Available online at www.sciencedirect.com

SciVerse ScienceDirect

Energy Procedia 17 (2012) 1102 – 1108

Energy Procedia

2012 International Conference on Future Electrical Power and Energy Systems

Application Research of HHT-IF Speech Feature Parameter in Speaker Recognition System

Liwei Liu,Feng Qian,Yao Zhang

College of Computer Science and Engineering,Changchun University of Technology ,Changchun, China

Abstract

Introduced the Hilbert-Huang transform (HHT) algorithm for nonlinear and non-stationary signal analysis. Specially to non-stationary speech signals, a new method of extracting the speech feature parameters is offered based on the HHT . The speaker identification system is designed based on the VQ and the experiments are carried out at different situations with both HHT-IF and LPCC. The results show that the HHT-IF is feasible for speaker recognition.

Keywords:Hilbert-Huang transform; feature parameter; Vector Quantization; speaker recognition

1.Introduction

Speaker Recognition is a biometric identification technology of distinguishing automatically speaker identity, which is according to speech parameter of the physical, psychological and behavioral characteristics of speaker which are reflected in waveform. The two key technologies in speaker recognition are individual feature parameter extraction and recognition model, particularly individual feature parameter extraction is the most critical. Currently, most of the speaker feature extraction methods are based on short-time stationary characteristic and Fourier transform of speech signals, losing the dynamic characteristics of speech signal. According to the research on speech signal, on the whole its features and parameters of characterizations are changing with time, which is a non-stationary process. With the appearance of signal processing methods of time-frequency analysis and wavelet analysis, we commence to study the method of speaker feature extraction using dynamic characteristics of speaker voice[1-4]. Hilbert-Huang transform(HHT)[5,6] is a new method to analyze nonlinear and non-stationary signal, it's signal processing method is different from the Fourier transform completely, which is not subject to the limitations of Fourier analysis, describing a spectrum of time-frequency amplitude. It is a more adaptive method of time-frequency analysis, it is considered as a major breakthrough to Fourier transform-based spectral analysis of linear and steady-state in recent years, it has been widely used in marine signal analysis, seismic signal analysis, graphics, image processing and other fields. This paper

 doi:10.1016/j.egypro.2012.02.214

describes based on HHT algorithms, it will be used in feature parameter extraction of speaker and applied to speaker identification, verifying the effectiveness of the parameters by experiment.

2.Hilbert-Huang Transform

Compared with various data analysis methods, the innovation of HHT is the introduction of IMF,which guarantees the physically meaningful Instantaneous Frequency. The HHT consists of two processes[1].

2.1.Empirical Mode Decomposition

The procedure of EMD decomposition is to shift the original data series until the signals are adaptively decomposed into a number of IMFs. Every IMF must satisfy two properties: (1) the number of extrema and the number of zero crossings are either equal or differ by one; (2) the mean value of the envelope defined by the local minima is constant zero. A special sifting process is employed to extract all of IMFs. This sifting process is described as follows.

Firstly, the upper envelopes and lower envelopes of signals $x(t)$, as well as their mean value $m_1(t)$,are calculated respectively. The first step of the sifting process is to calculate the difference:

$$h_1(t) = x(t) - m_1(t) \tag{1}$$

However, $h_1(t)$ rarely satisfies the two IMF properties and is taken as the first IMF of the signals straightway. Therefore, the sifting usually has to be implemented for more times, where the "difference" obtained in the previous sifting is taken as "signals" in present sifting. If after $(k+1)$ th sifting, corresponding difference $h_{1k}(t)$ satisfies the IMF properties,

$$h_{1k}(t) = h_{1(k-1)}(t) - m_{1k}(t) \tag{2}$$

then it can be taken as the first IMF component, denoted by $c_1(t)$, that is:

$$c_1(t) = h_{1k}(t) \tag{3}$$

In practice, to determine whether or not $h_{1k}(t)$ well satisfies the IMF properties, we usually use socalled standard deviation(SD) criterion, that is, to check if the following inequality holds[1]:

$$SD(k) = \sum_{t=0}^{T} \left[\frac{\left| h_{1(k-1)}(t) - h_{1k}(t) \right|^2}{h_{1(k-1)}^2(t)} \right] \le 0.2 - 0.3 \tag{4}$$

Where T is the length of data. Next, taking rest data

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

as "new" signals and implementing the sifting process on it, we can obtain the second IMF $c_2(t)$. This procedure should be repeatedly used for n times until the last residue $r_n(t)$ becomes a monotonic function. When the decomposition procedure finished, the signals then can be expressed as:

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

where $c_1(t), c_2(t), \cdots, c_n(t)$, are all of the IMFs included in the signals, and $r_n(t)$ is a negligible residue.

