Proceeding of the IEEE
International Conference on Information and Automation
Shenyang, China, June 2012

A Novel Instantaneous Frequency-based Voice Activity Detection for Strong Noisy Speech

Wei Shi
Advanced Digital Signal Processing Laboratory
Peking University Shenzhen Graduate School
shiwei@sz.pku.edu.cn

Yuexian Zou*
Advanced Digital Signal Processing Laboratory
Peking University Shenzhen Graduate School
zouyx@pkusz.edu.cn

Abstract - **The development of robust voice activity detection (VAD) for strong noisy speech is a challenging task. In this paper, we propose a novel voice activity detection method under Hilbert-Huang Transform (HHT) framework by using its good ability to automatically extract signal-frequency related intrinsic mode functions (IMF) by the empirical mode decomposition (EMD), which provides us a more flexible way to select different IMFs with strong speech components. With the Hilbert transform, the instantaneous frequency (IF) can be computed. Making use of the speech characteristics of each IMF, a weighted instantaneous frequency average (WIFA) measurement is proposed and the corresponding WIFA-VAD algorithm is developed, where the VAD threshold can be automatically estimated using the first noise frames. Experiments show that the proposed WIFA-VAD can achieve comparable results at high SNR. For low SNR (e.g., -5dB and below) conditions, it is able to achieve lower false alarm ration (FAR) and missing error ratio (MER) , compared with those of the conventional VAD algorithms.**

Index Terms - Voice activity detection, instantaneous frequency, strong noisy speech, weighted instantaneous frequency average.

I. INTRODUCTION

Voice activity detection (VAD), or speech endpoint detection, is a preprocessing part of many speech signal processing techniques, such as automatic speech recognition, speech enhancement, speech coding, echo cancellation, hands-free phone system, etc. The accuracy of VAD algorithm has a dominate affection on the performance of these techniques. The well-known VAD algorithms are mostly based on features like short-time energy and zero crossing rate [1], short-time autocorrelation, speech cesptrum [2], spectrum entropy [3], speech periodicity character [4], speech/noise statistic model [5]-[7] and so on, which were developed by taking the advantages of speech short-time stationarity, speech statistical characteristics and vocal modeling. These algorithms have achieved relatively good results under certain circumstances but still are confronted with one common problem: the performance will decline as the signal-to-noise ratio (SNR) of the noisy speech decreases. Sometimes the decline is so rapid that those algorithms are almost inapplicable. Thus, developing more robust VAD algorithms for strong noisy speech (e.g., -5dB and below) is still a valuable exploration.

In this paper, Hilbert-Huang transform (HHT) theory is explored to analyze noisy speech and one novel VAD algorithm is developed. In principle, HHT is an empirical, adaptive, data-driven signal analysis method, which was first proposed by N. E. Huang *et al*. in 1998 [8]. It is constituted by two parts: Empirical Mode Decomposition (EMD) and Hilbert Spectrum Analysis (HSA). EMD technique performs an adaptive decomposition of a signal into the elementary components that don't overlap in frequency domain. These components are named as Intrinsic Mode Function (IMF). The Hilbert transform can be applied on IMFs to provide the instantaneous amplitude and phase information of the original signal. Meanwhile, the instantaneous frequency components also can be computed. HHT has been viewed as a new, efficient data analysis method dealing with signals produced by nonlinear and non-stationary processes. Hence, we consider it as a good tool in analyzing the non-stationary noisy speech signals and provide more flexible way to develop an efficient VAD algorithm.

Literature study shows that there are some preliminary studies by exploring HHT to compute the speech endpoints for noisy speech. For each frame, Lu *et al*. took the energy of each IMFs to separate speech frames and noise frames, which gives a detection rate of 54.46% for noisy speech with SNR=-10dB [9]. It is noted that the IMF energy-based solution only exploits the amplitude of IMFs while ignoring the phase information and instantaneous frequency. After carefully evaluating the property of the EMD technique, it is noticeable that the EMD exhibits certain filter bank property. Moreover, the IF property of the speech frame and noise frame is shown with different characteristics, which motivated us to develop the novel VAD algorithm by jointly making use of these features in the instantaneous frequency domain under HHT framework. In the instantaneous frequency domain, a weighted instantaneous frequency average (WIFA), which has the ability to reflect the frequency distribution of IMFs, has been proposed to be a measurement to distinguish speech frames and noise frames, resulting in our proposed WIFA-VAD algorithm. Intensive experimental results show that the proposed WIFA-VAD algorithm obtains a remarkable improvement of VAD accuracy under low SNR environment (SNR<0dB) compared with that of the conventional VAD algorithms.

