Hilbert-Huang Transform Based Speech Enhancement

Liran Shen, Xueyao Li, Huiqiang Wang, Qingbo Yin, Rubo Zhang

College of Computer Science & Technology, Harbin Engineering University, Harbin, China, 150001
(yinq2003@eyou.com, yinq2004@21cn.com)

Abstract

Based on Hilbert-Huang Transform a novel speech signal enhancement was proposed. A new method called Hilbert-Huang Transform (HHT) was described. This technique allows the decomposition of one-dimensional signal into intrinsic oscillatory modes. The HHT allows us to study the different intrinsic oscillatory mode and instantaneous frequencies of the speech signal and noise. Applied to speech signal, it proves new time-frequency attributes. The results of experiments show that the new technique performed well, and its application foreground is hopeful.

Keywords: Hilbert-Huang Transform, speech enhancement, nonlinear and non-stationary signal processing

1. Introduction

The fast growing mobile communication of today demands increasingly better sound quality of the received speech signal. Disturbances that make the speech less intelligible often come from the background environment such as a car engine or humming people. There exist many methods for reducing such noise, e.g. spectral subtraction [1], Kalman filtering [2] and wavelet methods [3,4]. The noise reduction methods above have one thing in common. They often consider the signal in short time is stationary. Unfortunately, the signals are often unlinear and unstationary.

The HHT method is specially developed for analyzing nonlinear and non-stationary data. The method consists of two parts: (1) the empirical mode decomposition (EMD), and (2) the Hilbert spectral analysis. The key part of the method is the first step, the EMD, with which any complicated data set can be decomposed into a finite and often small number of intrinsic mode functions (IMF). An IMF is defined here as any function having the same number of zero-crossing and extrema, and also having symmetric envelopes defined by the local maxima, and minima respectively. The IMF also thus admits well-behaved Hilbert transforms. This decomposition method is adaptive and therefore highly efficient. Since the decomposition is based on the local characteristic time scale of the data, it is applicable to non-linear and non-stationary processes. With the Hilbert transform, the IMF yield instantaneous frequencies as functions of time that give sharp identifications of imbedded structures. The final presentation of the results is an energy–frequency–time distribution, which we designate as the Hilbert Spectrum.

In this paper, we use a novel data analysis method, which is not limited to linear and statistically stationary time series---noisy speech signal. This empirical mode decomposition (EMD) method extracts the energy associated with various intrinsic time scales in generating a collection of intrinsic mode function (IMF). The decomposition can also be viewed as an expansion of the data in terms of the IMFs. Then these IMFs, based on and derived from the data, serve as the basis of that expansion; they can be linear or nonlinear, as dictated by the data. The different IMFs correspond to the different physical time scales, which characterize the various dynamical oscillations in speech signal. The next section describes the Hilbert-Huang Transform (HHT). Section 3 using HHT to analyze speech signal, which constituted by a simple speech signal model. Section 4 describes the speech enhancement method.

2. Brief description of the HHT

Compare with various data analysis methods, the innovation of HHT is the introduction of IMF, which guarantees the physically meaningful instaneous frequency. HHT is composed of two procedures, i.e. (a) the EMD method, and (b) the associated Hilbert spectral analysis. The purpose of the EMD is to extract the IMFs from the data with sifting process.

2.1. Empirical Mode Decomposition

The empirical mode decomposition (EMD) was first introduced by Huang et al. [5]. The principle of this technique is to decompose adaptively a given signal $X(t)$ into oscillating components. These components are called intrinsic mode functions (IMFs) and are obtained from the signal X by means of an algorithm, called sifting. The essence of this algorithm

is to identify the IMF by characteristic time scales, which can be defined locally by the time lapse between two extrema of an oscillatory mode or by the time lapse between two zero crossings of such mode. The idea is then to extract for each mode locally the highest frequency oscillations out of X. To compute also the frequency behavior of each IMF in time, Huang proposed to use the instantaneous frequency of each IMF. However to calculate instantaneous frequencies we have to ensure that each IMF is symmetric with respect to its local mean, otherwise unwanted fluctuations in the instantaneous frequency will be induced by asymmetric waveforms in the IMF. In some sense the EMD can be seen as a type of adaptive wavelet decompositions, which was used for this problem. Each IMF replaces then the detail signals of X at a certain scale or frequency band. However, the EMD is adaptive since the frequency subbands in which the IMFs are built up as needed to separate the different oscillating components of X. Furthermore, the EMD does not use any pre-determined filter or wavelet functions. It is a fully data driven method.