2.2.Hilbert Transform

As mentioned above, the main purpose of the EMD is to conduct the HT and obtain the Hilbert spectrum which is similar to wavelet spectrum. After conducting HT to every IMF component, $c_i(t)$, we have a new data series $y_i(t)$ in the transform domain:

$$y_i(t) = \frac{1}{\pi} P \int \frac{c_i(\tau)}{t-\tau} d\tau \tag{7}$$

where P indicates the Cauchy principle value. With this definition, a complex series $z_i(t)$ is formed:

$$z_i(t) = c_i(t) + jy_i(t) = a_i(t)e^{j\theta_i(t)} \tag{8}$$

where

$$a_i(t) = \sqrt{c_i^2(t) + y_i^2(t)} \tag{9}$$

$$\theta_i(t) = \arctan \frac{y_i(t)}{c_i(t)} \tag{10}$$

and the IF is:

$$\omega_i(t) = \frac{d\theta_i(t)}{dt} \tag{11}$$

Compared with the traditional FFT, $a_i(t)$ and $\omega_i(t)$ derived by HHT are functions of time t, not constant, which are different from FFT, so the HT can present the varying of the power with time.

3.Speaker Recognition Feature Parameter Extraction Based on HHT

There are lots characteristic parameters in the speech signal, different characteristic parameter, the significance of different physical and acoustical. Speaker recognition system needs to get good performance and should be extracted to characterize the basic features of the speaker characteristics. As the frequency is an important feature of the signal, while the instantaneous frequency of signal can be obtained is a distinctive feature of HHT, therefore, this paper using the instantaneous frequency of the voice signal as the characteristic parameters of the speaker by HHT called HHT-IF(HHT-Instantaneous Frequency).

Experience mode decomposition (EMD) is entirely based on signal decomposition, IMF after its decomposition can be viewed as the basic function of the original signal, the basis function is different from the basis function of wavelet transform, they don't have a fixed function form, once the signal change, they follow to change, so they have a good adaptability. Speech signal obtains much IMF after EMD, frequency components included changes from high to low, amplitude energy from large to small, however it is not consistent decrease and will fluctuate in the first few levels. Therefore, extracting the instantaneous frequency coefficient using IMF needs to determine which should be retained and which should be discarded. In this paper, EMD aims to three different samples of pure speech signal, the IMF components and the energy distribution which are received are analyzed and studied. Voice sampling rate is 8KHz, content is chinese "qin fen hao xue", one female voice and two male voice. As the process of voice recording may lead to inconsistencies in the energy distribution of voice signal, so voice signals need to be normalized first, making the speech signal energy in the same range and then to be processed. Fig. 1 is the first 8 IMF components after EMD of men voice. Voice energy and the energy levels after decomposing the proportion of the total energy shown in Table I.

TABLE I. EACH LAYER OF THE IMF'S ENERGY DISTRIBUTION PROPORTION AFTER EMD(%)

Sample	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8
Men sample 1	9	41.40	32.16	13.51	2.13	0.21	0.13	0.10
Men sample 2	13.04	35.24	25.80	13.18	3.22	0.43	0.42	0.72
Women sample 3	20.33	36.04	27.16	8.09	1.07	0.63	0.68	1.27

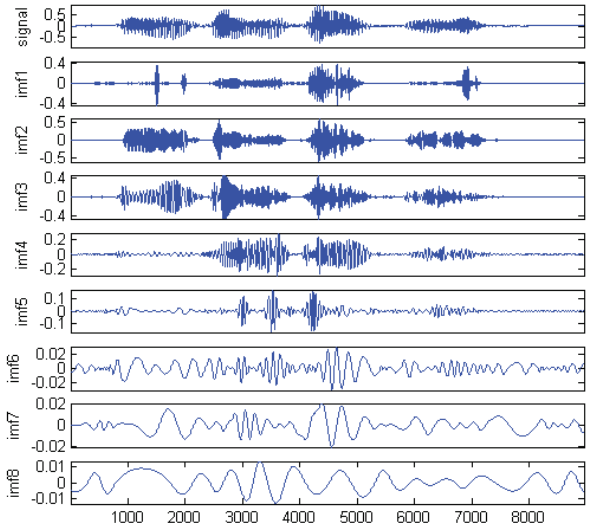

Figure 1. The first 8 IMF components after EMD

We can know though experimental analysis: Voice signal is broken down into sum of several IMF components by EMD. The first IMF contains mostly components of high frequency and has low energy, the main components of the signal distribute in the second, third and four IMF, the IMF components after IMF 6 contain no frequency components any more and the energy is almost 0. Therefore the text selects IMF1-IMF5 components to seek the instantaneous frequency coefficient, it as the feature parameters of the speaker in speaker recognition.

