IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingenierie des Systemes d'Information
Vol. 24, No. 1, February, 2019, pp. 57-60
Journal homepage: http://iieta.org/Journals/isi

Adaptive Speech Enhancement Algorithm Based on Hilbert-Huang Transform

Na Jiang[1*], Jiyuan Li[2]

[1] School of Automatic and Electrical Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China
[2] Business Support Center, China Mobile Communications Group Gansu Co., Ltd., Lanzhou 730070, China

Corresponding Author Email: najiang2008@163.com

https://doi.org/10.18280/isi.240108

Received: 13 November 2018
Accepted: 9 February 2019

Keywords:
HILBERT-Huang transform, empirical mode decomposition, intrinsic mode function, speech enhancement

ABSTRACT

In order to enhance the speech signal, a speech enhancement algorithm based on HHT transform is proposed. Firstly, this paper uses HHT to perform the empirical mode decomposition (EMD) of the noisy speech signal to obtain the intrinsic mode function (IMF) components of each order, then removes the high-frequency and low-frequency components and finally constructs the speech source signals with the residual components by the Hilbert transform. The experimental results show, due to the better adaptive ability of the speech enhancement algorithm based on HHT transform, so the algorithm greatly reduces the distortion of the speech signal. At the same time, it has better denoising effect in speech enhancement than wavelet transform method and the Xia's method, showing the advantage of HHT in processing non-stationary nonlinear signals. The effect of the obtained results is further recognized in the research field of non-stationary signals. The results of this study can be used as a reference algorithm for speech signal enhancement methods.

1. INTRODUCTION

In the natural environment of speech communication, speech signals can be easily interfered by environmental noise. Such noise can affect the quality of speech signals, and in severe cases, completely drown out the speech signals, making it impossible for the human ear to hear. Speech enhancement refers to the process of extracting clean original speech from the noisy speech signals. Its main purpose is to objectively eliminate noise and subjectively enhance the intelligibility of speech.

In recent years, Chinese American scientist Huang et al. proposed a time-frequency decomposition algorithm for analyzing non-stationary nonlinear signals using Hilbert-Huang transform (HHT) [1]. This algorithm consists of two parts - empirical mode decomposition (EMD) and Hilbert transform, of which, the former is the core. EMD is an adaptive signal decomposition method. Based on the local properties of data in time domain, it can decompose complex data into a set of intrinsic mode function (IMF) components, and differentiates the phase to obtain the instantaneous frequency through Hilbert transform, so that the concept of instantaneous frequency has an actual physical meaning. The biggest advantage of EMD over wavelet transform is that the latter is essentially a linear transform, which has shortcomings in dealing with nonlinear problems, and wavelet decomposition requires pre-setting of the wavelet basis (such as Morlet, Sym and Daubechies, etc.) and the decomposition hierarchy. EMD, on the other hand, overcomes the shortcomings of wavelet transform to a certain extent. It does not need to give the basis function and decomposition hierarchy in advance, but adaptively obtains the result based on signal characteristics by the iterative method. Since Hilbert-Huang transform is based on the local features of the signals in time domain, its decomposition process is adaptive, making it suitable for the analysis of non-stationary nonlinear time-varying processes. So far it has been applied in many fields [3-4]. At present, common speech enhancement methods include the noise cancellation method [5], the spectral subtraction method [6, 7], the priori information method based on the speech parameter model [8], the signal subspace method [9], the HMM method based on state transition [10] and the wavelet denoising method [11, 12], etc. Xia et al. [13] proposed a speech enhancement method based on Hilbert-Huang transform. The empirical mode decomposition method is applied to de-noise initially, then the spectral subtraction method is used to enhance each intrinsic mode function. Since speech signals are essentially non-stationary nonlinear signals, these methods all have their limitations, although they can address the denoising of noisy speech to a certain extent.