The algorithm to create IMFs is elegant and simple. First, the local extremes in the time series data $X(t)$ are identified, and then all the local maxims are connected by a cubic spline line $U_X(t)$, known as the upper envelope of the data set. Then, we repeat the procedure for the local minima to produce the lower envelope $L_X(t)$. Their mean $m_1(t)$ is given by:

$$m_1(t) = \frac{L_X(t) + U_X(t)}{2} \tag{1}$$

It is a running mean. We note that both envelopes should cover by constructing all the data between them. Then we subtract the running mean $m_1(t)$, from the original data $X(t)$, and we get the first component $h_1(t)$:

$$h_1(t) = X(t) - m_1(t) \tag{2}$$

To check if $h_1(t)$ is an IMF, we demand the following conditions: (i) $h_1(t)$ should be free of riding waves i.e. the component should not display under-shots or over-shots riding on the data and producing local extremes without zero crossing. (ii) To display symmetry of the upper and lower envelopes with respect to zero. (iii) Obviously the number of zero crossing and extremes should be the same in both functions.

The sifting process has to be repeated as many times as it is required to reduce the extracted signal to an IMF. In the subsequent sifting process steps, $h_1(t)$ is treated as the data; then:

$$h_{11}(t) = h_1(t) - m_{11}(t) \tag{3}$$

If the function $h_{11}(t)$ does not satisfy criteria (i) – (iii), then the sifting process continues up to k times, h_{1k}, until some acceptable tolerance is reached:

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \tag{4}$$

The resulting time series is the first IMF, and then it is designated as $C_1(t) = h_{1k}$. The first IMF component from the data contains the highest oscillation frequencies found in the original data $X(t)$.

The first IMF is subtracted from the original data, and this difference, is called a residue $r_1(t)$ by:

$$r_1(t) = X(t) - C_1(t) \tag{5}$$

The residue $r_1(t)$ is taken as if it was the original data and we apply to it again the sifting process. The process of finding more intrinsic modes $C_j(t)$ continues until the last mode is found. The final residue will be a constant or a monotonic function; in this last case it will be the general trend of the data.

$$X(t) = \sum_{j=1}^{n} C_j(t) + r_n(t) \tag{6}$$

Thus, one achieves a decomposition of the data into n-empirical IMF modes, plus a residue, $r_n(t)$, which can be either the mean trend or a constant. We must point out that this method does not require a mean or zero reference, and only needs the locations of the local extremes.

2.2 The Hilbert Transform

Having obtained the IMF, one applies the Hilbert transform to each IMF component.
The Hilbert transform $Y_j(t)$ of $C_j(t)$ is

$$Y_j(t) = \frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{C_j(\tau)}{t - \tau} d\tau \tag{7}$$

Where P indicates the Cauchy principal value. With this definition $Y_j(t)$ and $C_j(t)$ form the complex conjugate pair, so we can get an analytic signal, $Z_j(t)$

$$Z_j(t) = C_j(t) + iY_j(t) = a_j(t)e^{i\theta_j(t)} \tag{8}$$

Where
$$a_j(t) = \left[C_j^2(t) + Y_j^2(t)\right]^{1/2} \tag{9}$$

And
$$\theta j(t) = \arctan\left[\frac{Y_j(t)}{C_j(t)}\right] \tag{10}$$

The corresponding frequency is:

$$\omega_j(t) = \frac{d\theta_j(t)}{dt} \tag{11}$$

Compared with the classic FFT, where amplitude and frequency are time independent, the amplitude and frequency derived by HHT are functions of time. In amplitude-frequency-time plot, the amplitude can be contoured on the frequency –time plane. This frequency-time distribution of the amplitude is designated as the Hilbert amplitude spectrum, $H(t,\omega)$, or simply Hilbert spectrum. The square of the amplitude is a common representation of the energy density.

