TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 14, No. 1, April 2015, pp. 116 ~ 122
DOI: 10.11591/telkomnika.v14i1.7216

Robust Pitch Detection Based on Recurrence Analysis and Empirical Mode Decomposition

Jingfang Wang
School of Information Science & Engineering, Hunan International Economics University,
Changsha, China, postcode: 410205
E-mail: matlab_bysj@126.com

Abstract

A new pitch detection method is designed by the recurrence analysis in this paper, which is combined of Empirical Mode Decomposition (EMD) and Elliptic Filter (EF). The Empirical Mode Decomposition (EMD) of Hilbert-Huang Transform (HHT) is utilized tosolve the problem, and a noisy voice is first filtered on the elliptic band filter. The two Intrinsic Mode Functions (IMF) are synthesized by EMD with maximum correlation of voice, and then the pitch be easily divided. The results show that the new method performance is better than the conventional autocorrelation algorithm and cepstrum method, especially in the part that the surd and the sonant are not evident, and get a high robustness in noisy environment.

Keywords*: empirical mode decomposition, recursive analysis, elliptic filter, intrinsic mode function, pitch detection*

1. Introduction

Pitch refers to the hair caused by vocal fold vibration during voiced periodicity pitch is the reciprocal of the frequency of vocal fold vibration [1]. Speech signal pitch is to describe one of the important parameters, in the tone recognition, emotion recognition, speech recognition, speaker recognition, speech synthesis and coding, music retrieval, sound system, diagnosis, hearing impairment and many other areas of language instruction has wide range of applications [2]. Because speech is a dynamic process is non-stationary random process, so changes in the waveform is extremely complex, not only the size of the pitch period length of individual vocal, thickness, toughness and pronunciation habits, but also with the pronunciation of age, gender, pronunciation , the intensity and emotional articulation, and many other factors. At present, the harder to find a common approach to extract accurately and reliable voice in any case, the pitch period, so the estimated pitch period is the study of speech processing field has been hot and difficult one.

We use elliptic band-pass filter (Elliptic Filter) [3] to preprocess the signal to eliminate the introduction of high-order harmonic distortion and noise, the singular point, so has the physical meaning of the signal component of a complete linear superposition stand out the way, and then use the EMD method selected correlation with physical meaning we need the mode signal. Then with voice recursive analysis of dynamic characteristics of pitch detection is combined. Additive broadband noise with a variety of voice test, the method can accurately detect the pitch period, so that further reduce the detection error, and has good robustness.

2. Elliptic Filter with the Pitch Detection Process

2.1. Elliptic filter

Elliptic filter (Elliptic filter) [3], also known as Kaul filter (Cauer filter), is in the passband and a stopband equiripple filter. Elliptic filter compared to other types of filters, in order under the same conditions with the minimum passband and stopband, fluctuations in transition zone decreased rapidly; the transition zone is very narrow. It is in the passband and stopband of the fluctuations in the same, which is different from the passband and stopband are flat Butterworth filter, and a flat passband, stopband equiripple stop-band or a flat passband ripple, etc. Chebyshev filter.

Received December 20, 2014; Revised February 26, 2015; Accepted February 15, 2015

This 4-order elliptic band-pass filter, the maximum attenuation of 0.05dB passband and minimum stopband attenuation of 80dB, passband region 2 * [75,500] / fs, fs the sampling frequency (Hz). When the fs = 19.98kHz to obtaining the filte (1) (Omission).

2.2. Pitch Detection Process

Noisy speech underwent a 4-order elliptic band-pass filter, filter out high frequency and low frequency below 60Hz, and calculated to the first N_0 elliptical filtering of data as the initial standard deviation of the noise section of Q0 (EMD as a basis for access); then 20-30ms long framing; of each frame signal of a recursive two-degree threshold frame voicing decision, voiceless frame zero pitch, or determine the initial section of the standard deviation of the noise Q0 size. If Q0 <α (eg α = 0.15), the recursive analysis of direct access to pitch, voiced frame or quasi-variance calculations Q. When Q <kQ0 (k constant), voiceless frame pitch to zero, otherwise the EMD decomposed IMF components on different scales associated with the decomposition of the signal prior to calculation, take the maximum correlation of the two modes (IMF) synthetic pitch signal, again Synthetic spectrum calculated for the second request signal pitch.