This paper studies the application of HHT in speech enhancement. First, the author added Gaussian white noise into a section of clean speech signals, then used EMD to decompose the noisy speech signals to obtain each IMF component, and finally removed the first few orders of IMF high-frequency components (such as IMF_1 and IMF_2) and margin (trend term). After that, the author performed Hilbert transform of the remaining IMF components, and reconstructed the enhanced speech signals after further removing the noise. The simulation results indicate that the HHT-based adaptive speech enhancement algorithm is better than the wavelet-transform-based speech enhancement method, showing the advantage of Hilbert-Huang transform in dealing with non-stationary nonlinear signals.

This paper will be introduced in four parts. These four parts introduce the background and significance of speech enhancement, the basic theory of HHT transform, the algorithm flow and conclusions of this paper and the summary of the algorithm. From the actual meaning and the detailed implementation process of the algorithm, we have found that

the algorithm has better speech enhancement effect.

2. HILBERT-HUANG TRANSFORM

2.1 Huang transform

The principle of the Huang transform, i.e. the EMD, is to decompose the complex signal into the sum of several intrinsic mode functions (IMFs) with physical meanings, that is, to decompose the signal into a group of IMF signals, each of which is a narrow-band signal. The IMF component *f(t)* must meet two basic conditions, as follows:

(1) In the entire data sequence range, the number of zero crossing points N_s and the number of extreme points N_e are equal or have a difference of 1 at most, as follows:

$$(N_s - 1) \le N_e \le (N_s + 1) \tag{1}$$

(2) At any point of time t_i, the mean of the upper envelope $f_{max}(t)$ obtained from the local maximum value of the signal and the lower envelope $f_{min}(t)$ from the local minimum value of the signal is zero, as follows:

$$[f_{\min}(t_i) + f_{\max}(t_i)]/2 = 0, \quad t_i \in [t_a, t_b] \tag{2}$$

where, $[t_a, t_b]$ is a time interval.

The first constraint is already evident, similar to the distribution of the traditional stationary Gaussian process; the innovative part is the second condition, which changes global limit to local limit. Such limitation is necessary to prevent the fluctuations in the instantaneous frequency caused by the asymmetry of the signal waveform.

The EMD process of any original signal *x(t)* is as follows:

(1) Calculate all local extremum points of the original signal *x(t)*, and then use the cubic spline curve to connect all the minima and maxima points to obtain the envelopes of *x(t)*, so that all data points of the signal are between in the upper and lower envelopes.

(2) Calculate the mean of the upper and lower envelopes to obtain the sequence *m(t)*, and subtract *m(t)* from the original signal *x(t)* to get the following equation:

$$h_1(t) = x(t) - m(t) \tag{3}$$

Then check if $h_1(t)$ meets both conditions for the intrinsic mode function component at the same time. If it does, take $h_1(t)$ as a processed signal, and repeat the above steps until $h_1(t)$ becomes the IMF component, which is denoted as:

$$c_1(t) = h_1(t) \tag{4}$$

(3) $c_1(t)$ is the first IMF component decomposed from *x(t)*. Subtract $c_1(t)$ from *x(t)* to obtain a sequence of residual values $r_1(t)$, as follows:

$$r_1(t) = x(t) - c_1(t) \tag{5}$$

(4) Take $r_1(t)$ as the new original signal and repeat the above steps to obtain the 1st, 2nd, ..., nth IMF components, which are sequentially denoted as $c_1(t)$, $c_2(t)$, ..., $c_n(t)$, until the preset stop criterion is satisfied (if the IMF component $c_n(t)$ or the residual $r_n(t)$ is sufficiently small or the residual $r_n(t)$. The result is the remainder $r_n(t)$ of the original signal.

In this way, the original signal *x(t)* is decomposed into several IMF components plus one remainder:

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{6}$$

The original signal *x(t)* is decomposed into *n* IMF components and one residual $r_n(t)$. The decomposition process is based on the local features of the data signal, so it is empirically and adaptively decomposed into stationary IMF components. The result of EMD decomposition reflects the real physical process, indicating this method can better deal with non-stationary nonlinear time-varying signals.

2.2 Hilbert transform

When we discuss the EMD method, we also need to understand what is the instantaneous frequency. As a matter of fact, there was no common understanding about this concept until the Hilbert transform method emerged.