* This work was supported by Shenzhen Science & Technology Fundamental Research Program (2011)

978-1-4673-2237-9/12/$31.00 ©2012 IEEE

II. HILBERT-HUANG TRANSFORM METHODOLOGY

In this section, we will introduce the HHT algorithm, which is constituted by the empirical mode decomposition and the Hilbert spectrum analysis, in details.

A. *The Empirical Mode Decomposition Method*

EMD is based on the assumption that any data consists of different simple IMFs so that the data can be represented by a linear combination of these IMFs and a residue. The definition of IMF follows two constraints [8]. First, among the whole dataset, the number of extrema and zero-crossing points must be equal, or only differ by one. Second, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima should be constant zero.

Let us take $x(t)$ to represent the target signal. EMD algorithm is described as follows [8]:

1) Identify all the maxima and minima of $x(t)$.
2) Generate the upper envelop of $x(t)$ via cubic spline interpolation with all the maxima; similarly generate the lower envelop with all the minima.
3) The mean of the upper and lower envelops is designated as m_1, and the difference between the data and m_1 is the first IMF candidate h_1, $h_1 = x(t) - m_1$.
4) Replace $x(t)$ by h_1 and repeat from step 1 to step 3, then $h_{11} = h_1 - m_{11}$. This is called the sifting process.
5) Repeat the sifting process for k times, that is $h_{1k} = h_{1(k-1)} - m_{1k}$. Stop until h_{1k} satisfies the previously defined constrains. Then the first order IMF is $c_1 = h_{1k}$, and the residue is $r_1 = x(t)$-c_1.
6) Repeat the procedure from step 1 to step 5 by sifting the residual data. The process will end when the residue satisfies a predefined stopping criterion, which is usually that the residue becomes monotonic or a constant. And then the original signal can be written in the form of (1). In (1), c_j is the j-th order IMF, r_N represents residue after N times of sifting.

$$x(t) = \sum_{j=1}^{N} c_j + r_N \tag{1}$$

In step 5), it states that h_{1k} should satisfy the previously defined constrains so that we can determine an IMF and move on to the next loop. But this statement is not practical yet. A mathematical loop stop criterion must be preset. There are two criteria that have ever been used. The first one requires SD_k should be small enough [8], which is the normalized squared difference between two successive sifting operations defined as

$$SD_k = \frac{\sum_{t=0}^{T} \left| h_{k-1}(t) - h_k(t) \right|^2}{\sum_{t=0}^{T} h_{k-1}^2} \ . \tag{2}$$

The second criterion is that the sifting process will stop only after S consecutive times, when the numbers of zero-crossings and extrema stay the same and are equal or differ at most by one [10]. S is a preset value. In this paper, the first loop stop criterion is chosen and SD_k is set to be 0.3.

B. *The Hilbert Spectrum Analysis*

After the acquisition of N order IMFs, it is of no difficult to perform Hilbert transform on IMFs. The complex conjugate $c_j'(t)$ of the j-th order IMF $c_j(t)$ can be determined by (3).

$$c_j{}'(t) = \mathrm{H}\left\{c_j(t)\right\} = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{c_j(\tau)}{t-\tau} d\tau \tag{3}$$

Then analytic function is defined as

$$z_j(t) = c_j(t) + ic_j'(t) \, . \tag{4}$$

The instantaneous amplitude $a_j(t)$ and the instantaneous phase function $\varphi_j(t)$ can be determined as

$$a_j(t) = \sqrt{c_j{}^2(t) + c_j'^2(t)} \ , \quad \text{and} \quad \varphi_j(t) = \arctan\left(\frac{c_j'^2(t)}{c_j{}^2(t)}\right) . \tag{5}$$

The instantaneous frequency can be simply represented by

$$f_j(t) = \frac{1}{2\pi}\frac{d\varphi_j(t)}{dt} \, . \tag{6}$$

According to Huang [11], for an IMF, IF reflects the frequency distribution, related to the signal energy, which motivates us to explore the novel approach for the voice activity detection technique, as described in section 3.