3. Application of HHT to speech signal analysis

The so-called zero-phase harmonic representation [6] is among the simplest sinusoidal models for speech analysis and synthesis. Its elegance is in the use of the same expression for both voiced and unvoiced segments and allowing for a soft decision whereby a frame may contain both types. The model is characterized by sine-wave amplitudes, a voicing probability, and a fundamental frequency. In this representation, a given frame is represented by a sum of harmonically related sine waves. A synthetic phase function is used such that during voiced speech, the sine waves are coherent (in phase) and during unvoiced speech they are incoherent. The speech signal over a short-term window is then

$$s(n) = \sum_{m=1}^{M} a_m \cdot \cos[(n-n_0)w_m + \psi_m + \theta_m] \tag{12}$$

Where n_0 is the voice onset time, M the number of sinusoids, a_m the amplitude and w_m the excitation frequency of the mth sine wave. For a stationary periodic frame, these are harmonically related, i.e., $w_m = mw_0$ with w_0 being the fundamental frequency. The first phase term in Eq. (12) is due to the onset time n_0 defined as the time when the pitch pulse occurred relative to the beginning of the frame. The second phase component depends on a frequency cutoff w_c and a voicing probability p_v, so that the higher the voicing probability the more sine waves are declared voiced with zero phase. The third phase component is the system phase θ_m along frequency track m, often assumed zero or a linear function of frequency.

If speech were divided in narrow bands such that at most two harmonics fall in each band, then in light of the model, voiced speech is expressed as the sum of two sinusoids with deterministic phase and unvoiced speech as the sum of two sinusoids with random (and uncorrelated) phases. Thus,

Voice speech:

$$s(n) = a_1 \cos(npw_1 + \varphi_1) + a_2 \cos(npw_2 + \varphi_2) \tag{13}$$

With $\varphi_1 = cw_1$, $\varphi_2 = cw_2$ both deterministic

Unvoice speech:

$$s(n) = a_1 \cos(npw_1 + \varphi_1) + a_2 \cos(npw_2 + \varphi_2) \tag{14}$$

With $\varphi_1, \varphi_2 \in [-\pi, \pi]$ uniformly distributed,

Where p is the sampling period. To account for transitional segments, transient speech is modeled here as an exponentially decaying (or growing) sinusoid. Transient voiced speech:

$$s(n) = ae^{-\alpha nP} \cos(npw_0 + \phi) \tag{15}$$

With φ is a constant, $\phi = cw_0$.

It is worth noting that the term 'voiced speech' used here refers to a simple two-harmonic signal in a narrow band. This is unlike the conventional usage of this term, which usually refers to a speech signal with a fundamental frequency and many harmonics. Similarly, the model assumed for transient speech is a simplistic model in that it only uses a single sinusoid and assumes an exponential decay. These restrictions are made to simplify the mathematics while providing a third simple signal model, which is different from the steady two-sinusoid models for voiced and unvoiced speech.

In this paper, we construct a speech signal based on the sinusoidal model which the sample rate is $8000Hz$, and $\omega_1 = 100Hz$, $\omega_2 = 200Hz$, $a_1 = a_2 = 1$, $\phi_1 = \phi_2 = 0$. The Fig.1 show the original single — o_signal and the result of EMD and the reconstruct signal—R_signal.

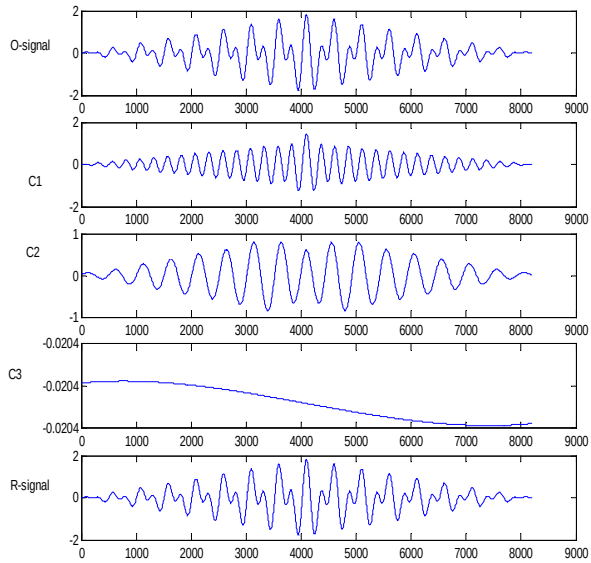

Fig.1 the decomposing and reconstructing using EMD

From the Fig.1 we can see the original signal was

decomposed into the two IMFs of O-signal C1 and C2. C1 and C2 are the two different components of O_signal which frequency are $\omega 1$ and $\omega 2$ respectively. C3 is residual component. We can reconstruct the original signal using C1and C2. The reconstruct method is

$$R\_signal = \sum_{i=1}^{2} C_i \tag{16}$$

The reconstruct error is

$$ERR = \frac{(\sum_{j=1}^{N} [R\_signal(j) - o\_signal(j)]^2)}{N} \tag{17}$$

So, we can get the reconstruct error, which is 3.048e-11.

