解成若干个 IMF。

(2)对每个 IMF 进行短时分帧处理，根据耳语音的特点，其发音状态变化相对缓慢，所以帧长可选择较长(取 60ms)，而帧移较小(取 5ms)。

(3)求出每个 IMF 的分帧能量 EIMF(i,j)：

$$\mathrm{EIMF}(i,j)=\sum_{l=1}^{L}|\,\mathrm{IMF}(i,l)\,| \tag{2}$$

其中 L 为帧长，l 为帧内样点序号，i 为 IMF 对应的序号，j 为帧号。

(4)将包含高频信息的几个 EIMF 能量归一化，并且累加求和：

$$\mathrm{SEIMF}(k,j)=\frac{\sum_{i=1}^{k}\mathrm{EIMF}(i,j)}{\sum_{i=1}^{m}\mathrm{EIMF}(i,j)},\quad k=1\sim m \tag{3}$$

m 为选取 EIMF 的个数，j 为帧号。

(5)取 $X=[1/m,2/m,\cdots,1]$，利用最小二乘法对每一帧 j 进行拟合，求出 X 与 SEIMF(i,j)的拟合斜率 $D(j)$，$D(j)$即为该帧语音的 EMD 拟和特征：

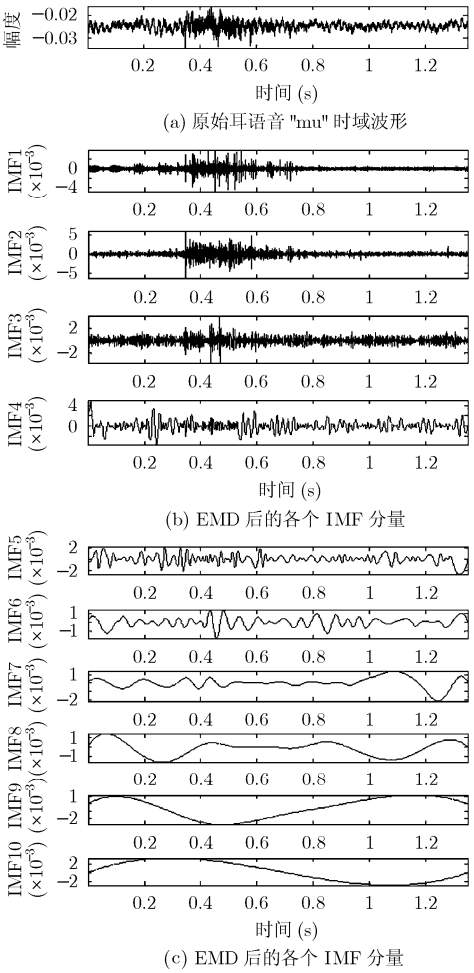
$$D(j)=\frac{\sum_{i=1}^{m}\mathrm{SEIMF}(i,j)\sum_{i=1}^{m}X(i)-m\sum_{i=1}^{m}\mathrm{SEIMF}(i,j)X(i)}{\left(\sum_{i=1}^{m}X(i)\right)^{2}-m\sum_{i=1}^{m}X(i)^{2}} \tag{4}$$

(6)由此根据计算出的拟合特征参数进行耳语音端点划分。

3 实验结果与分析

本文进行了大量实验来验证本文提出算法用于耳语音端点检测的有效性。在耳语音库中，根据汉语普通话基本音节表，选取 400 个音节，每个音节取 3 遍发音，总共 1200 个样本，耳语音的采样频率为 8kHz，信噪比约为 2~10dB。

图 1 示例了耳语音信号经 EMD 分解后得到的 IMF 分量：



(a) 原始耳语音 "mu" 时域波形

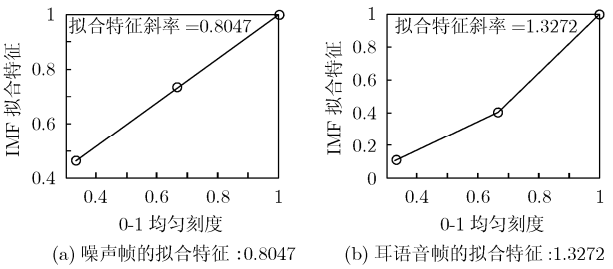
(b) EMD 后的各个 IMF 分量

(c) EMD 后的各个 IMF 分量

图 1

由图 1(b)，1(c)可知，分解后的语音能量主要集中在前几个 IMF 分量中，而且从能量大小角度观察，有声段和噪声段的分界较为明显；对于后几个 IMF 分量从波形上就很难分辨语音的端点了。单纯从能量大小的角度去划分耳语音的端点意义不大，也不利于研究耳语音声学特性。所以如前所述，将 IMF 能量进行归一化，这使得提取到的拟合特征参数只与其能量的分布有关。图 2 分别是一帧噪声数据和一帧语音数据提取得到的拟合特征：