III. PROPOSED VOICE ACTIVITY DETECTION ALGORITHM

Evaluating the principles of the EMD technique, it is noted that EMD sequentially extracts the local, relatively high frequencies of the signals. This property endues EMD with some similar characteristics with a filter bank, which can be intuitively observed from Fig. 1. For a segment of noisy speech heavily polluted by Gaussian white noise (SNR=-5dB), the spectrogram of the original data and its first five IMFs extracted by EMD are displayed in Fig. 1, respectively. From Fig. 1 (a), we can see that the frequency band of the clean speech mainly ranges from 200-2500Hz, lasting from 0.3s to 1.2s. Moreover, from Fig.1 (b), the IMF$_1$ presents the highest frequencies from 2 kHz to 4 kHz and has little information of the speech. From Fig. 1 (c) to (f), it is noted that each IMF shows certain filter bank property. For example, let's look at the speech frame around 0.3s, from Fig. 1 (b) to (f), it is clear to see that the order of IMF increases, the highest speech frequency component goes lower at the rate of 2^{-j} (j=1,2…). This observation verified that EMD works similarly as a 2-base filter bank or 2-base discrete wavelet decomposition [8]. Moreover, part of the IMF$_2$, its spectrogram and the associated instantaneous frequency (IF) computed by (6) are shown in Fig. 2 (a) to (c), respectively. From Fig. 2 (c), it is also noted that, the instantaneous frequency varies with time and shows some random distribution properties. But we can see that the low IF value corresponds to high possibility of the presence of the speech. Similar conclusions can be drawn for other IMFs. More specifically, the instantaneous frequency has a decreasing trend when speech presents, therefore the instantaneous frequency average (IFA) over one frame may be a good indication to detect the activity of voice, which is defined as in (7),

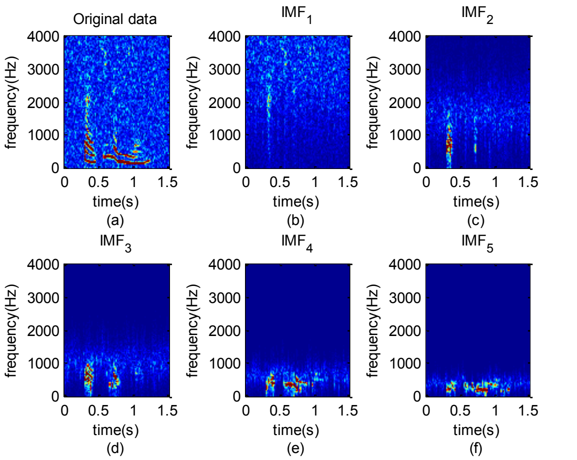

Figure. 1. Illustration of the filter bank property of EMD (SNR=-5dB)

$$p_{j,l}=\sum\nolimits_{t=0}^{T_f}f_{j,l}(t)/T_f \tag{7}$$

where T_f is the time duration of one frame, $f_{j,l}(t)$ is the instantaneous frequency, j is the IMF order and l is the frame index. From (7), the IFA ($p_{j,l}$) is essentially the mean value of the instantaneous frequency for one frame. Simulations show that the IFA of speech frames are most likely smaller than that of noise frames. To get better understanding, four IFA curves are computed and plotted in Fig. 3, where the speech frames and the noise frames are randomly selected. From Fig. 3, it is obvious that, for all IMF levels, the IFA for speech frame is lower than that of the noise frame, but the larger difference occurs at IFA at 2 and 3 IMF levels. This property can be further incorporated to develop a more robust VAD algorithm.

According to the discussions above, the weighted sum of IFA of each IMF, which is called weighted instantaneous frequency average (WIFA), is proposed as

$$P_l=\sum\nolimits_{j=2}^{N}2^{j-1}p_{j,l} \tag{8}$$

where N is order of IMF used.

For presentation clarity, the proposed VAD algorithm can be termed as WIFA-VAD, which is summarized as follows:

1) Divide noisy speech into N_f frames.

2) For each frame, calculate the first N order IMFs.

3) Perform Hilbert transform to the 2~N order IMFs and compute the instantaneous frequency $f_{j,l}(t)$ according to (3)-(6), where j is the IMF order and l is the frame index (l=1,…, N_f). Since IMF$_1$ mainly consists of noise information under low SNR condition, it is not taken into account.

4) Calculate IFA ($p_{j,l}$) and WIFA P_l by (7) and (8) , respectively. To remove the impact of outliers, a smoothing approach can be applied to get the final P_l curve, such as the median filtering.

5) Estimate the threshold T_S by using the first 10 noise frames, and which is computed by

$$T_S=\alpha E\left[P_l\right]. \tag{9}$$

where α is an controlling factor and can be selected as 0.92 to 0.95 according to some experiments.

6) For the l-th frame, if P_l exceeds $T_{S,}$ it will be marked as 1 (speech frame), otherwise as 0 (noise frame). Repeat till l=N_f.