Voiced frame of recursive frequency signal analysis and calculation: Statistics recurrence plot parallel to the main diagonal length of each DG (k), k = 1,2, .., N-1, N is frame length. $n_0 = [\frac{f_s}{f_0}]$, f_s is the sampling frequency, f_0 upper frequency limit for the pitch, [x] said that the greatest integer not exceeding z, find the max (DG (k> n_0)) corresponds to the number n, the pitch frequency: $f_J = \frac{f_s}{n}$.

3. Empirical Mode (EMD) Decomposition and Pitch Automatic Synthesis

3.1. Empirical Mode Decomposition (EMD)

Assumption of signal, EMD IMF component selection to achieve the following steps:

First find the signal maximum points and minimum of all data points, fitted by cubic spline interpolation to obtain the signal envelope and the next on the envelope, to ensure that all points on the two envelopes in the Between the upper and lower envelope by calculating the mean of each point, to obtain a mean curve, and define the signal minus the corresponding point of the sequence of the new data available $h_1^{(1)}(t)$:

$$x(t) - m_1(t) = h_1^{(1)}(t) \tag{2}$$

If $h_1^{(1)}(t)$ meet the conditions of IMF components, $h_1^{(1)}(t)$ is I the first order IMF component. Otherwise, $h_1^{(1)}(t)$ continue to repeat the process times, until $h_1^{(n)}(t)$ meet the convergence criteria, then the first order component of the $x(t)$'s IMF:

$$C_1(t) = h_1^{(n)}(t) \tag{3}$$

$C_1(t)$ is the most high-frequency components. Subtracted $C_1(t)$ from the original signal to obtain first-order residual term $r_1(t)$:

$$x(t) - C_1(t) = r_1(t) \tag{4}$$

Then, $r_1(t)$ repeat the process to get the second order IMF component $C_2(t)$. This continued through the EMD decomposition of the signal a second round selection to get some order IMF

components and a residual component r_n, the entire decomposition process is complete. After the decomposition, the original signal $x(t)$ can be expressed as:

$$x(t) = \sum_{i=1}^{n} C_i(t) + r_n(t) \quad (5)$$

Finally, the EMD decomposed IMF components $C_i(t)$ of each order contained in the signal reflects the characteristics of different time scales, on behalf of non-linear signal from the high-frequency modes to low frequency vibration modes inherent characteristics, so that you can make in different signal characteristics Resolution display, in order to achieve multiresolution signal capacity; that $r_n(t)$ is the trend term or mean of $x(t)$. EMD decomposition to avoid the energy loss caused by the wavelet transform to overcome the energy leakage. Using (5) can reconstruct the original signal.

3.2. Automatic Synthesis Pitch

Elliptic filter through the noisy speech (1) filtering after that x (t), the main ingredients for the pitch; noise when the band is still strong (Q0 larger), the use of EMD (5) decomposition. Calculated the correlation coefficient:

$$R(i) = \frac{\mathrm{cov}(x, C_i)}{STD(x) * STD(C_i)} \qquad i = 1,2,\ldots,n \quad (6)$$

Where cov is the covariance, STD is standard deviation. Let R (i) by order of the first two serial number for i (1), i (2), the synthetic pitch is:

$$x_J(t) = C_{i(1)}(t) + C_{i(2)}(t) \quad (7)$$

4. Recursive Analysis

Recursive analysis is a nonlinear dynamic analysis method, it is based phase space reconstruction, reflecting the recovery after the chaotic attractor has a law. Different nature of the state of the signal characteristics of the track not the same as, and in the recurrence plot (Recurrence Plot RP) of the structure is different [4, 5]. Thought algorithm described as follows:

(1) Select the appropriate time delay τ and embedding dimension m, the one-dimensional reconstruction of nonlinear time series, the resulting dynamic system is as follows:

$$X_i = (x(i), x(i+\tau), \ldots, x(i+(m-1)\tau)) \quad (8)$$

More than one-dimensional time series that is re-pose-dimensional phase space trajectory, from the perspective of dynamical systems to achieve a recovery in the high dimensional space attractor.