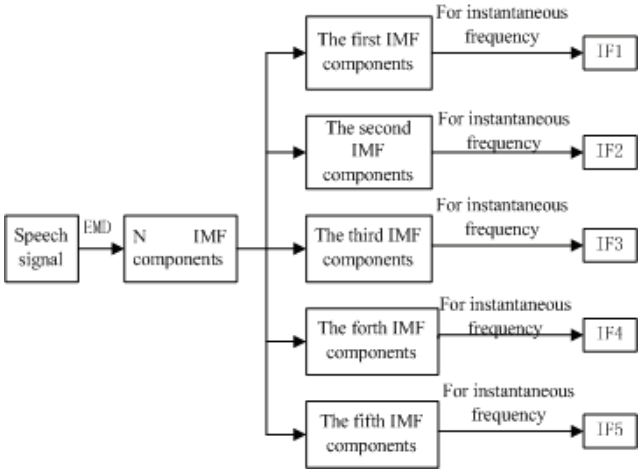

Figure 2. The block diagram of calculating the instantaneous frequency based on HHT

HHT-IF extraction block diagram shown in Fig.2, there are two sides in operation: 1) There are N IMF components after EMD of voice signal. As the rate of EMD decreases with the length of the data increases, voice signal may be separated to improve processing speed, each piece of signal stops after factoring out the five IMF, then link up the five IMF of every speech signal along the direction of time, thus we can obtain the required five IMF components of speech signal. 2) Hilbert transform to each IMF component and calculating the instantaneous frequency.

4.Experiment and Data Analysis of Speaker Recognition System

At present the main speaker recognition methods are Hidden Markov Model (HMM) based on parameter model and Vector Quantization (VQ) based on non-parameter model[7]. HMM-based method requires more model training data, longer training time, recognition time and large memory space; VQ-based method requires is small, so it is easy to implement. Therefore we use the method of VQ-based to implement speaker identification systems.

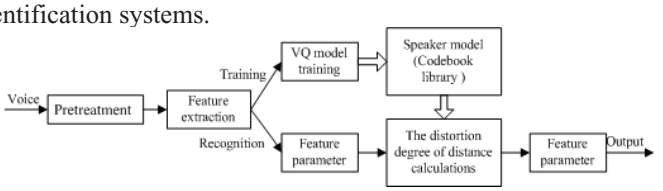

Figure 3. Diagram of speaker recognition system based on VQ

Fig.3 is the system block diagram of speaker recognition based on VQ, it consists two stages, training and recognition, everyone's feature vector data set should be clustered in training, cluster center as the code book to show speaker mode; According to voice will be measured and distortion measure of speaker codebook in recognition, according to some decision rules to identify or confirm. There are two methods to extract characteristic parameters in the design process in system: one is LPC cepstrum coefficients which is more mature, one is instantaneous frequency coefficients based on HHT, then speaker identification separately, In order to compare the new feature parameters-performance of HHT-IF which is put forward.

In order to verify the performance of speaker recognition system, it is needed to establish a voice database for evaluation. We record a 40 voice library from the actual situation. Recording officers are college students from the country who are about 20 and speak Mandarin. They sound more natural and individual local color slightly. Recorded under conditions of normal laboratory environment, using the ordinary sound card, sampling frequency is 8KHz, PCM mode, accuracy of quantify is 16bits, recording number are 14 (24 men, 16 women).In order to show the characteristics having nothing to do with the text, everyone reads 2 minutes to obtain data from different texts, a part of these as the training data set, the other as the test data set.

The performance evaluation of speaker recognition system has a variety of indicators, the most important is the recognition accuracy of the results. For the speaker identification systems of this thesis, the performance evaluation is the rate of correct recognition, error rate ($E_{ID} = n_{err} / n_{tot}$) or correct rate ($C_{ID} = n_{cor} / n_{tot}$) can be directly used, n_{tot}, n_{err} and n_{cor} are the total numbers, errors or correct numbers. Recognition rate in different environment and different speakers may be different, but the basic performance evaluation is consistent. The performance of speaker identification system is analyzed in experiment from different perspective in this article.