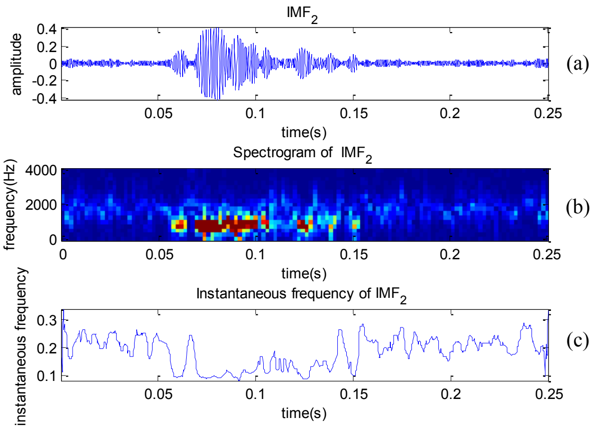

Figure. 2. IMF$_2$ and its instantaneous frequency (SNR=-5dB)

IV. Simulation Results and Analysis

Speech signals from the 863 Speech Corpus of Chinese Academy of Sciences, as well as Gaussian white noise from the NoiseX-92 Corpus, were adopted to evaluate the performance of the proposed WIFA-VAD algorithm. Twenty segments of male and female utterances spoken by different speakers were randomly selected. The sampling rate is 8 kHz and the frame length is of 32ms (T_f = 0.032). The 50% overlapping is applied. The testing noisy speech signals were generated by mixing speech signals with Gaussian white noise at different SNR levels (0dB, -5dB and -10dB). The performance of the VAD algorithm is measured by False Alarm Rate (FAR, the ratio of the number of falsely alarmed frames and the number of non-speech frames) and Missing Error Rate (MER, the ratio of the number of missed speech frames and the number of speech frames). FAR and MER were calculated by the VAD experimental result and the manually marked speech endpoints. For a robust and efficient VAD, both FAR and MER should be small enough.

Experiment 1: In this experiment, to visualize the performance of the proposed WIFA-VAD algorithm, the 15 second noisy speech is generated with SNR=-5dB, where the clean speech and noise are randomly selected. Fig. 4 shows the clean speech, noisy speech marked with the VAD result and the corresponding WIFA curve, respectively. From Fig. 4 (b), it is difficult to identify the speech frames since the clean speech is heavily polluted by noise. However, from Fig. 4 (c), the WIFA curve is able to detect probably the presence of the speech. The red line in Fig. 4 (c) is computed by 10 noise frames by using (9) with α=0.95. The WIFA-VAD result is plotted in Fig. 4 (b) shown as the red line, which corresponds to the clean speech quite well.

Experiment 2: This experiment is carried out to compare the

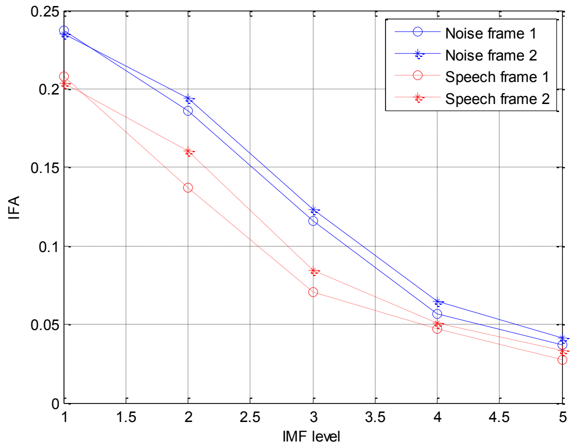

Figure. 3. IFA comparison between noise frames and speech frames (SNR=-5dB)

TABLE I. VAD PERFORMANCE COMPARISON

Algorithms	EZCR		LRT		LTSD		WIFA	
Criterions(%)	*FAR*	*MER*	*FAR*	*MER*	*FAR*	*MER*	*FAR*	*MER*
SNR=0dB	6.8	21.1	17.3	16.5	22.6	14.2	14.7	26.8
SNR=-5dB	3.8	56.5	16.3	32.4	12.9	37.4	9.7	29.3
SNR=-10dB	0.8	94.7	16.8	57.5	10.5	70.1	7.7	41.0

VAD performance among several VAD algorithms, which are WIFA-VAD, Sohn's VAD based on likelihood ratio test (LRT) [7], the long term spectral divergence (LTSD) VAD [10], as well as the classic VAD based on energy and zero-crossing rate (EZCR). The experimental results are shown in TABLE I. For each algorithm, all 20 segments have been applied for test and the average FAR and MER is recorded. For WIFA-VAD algorithm, the threshold is determined as the same as that in Experiment 1. From TABLE I. for SNR=-5dB and -10dB, the proposed WIFA-VAD outperforms than all other three VAD algorithms. But for SNR=0dB, the proposed WIFA-VAD has relatively high MER. This indicates that WIFA only has superiority when dealing with strong noisy conditions (SNR < 0dB). For FAR, the WIFA-VAD also outperforms the LRT and LTSD methods. Although the EZCR gives the lowest FAR, but its MER is worst, which means EZCR almost cannot detect the activity of voice for strong noisy speech.