4. Speech enhancement by using threshold in IMFs domain

As speech has evolved from laboratory demonstrations to real-world application, the need to maintain performance in a wide variety situation has emerged. Speech enhancement provides one way of compensating for different environments. In this paper, we study enhancement of speech corrupted by additive noise.

Assume we have the signal model

$$y(n) = x(n) + e(n) \tag{18}$$

In time domain where $y(n)$ is the noisy speech signal, $x(n)$ is the noise free speech signal and $e(n)$ is the added background noise.

4.1 The Enhancement Method

In this section, we propose a novel strategy for speech enhancement based on HHT. The novel method is as follows:

(1) According to section 2, calculate the IMFs and Hilbert-Huang spectrum.

(2) Calculate the marginal spectra of the IMFs, i.e. we have computed for each *j*th IMF, its Hilbert amplitude spectrum $H(\omega, t)$ or Hilbert spectrum. The marginal Hilbert spectrum for the *j*th IMF is given by:

$$h_j(\omega) = \int_0^T H_j(\omega, t)dt \tag{19}$$

(3) Find the maximum of *j*th IMF marginal Hilbert spectrum.

(4) Find the frequency which corresponding the maximum.

(5) Compute the gain value of the *j*th IMF:

$$G(j) = \frac{\max_t t(h_j(t))}{\arg(\max(h_j(t)))} \tag{20}$$

(6) Compute all the gain of IMFs based on (2)-(5). And obtain gain curve $G(i), i = 1 \cdots N$, N is the total of IMFs.

(7) Reconstruct signal:

$$r\_signal = \sum_{i=1}^{N} G(i) \cdot IMF_i(t) \tag{21}$$

4.2 Experiments and Analysis

In this section, we take two examples. The first one is the o_signal added white noise, namely n_signal. And the SNR is −5dB.

Fig.2 display the o_signal, n_signal and r_signal, which obtain by algorithm 4.1. Fig.3 displays the decomposition in eleven IMFs of the n_signal. With the help of sifting algorithm explained in section 2.

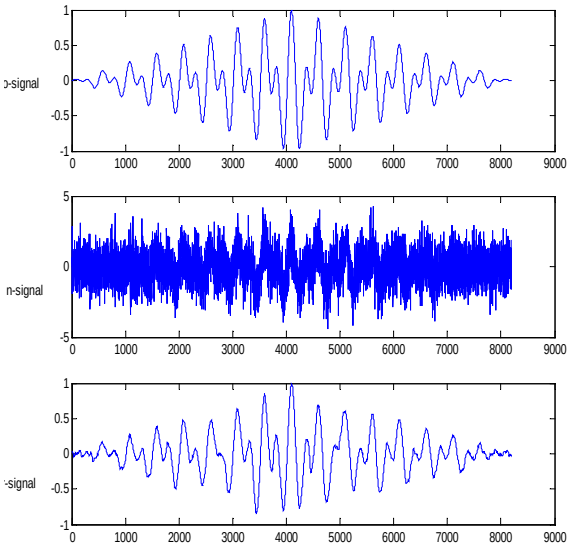

Fig.2 o_signal, n_signal and r_signal

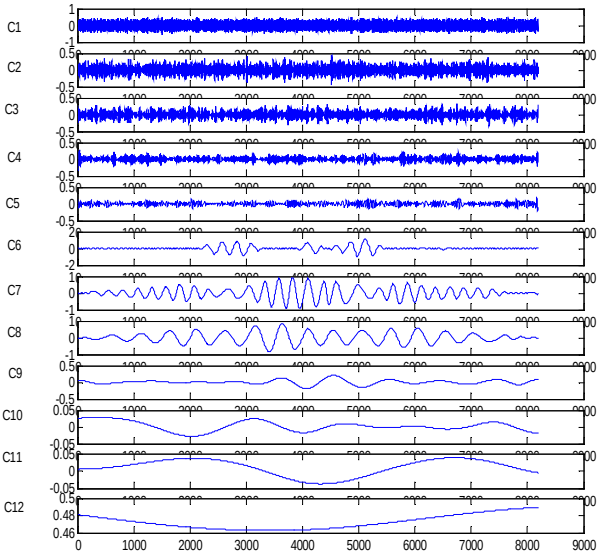

Fig.3 Decomposition by the sifting method of n_signal.