For any time sequence *x(t)*, you can obtain its Hilbert transform:

$$y(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau \tag{7}$$

Construct the analytic signal:

$$z(t) = x(t) + iy(t) = a(t)e^{i\Phi(t)} \tag{8}$$

where,

$$a(t) = \sqrt{x^2(t) + y^2(t)} \tag{9}$$

$$\Phi(t) = \arctan \frac{y(t)}{x(t)} \tag{10}$$

Equation (8) represents the instantaneous amplitude and instantaneous phase, which are the instantaneous characteristics of the signal. Therefore, the definition of instantaneous frequency is as follows:

$$\omega(t) = \frac{d\Phi(t)}{dt} \tag{11}$$

Perform Hilbert transform on equation (6) as follows:

$$H(t) = \sum_{j=1}^{n} a_j(t) e^{i\int \omega_j(t) dt} \tag{12}$$

where, $a_j(t)$ is the amplitude of the analytic signal of the jth-order IMF component $c_j(t)$. Here the *n*th-order residual $r_n(t)$ is omitted because $r_n(t)$ is a constant or monotonic function.

H(t) in equation (12) is a function of time t and also a function of frequency ω. Take the real part of equation (12), and we can obtain the expression of the signal *x(t)*:

$$x(t) = \mathrm{Re}\sum_{j=1}^{n} a_j(t) e^{i\int \omega_j(t)dt} \tag{13}$$

The above formula is called Hilbert-Huang Transform (HHT), which is an EMD-based Hilbert analysis method. HHT is the Hilbert-Huang transform expression of the original signal *x(t)*, which contains the instantaneous frequency and amplitude of *x(t)*. It is a fully adaptive time-frequency analysis method that can analyze not only stationary and linear signals but also non-stationary and nonlinear ones.

3. ALGORITHM SIMULATION

In order to verify the effectiveness of the proposed algorithm, the selection of speech and noise materials is the basis for the entire experimental process. In this experiment, the clean speech signal "Zhao Ci Bai Di Cai Yun Jian" (in the morning one departs from Baidi, amongst rosy clouds) was downloaded from the "Download Centre" on the website of Beihang University Press as the simulation object. It was sampled at a frequency of 8KHz, with a duration of 3.5 s. The noise material was Gaussian white noise taken from the NOISEX-92 database. The simulation experiment was carried out in the MatlabR2010a software environment.

(1) Gaussian white noise was added to the clean speech signal at a signal-to-noise ratio (SNR) of 5 dB to obtain a noisy speech signal. Figure 1 shows the waveforms of the clean speech signal and the noisy speech signal.

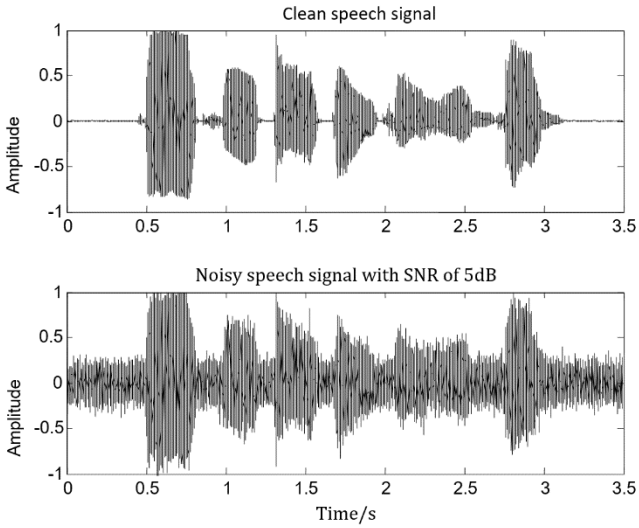

Figure 1. Waveforms of clean speech signal and noisy speech signal

Firstly, HHT was used to decompose the noisy speech signal with a signal-to-noise ratio of 5 dB. Through this, the IMF components of each order and margin are obtained, as shown in Figure 2.