(2) Calculate the phase space rows X_i, columns X_j, and the distance between vectors:

$$S_{ij} = \left\| X_i - X_j \right\| \quad (9)$$

Where $\|x\|$ is Euclidean norm.

(3) Recursive calculation

$$R_{ij} = \Theta\left(\varepsilon - S_{ij}\right) \quad (10)$$

Where, ε is the critical distance, $\Theta(x)$ is said step (Heaviside) function, $\Theta(x)=\begin{cases}0 & x\le 0\\ 1 & x>0\end{cases}$. Nodes use the phase space can be described from two-dimensional graphics on the internal dynamics of nonlinear time series matrix of the mechanism, the recurrence plot (RP). $R_{ij}=1$ is corresponding position at the time that the black point, $R_{ij}=0$ is white point when you said, RP is through the black point and figure point to describe the white graphics to reflect the time series.

In order to quantitatively from a statistical point of view of signal analysis, in the recurrence plot is introduced based on the signal to be measured degree of recursion [5]:

$$R_{\varepsilon}=\frac{1}{N^{2}}\sum_{i,j=1}^{N}R_{ij} \tag{11}$$

Where, N for RP map up, the column vector of nodes. Clearly, R_{ε} is a cumulative distribution function, which describes the phase space attractor is less than the distance between two points on the probability of ε, depicted relative to the phase space of a reference point Xi in the ε phase points within the aggregate level. So, R_{ε} as the correlation integral function attractor [6]. If ε is too small to obtain a result from the large $\| X_i - X_j \|$ than ε, then $\Theta(x)$ = 0, summation, R_{ε} = 0 indicates the distribution of phase points outside the ε. If ε too big election, all "points of " no more than the distance from it, then R_{ε} = 1. Therefore, ε is too large or too small can not reflect the system's internal nature. In general, ε the emulated to make 0 ≤ R_{ε} ≤ 1 makes sense. Proposed degree of signal recursion system for the analysis of signals in the dynamic complexity provides a theoretical method.

The main diagonal straight line parallel to the R_{ε} = 0 we call recurrent points:

$$D_{\varepsilon}(k)=\sum_{i=1}^{N-k}R_{i,i+k}\qquad k=1,2,\cdots,N-1 \tag{12}$$

Its size reflects the strength of the system periodically. In speech signal processing, we take the embedding dimension m = 1, time delay τ = 0, the sampling frequency fs (Hz), pitch frequency limit 500Hz, $ks=[\frac{fs}{500}]$.

$$D_{\varepsilon}(k_0)=\max\{D_{\varepsilon}(k)\quad k=ks,ks+1,\cdots,N-1\} \tag{13}$$

5. Experimental Evaluation

Background noise taken from Noisex-92 database [9], and its sampling frequency fs = 19.98kHZ. Here we have the same sampling frequency fs, the noise in the computer record and interior noise environment, "language, tone, end point" sound shown in Figure 1(a), the method frame Voicing line for the verdict. Process in the voice sub-frames, each frame taking 25ms, the frame length N = [0.025fs] point, frame shift $\frac{N}{2}$.