On the one hand: Parameters based on 16-order LPCC and HHT-IF, the size of codebook is 32, identification number is 20, speech length of training and identification are in different situations, recognition of speaker identification system which is having nothing to do with the text, data as Table II.

TABLE II. THE CORRECT RECOGNITION RATE OF THE SPEAKER RECOGNITION SYSTEM IN DIFFERENT TRAINING AND RECOGNITION TIME SITUATION

	Training 2s Identify 2s	Training 4s Identify 2s	Training 4s Identify 4s	Training 8s Identify 2s	Training 8s Identify 4s	Training 8s Identify 8s
HHT-IF	65%	75%	90%	85%	100%	95%
LPCC	70%	80%	90%	85%	100%	95%

On the other hand: Parameters based on 16-order LPCC and HHT-IF, the length of training speech is about 8 sec, the length of recognition speech is about 4 sec, identification number is 40, the size of codebook is in different situations, recognition of speaker identification system which is having nothing to do with the text, data as Table III.

TABLE III. THE CORRECT RECOGNITION RATE OF THE SPEAKER RECOGNITION SYSTEM IN DIFFERENT CODEBOOK SIZE SITUATION

	Codebook size 16	Codebook size 32	Codebook size 64
HHT-IF	77.5%	87.5%	95%
LPCC	80%	85%	92.5%

On the third hand: Parameters based on 16-order LPCC and HHT-IF, the size of codebook is 32, the length of training speech is about 8 sec, the length of recognition speech is about 4 sec, identification number is in different situation, recognition of speaker identification system which is having nothing to do with the text, data as Table IV.

TABLE IV. THE CORRECT RECOGNITION RATE OF THE SPEAKER RECOGNITION SYSTEM IN DIFFERENT RECOGNITION NUMBER SITUATION

	5 people	10 people	15 people	20 people	25 people	30 people	35 people	40 people
HHT-IF	100 %	100 %	100 %	100 %	96%	93.3 %	91.4 %	87.5 %
LPCC	100 %	100 %	100 %	100 %	92%	90%	85.7 %	85%

Though the above experimental data of the speaker identification system from different angles shows, the performance of the speaker identification system of the design based on HHT-IF characteristic is better than the LPCC characteristic. In addition, for the speaker identification system, the longer of the use of training and the voice length of recognition, the higher of recognition rate of the system; the bigger of codebook size, the higher the recognition rate of the system, however due to the LBG vector quantization algorithm determines the code size increase is the increasing exponentially, it allows training speed and recognition respond speed of the system to slow, the increase of recognition rate is not significant, so the codebook size is 32 what this paper used. For speaker identification system, with the increased of the participation of a closed set identification number, the system recognition rate will decrease, for less than 20 cases, identification rate of identification system reached 100% based on LPCC and HHT-IF feature parameter.

5.Conclusion

This article describes a new method to deal with the leaner and non stationary signal—Hilbert-Huang Transform, using the distinctive feature of it can be obtained the instantaneous frequency of signal, proposing the HHT-IF to extract the feature parameters of speech signal based on HHT, and through experiments using vector quantization, we design and implement speaker identification system having nothing to do with this text based on HHT-IF and LPCC. Though the data in voice set, from different angles of speaker identification results show that HHT-IF feature parameters extracted based on HHT for speaker recognition is feasible. As the voice is a typical non-stationary signal, HHT in the voice signal processing has a broad development.

References

[1] Zhao Zheng and Hou Boheng,"Research of speaker recognition technology based on wavelet sransform", Xidian university journals, vol 27,no 4,pp.437-441,2000.

[2] S.Samuel. Speaker Identification Using Neural Networks and Wavelets[J]. IEEE Trans. On Engineering in Medicine and Biology(S1524-1904), 2000(1,2):92-100.

[3] S.C.Woo, C.P.Lim,R.Osman. Development of A Speaker Recognition System Using Wavelets and Artificial Neural Networks[C]. International Symposium on Intelligent Multimedia, Video and Speech Processing, 2001, Hong Kong: 413-416.

[4] Yu Lingli."Wavelet transform processing method of speech signal", Changchun university of techonology journals,vol 26,no 3, pp 229-232, 2005.

[5] Norden E. Huang. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-stationary Time Series Analysis[J]. Proc.R.Soc.Lond.A(S1364-5021), 1998, 454:903-995.

[6] Norden E. Huang, Shen Z, Long S R. A New View of Nonlinear Water Waves: the Hilbert Spectrum[J]. Annual Review of Fluid Mechanics(S1091-6490),1999,31:417-457.

[7] Zhao Li, The speech signal processing Beijing: Mechanical industry press,pp244-249,2003.