The frequency range of the audio signals that can be heard by the human ear is 20 Hz ~ 20 kHz, within which, the frequency range of speech signals is 300 Hz ~ 3400 Hz. According to this feature and the IMF components obtained through EMD, the high frequency component imf_1, the low-frequency components imf_{11}~imf_{14} and the residual r_{14} could be removed in the experiment, and the remaining IMF components were further denoised through Hilbert transform to reconstruct the original speech signal. The simulation results also show that the speech signal re-constructed with the remaining components after removal of the above components has the highest signal-to-noise ratio - up to 7.5293 dB, an increase of 2.5293 dB. The reconstructed speech signal waveform is shown in Figure 3.

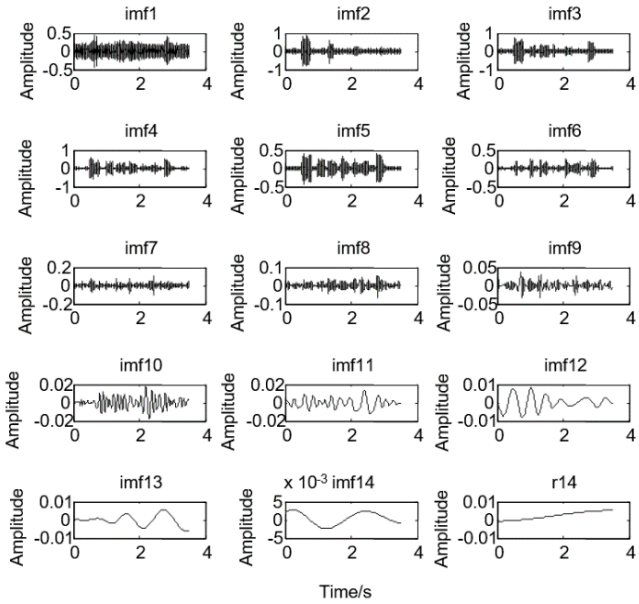

Figure 2. IMF components and margin of the noisy speech signal after EMD

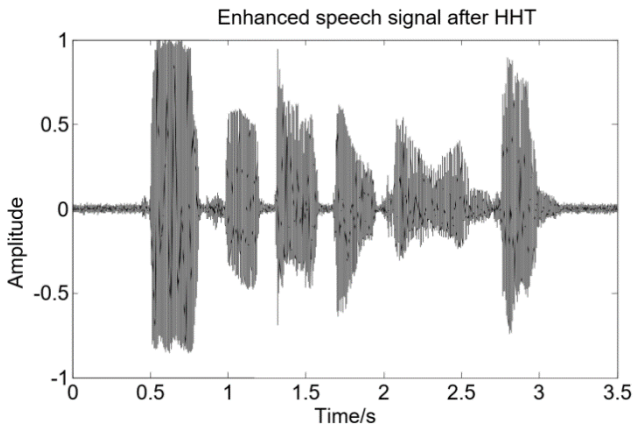

Figure 3. Enhanced speech signal waveform after de-noising and reconstruction

It can be seen from the simulation results in Figure 1 and Figure 3 that, after most of the background noise is eliminated by the proposed algorithm, the waveform of the reconstructed enhanced speech signal is already very close to that of the original clean speech signal. The audiometry of the experiment results shows that the enhanced speech signal obtained by the proposed algorithm has higher intelligibility and naturalness, which proves the effectiveness of the proposed algorithm.

(2) The experiment compares the speech enhancement performance of the wavelet transform (Daubechies 3 wavelet base, decomposed into 5 layers) and Xia's method, compares the experimental results with different signal-to-noise ratios, as shown in Table 1.

Table 1. Output signal-to-noise ratios at different input signal-to-noise ratios/dB

Input SNR/dB	Speech enhancement by Wavelet transform	Speech enhancement by Xia's method	Speech enhancement by the proposed algorithm
input SNR=-5	-0.3027	-0.1937	-0.0415
input SNR=-3	0.4272	0.8428	1.8821
input SNR=-1	0.9125	2.1086	3.1126
input SNR=0	2.0125	2.9479	3.7542
input SNR=1	3.8023	4.1038	4.8977
input SNR=3	4.9642	5.2980	6.2788
input SNR=5	6.2935	7.1458	7.5293

It can be seen from Table 1 that when the input SNR is increasing, the output signal-to-noise ratios of these algorithms are also increasing. However, when the input SNR is the same data, the output SNR of the speech signal by the algorithm is higher than the other two methods. Therefore, the speech enhancement algorithm proposed in this paper is better than the other methods.