V. CONCLUSION

This paper aimed at developing a robust VAD algorithm for strong noisy speech, particularly for conditions with SNRs lower than -5dB. In order to find more discriminative characteristics for speech/non-speech frame detection, HHT is used to analyze noisy speech signals. Based on the observations of filter bank property of EMD and IMF characteristics when speech presents, the WIFA is proposed to be the right characteristic for VAD. The threshold is estimated by using the noise frames, and then a simple binary rule is applied as the decision rule. The experiment results show the proposed WIFA-VAD algorithm has better anti-noise ability than the conventional VAD methods. But for higher SNRs (i.e. 0dB and above), WIFA are not superior to the existing VAD measurements. In our further work, the WIFA and other features of the IMFs, such as the energy of IMFs, will be jointly explored to improve the VAD performance for strong noisy speech.

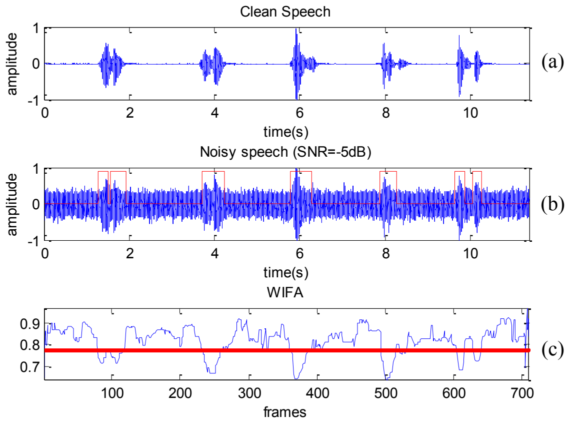

Figure. 4. VAD result of the proposed method (SNR=-5dB)

REFERENCES

[1] L. R. Rabiner, and B. H. Juang, *Fundaments of Speech Recognition*. Prentice Hall, 1993.

[2] J. A. Haigh, and J. S. Mason, "Robust Voice Activity Detection using Cepstral Features," *Proc. IEEE TEN-CON*, pp. 321–324, 1993.

[3] J. L. Shen, J. W. Hung, and L. S. Lee, "Robust entropy based endpoint detection for voice recognition in noisy environments," *Proc. ICSLP'96*, 1996.

[4] R. Tucker, "Voice Activity Detection Using A Periodicity Measure," *Proc. Inst. Electr. Eng.*, vol. 139, pp. 377~380, 1992.

[5] L. R. Rabiner, and M. R. Sambur, "Voiced - unvoiced-silence Detection Using the Itakura LPC Distance Measure," *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pp. 323–326, 1977.

[6] N. B. Yoma, F. Mciness, and M. Jack, "Robust Speech Pulse-detection Using Adaptive Noise Modeling," *Electron. Letter*, vol. 32, pp. 1350~1352, 1996.

[7] S. M. Jongseo Sohn, N. S. Kim, W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE signal processing letters*, vol. 6, No. 1, 1999.

[8] N. E. Huang, Z. Shen, S. R. Long, M. L. C. Wu, H. H. Shih, Q. N. Zheng, *et al*, "The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Nonstationay Time Series Analysis," *Proc. R. Soc., London*, vol. 454, pp. 903-995, 1998.

[9] Z. Lu, B. Liu, and L. Shen, "Speech Endpoint Detection in Strong Noisy Environment Based on the Hilbert-Huang Transform," *IEEE International Conference on Mechatronics and Automation*, Changchun, China, pp. 4322-4326, Aug. 9 - 12, 2009.

[10] N. E. Huang, M. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, *et al*,"A confidence limit for empirical mode decomposition and Hilbert spectral analysis," *Proc. R. Soc., London*, vol. 459, pp. 2317–2345, 2003.

[11] N. E. Huang, Z. Wu, S. R. Long, *et al*," On instantaneous frequency", *Advances in Adaptive Data Analysis*, Vol. 1, No. 2, pp. 177 – 229, 2009.

[12] J. Ramírez, J. C. Segura, C. Benítez, Á. de la Torre, and A. Rubio, "Efficient Voice Activity Detection Algorithms using Long-Term Speech Information", *Speech Communication*, Vol. 42, pp. 271-287, 2004.