从图 2 中可以得知，宽带噪声在频域中能量均匀分布，起始 IMF 能量相对大，而各个 IMF 增长较为缓慢，所以

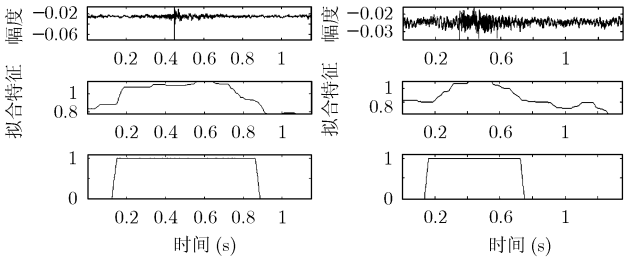


(a) 噪声帧的拟合特征 :0.8047 (b) 耳语音帧的拟合特征 :1.3272

图 2

获得的拟合斜率一般低于 1；而对于耳语音段，由于声门气流因声道的调制作用，以及涡流的存在和口腔的辐射影响等等，使得耳语音信号能量的分布不均，故而各个 IMF 增长较快，拟合斜率一般高于 1。

图 3 示例了本文算法进行耳语音端点检测的结果：



(a) (从上至下) 耳语音 "lu" 的时域波形 ;EMD 拟合特征 ; 端点检测结果

(b) (从上至下) 耳语音 "mu" 的时域波形 ;EMD 拟合特征 ; 端点检测结果

图 3

为进一步验证本文方法的有效性，对相同 1200 个样本采用了 3 种不同方法进行了实验比较；从结果中可知，本文方法均优于其它两种，如表 1 所示(零能积和熵函数方法参见文献[10])。

表 1　3 种端点检测方法实验结果(%)

方法	零能积方法	熵函数方法	本文方法
正确率	94.5	97.6	98.25

判断语音端点检测结果是否正确的标准一直都是值得探讨的问题。本文将检测到的语音起止端点在手工标定端点的前后 3 帧范围内，作为正确的检测结果。根据 2.3 节中所述，此标准估计耳语音端点检测误差最大不超出 15ms(帧移 5ms×3 帧)，根据文献[10]提供的汉语辅音平均音长数据，该误差确保辅音数据的完整性，且由于元音发音时长更长，更不会造成损失。

以上 3 种方法各有优劣，主要应用于不同的研究目的。零能积的方法较为传统也最为成熟，计算简单、耗时少，适用于实时系统。但是由于耳语音的能量较低，发音很不稳定，故而该方法并不适合耳语音的端点检测。频域熵函数的方法，以语音信号的频谱分布为依据，利用熵函数的性质将其量化，从而分离出有声段。该方法避免了耳语音能量较低的不利因素，突出了耳语音的频谱特征，是比较理想的方法。但是该方法是对语音信号的频域数据运算，增加了计算复杂度和运算时间。以上两种方法都是以纯粹信号处理的手段进行耳语音的端点检测，与汉语耳语音的发音特性以及人耳的感知特性并无联系。因此本文使用了当前研究的热点 EMD 方法，该方法虽然计算量大、耗时长，相对于其他方法不占优势，目前也很难实时运行，但 EMD 是对整体耳语音信号的分解，而且是根据耳语音信号自身的时域特征自适应的分解，这是其他分析方法诸如傅里叶分析和小波分析所无法完成的。目前 EMD 分解得到的 IMF 已证明具有调频调幅的特性，而调频调幅特征也是语音发音的特征，在这一点上算法与实际情况是统一的。而且人类对自身的听觉感知原理的认识还比较有限，故而将语音信号进行自适应分解后研究人类听觉感知原理也不失为一种值得探索的手段。这也为以后的汉语耳语音的声韵分离以及声调感知提供了研究基础。

4　结束语

基于汉语耳语音非线性、非平稳的特点，本文成功地运用了经验模态分解(EMD)方法，分离出信号的各个内禀模态函数(IMF)，并利用耳语音信号各个 IMF 能量增长非均匀的特点，引入了一种拟合特征参数作为信号的特征参量，并且将其应用在耳语音信号的端点检测中，有效克服了耳语音因信噪比低造成的端点检测易造成错误的困难，得到了很好的实验结果。

致谢　感谢南京大学声学所提供的耳语音数据测试样本。

参 考 文 献

[1] 拉宾纳, 谢弗著. 朱雪龙等译. 语音信号数字处理 [M]. 北京: 科学出版社, 1983: 100–105.
Written by Rabiner L R and Schafer R W. Translated by Zhu X L. Digital Processing of Speech Signals [M]. Beijing: Science Press, 1983: 100–105.