4. CONCLUSION

In recent years, the Hilbert-Huang transform has been used in the nonlinear non-stationary signal processing widely. And has been applied in speech signal processing gradually. Due to its own shortcomings in EMD theory, HHT still has some shortcomings in signal processing. For example, EMD produces some false frequency components in the low frequency band; so far, the EMD decomposition theory cannot be well explained, so the components can only be decomposed sequentially from high to low frequency, and there is no way to factor out one or some of them directly; and processing cannot be carried out in real time, slowing down the speed of the algorithm. In addition, as the algorithm proposed in this paper cannot completely remove the background noise, it needs to be combined with other methods in order to further remove the noise. Despite these problems, Hilbert-Huang transform, as a new signal analysis method, has a broad application prospect. The simulation results show that the proposed algorithm can remove most of the noise in the speech signal. Compared with the wavelet transform speech enhancement algorithm and the Xia's method, the proposed algorithm increased the signal-to-noise ratio of speech signals and improved the quality of the speech signal.

ACKNOWLEDGMENT

This work is supported by Natural Science Foundation of Gansu Province (Grant No.: 17JR5RA101), and Gansu "13th Five-Year" Planned Education Science Research Topic (Grant No.: GS[2016]GHB0217).

REFERENCES

[1] Huang NE, Shen Z, Long SR, Wu MC, Shih HH. (1998). The empirical mode decomposition and the Hilbert spectrum for nonlinear and nonstationary time series analysis. Proceedings of the Royal Society of London 454(1971): 903-995. https://doi.org/10.1098/rspa.1998.0193

[2] Xu XG, Xu GL, Wang XT. (2009). Empirical mode decomposition and its application. Acta electronica sinica 37(3): 581-585.

[3] Hao H, Wang HL, Wei Q. (2016). Theory of empirical mode decomposition and its application. Chinese High Technology Letters 26(1): 67-80.

[4] Huang CJ. (2013). The study of interferogram denoising method based on empirical mode decomposition. International Journal of Computer Science Issues 10(1): 750-756.

[5] Srinivasan A. (2009). Adaptive echo noise elimination for speech enhancement of Tamil letter. International Journal of Engineering Science and Technology 1(3): 91-97.

[6] Li Z, Wu WJ, Zhang Q. (2017). Multi-band spectral subtraction of speech enhancement based on maximum posteriori phase estimation. Journal of Electronics & Information Technology 39(9): 2283-2286.

[7] Suma MO, Madhusudhana RD, Rashmi HN, Manjunath BS. (2013). Speech enhancement using spectral subtraction. International Journal of Advanced Electrical and Electronics Engineering 2(4): 1-6.

[8] McCallum M, Guillemin B. (2013). Stochastic-deterministic MMSE STFT speech enhancement with general a priori information. IEEE Transactions on Audio, Speech, and Language Processing 21(7): 1445-1457. https://doi.org/10.1109/TASL.2013.2253100

[9] Wang H, Wang AP, Zou H. (2017). Speech enhancement algorithm based on signal subspace with low signal-to-noise ratio. Microelectronics & Computer 34(2): 43-47.

[10] Cassia VB, Junichi Y, Simon K. (2014). Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion. Computer Speech and Language 28(2): 665-686. https://doi.org/10.1016/j.csl.2013.06.001

[11] Tang P, Guo BP. (2017). Wavelet denoising based on modified threshold function optimization method. Journal of Signal Processing 33(1): 102-110.

[12] Zhang SQ, Li JZ, Guo R, Tang H, Zhao T. (2014). Study on white noise suppression using complex wavelet threshold algorithm. Applied Mechanics and Materials 521(3): 347-351. http://dx.doi.org/10.4028/www.scientific.net/AMM.521.347

[13] Xia ML, Xu Z, Yu QY, Fan YL. (2010). Study on speech enhancement based on Hilbert-Huang transform. Jisuanji Gongcheng yu Yingyong (Computer Engineering and Applications) 46(17): 139-141.