From Fig.3 we can see the intrinsic mode function $C_j(t)$, or IMFs represent the oscillation modes embedded in the signal. Therefore the decomposition of the n_signal in IMFs, will have two advantages, the first one is that will allow us to examine the physical meaning of each IMF component. The second one is that the residual will contain the data trend, and in this way we can perform the analysis even if the data are not stationary. Through analyze the IMFs and it marginal Hilbert spectrum using algorithm 4.1, we can get gain curve showed in Fig.4.

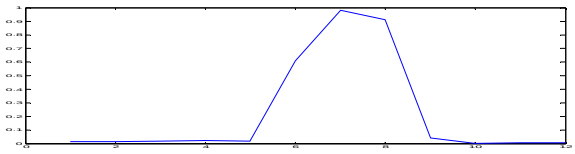

Fig.4. The gain curve of n_signal.

Based on the gain curve, we can reconstruct the signal r_signal.

The second one is the real-world signal, which come from short-wave channel. Fig.5 displays the original signal o__signal and the reconstructed signal r__signal. The 0__signal is the voice ‘ ye’ said by one man and contaminated by noise.

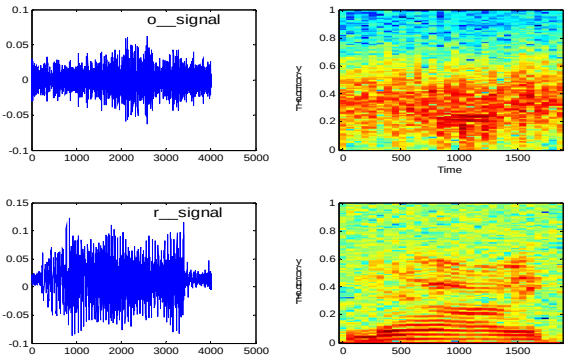

Fig.5. o__signal, r__signal and their spectrum

Using the algorithm 4.1 we can obtain the gain curve showed in Fig.6.

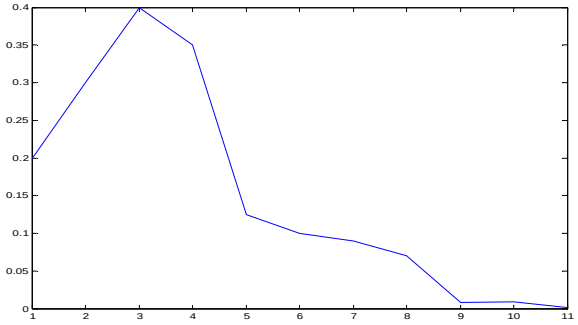

Fig.6 the gain curve of o__signal

From Fig.5 we can see the r__signal’s spectrum is clearer than the o__signal. Actually, the signal of r__signal is heard clearly.

5.Conlusion

In this paper, we describe a new technique HHT that is used to process non-linear and non-stationary signal. For speech signal contaminated by noise we proposed a novel method to enhance it. From the algorithm and the two examples, we can see that this method can adaptively extract IMFs form itself, and then obtain the gain curve. All the processing of the method based on the characteristics of signals itself. However other processing such as fourier or wavelet need to designate basic function. The results of experiments show that the new technique performed well and its application foreground is hopeful.

Reference:

[1] Boll, S. 1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. On Acoustics, Speech, and Signal Processing*, 1979,**ASSP- 27**(2), 113–120.

[2] Sörqvist, P., Händel, P., & Ottersten, B. Kalman filtering for low distortion speech enhancement in mobile communication. In *Proc. ICASSP ’97*, 1219–1222, Munich, Germany.

[3] Storm, H. Noise reduction of speech signals with wavelets. Tech. rep. NO 1998-02/- ISSN 0347-2809, Department of Mathematics, Chalmers University of Technology and Gothenburg University, Gothenburg, Sweden.

[4] Holst, J., Lindoff, B.,&Svensson, N. Wavelets and noise cancellation in mobile communication.In The International Conference on Signal Processing Applications and Technology, Toronto, Canada. 1998,1423-1428.

[5] Huang NE, Shen Z, Long SR, Wu MC etal. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc.R.Soc., 1998,A(454):903:995.

[6] McAulay, R., Quatieri, T. Speech analysis/synthesis based on a sinusoidal representation.1986, IEEE Trans. ASSP 34 (4), 744.