[2] 陈四根, 和应民. 一种基于信息熵的语音端点检测方法 [J]. 应用科技, 2001, 28(3): 13–14.
Chen S G and He Y M. A scheme of speech endpoint detection based on information entropy [J]. *Applied Science and Technology*, 2001, 28(3): 13–14.

[3] 胡光锐, 韦晓东. 基于倒谱特征的带噪语音端点检测 [J]. 电子学报, 2000, 28(10): 95–97.
Hu G R and Wei X D. Endpoint detection of noisy speech based on cepstrum [J]. *Acta Electronica Sinica*, 2000, 28(10): 95–97.

[4] Drouiche K, Gomez P, Alvarez A, Martinez R, Rodellar V, and Nieto V. A spectral distance measure for speech detection in noise and speech segmentation [C]. Proceedings of the 11th IEEE Signal Processing Workshop on Statistical Signal Processing, Singapore, 2001: 500–503.

[5] 韦岗, 陆以勤, 欧阳景正. 混沌,分形理论与语音信号处理 [J]. 电子学报, 1996, 24(1): 34–39.
Wei G, Lu Y Q, and Ouyang J Z. Chaos and fractal theories for speech signal processing [J]. *Acta Electronica Sinica*, 1996, 24(1): 34–39.

[6] Chen S H, Liao Y F, and Chiang S M, *et al.*. An RNN-based pre-classification method for fast continuous mandarin speech recognition [J]. *IEEE Trans. on Speech and Audio*

Processing, 1998, 6(1): 86–90.

[7] 朱杰, 韦晓东. 噪声环境中基于 HMM 模型的语音信号端点检测方法[J]. 上海交通大学学报, 1998, 32(10): 14–16.
Zhu J and Wei X D. Speech signal endpoint detection method based on HMM in noise [J]. *Journal of Shanghai Jiaotong University*, 1998, 32(10): 14–16.

[8] Robert W M and Mark A C. Reconstruction of speech from whispers [J]. *Medical Engineering & Physics*, 2002, 24(8): 515–520.

[9] Higashikawa M. Perceived pitch of whispered vowels – relationship with formant frequencies: a preliminary study [J], *Journal of Voice*, 1996, 10(2): 155–158.

[10] 栗学丽, 丁慧, 徐柏龄. 基于熵函数的耳语音声韵分割法 [J]. 声学学报, 2005, 30(1): 69–75.
Li X L, Ding H, and Xu B L. Entropy-based initial/final segmentation for Chinese whispered speech [J]. *Acta Acoustica*, 2005, 30(1): 69–75.

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

[12] Liu Z F, Liao Z P, and Sang E F. Speech enhancement based on Hilbert-Huang transform [C]. Proceedings of 2005 International Conference on Machine Learning and Cybernetics, Guangzhou, China, 2005, 8: 4908–4912.

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

[14] 杨莉莉, 李燕, 徐柏龄. 汉语耳语音库的建立与听觉实验研究[J]. 南京大学学报(自然科学), 2005, 41(3): 311–317.
Yang L L, Li Y, and Xu B L. The establishment of a Chinese whisper database and perceptual experiment [J]. *Journal of Nanjing University* (*Natural Sciences*), 2005, 41(3): 311–317.

[15] Taisuke I, Kazuya T, and Fumitada I. Analysis and recognition of whispered speech [J]. *Speech Communication*, 2005, 45(2): 139–152.

[16] 吴宗济,林茂灿主编. 实验语音学概要[M]. 北京: 高等教育出版社, 1989: 112–152.
Wu Z J and Lin M C. Experiment Phonetics [M]. Beijing: Higher Education Press, 1989: 112–152.

潘欣裕： 男，1981 年生，硕士生，研究方向为语音信号处理.

赵鹤鸣： 男，1957 年生，苏州大学电信学院院长，教授，博士生导师，研究领域为语音信号处理、多媒体处理、神经计算.

陈雪勤： 女，1974 年生，讲师，博士生，研究方向为语音信号处理.

徐 敏： 女，1982 年生，硕士生，研究方向为语音信号处理.