Experiment 1: The original voice, original voice and noise Noisex-92 library of white noise (white) were used in this method signal to noise ratio 10db, 5db, 0db,-5db, respectively, under the pitch detection shown in Figure 4, the figure Left part of the horizontal axis is time (seconds), vertical axis is amplitude, the right side of the abscissa is the number of frames,

respectively, the vertical axis pitch frequency (Hz) signal with the elliptical filter recursive degrees. Ministry left diagram of voice, speech mixed with different noise (blue), elliptical filtered signal (black) and voicing their discriminant results, the algorithm for the detection of the central figure to the pitch frequency, the corresponding figure for the right of the elliptical filter Voicing signal recursive degrees and split double threshold discriminant line.

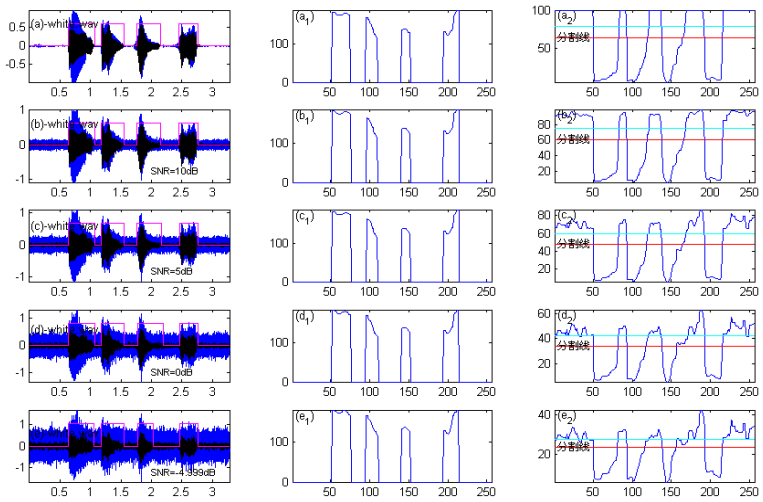

Figure 1. The original voice mixed white noise (white) with different SNR Comparison of Fundamental Frequency Detection algorithm

Experiment 2: For non-stationary noise. The original voice, original voice and noise Noisex-92 library in the car noise (volvo), burst engine (destroyerengine) noise, factory noise (factory), were noisy noise (babble), respectively, the method used in the signal to noise ratio (SRN) Pitch detection under the 0db were shown in Figure 5, the legend above.

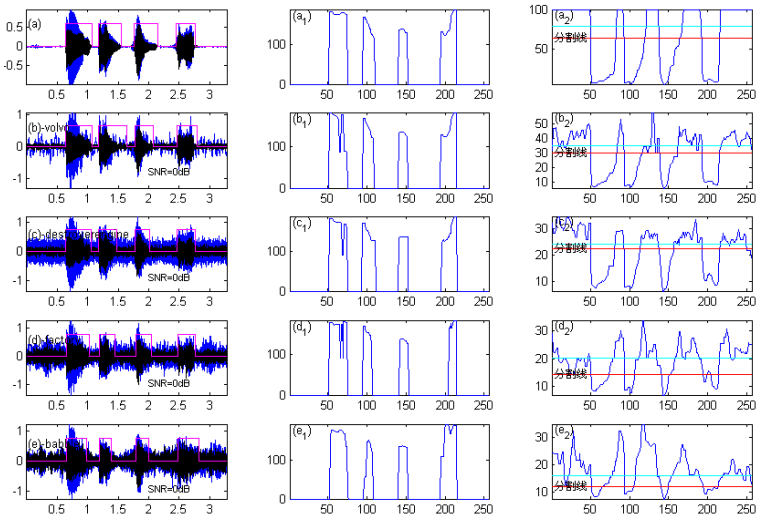

Figure 2. Original speech mixed with different noise (SNR = 0dB) algorithm under the Fundamental Frequency Detection with Comparison

(a) Original speech and the voicing decision, (a1) of the original voice pitch frequency detection, (a2) of the original audio frequency signal recursive degrees;

(b) Hybrid car noise (volvo) speech (SNR = 0dB) voicing decision, (b1) hybrid vehicle noise tone pitch frequency detector, (b2) Recursive hybrid vehicle noise level low frequency signal;

(c) Hybrid motor (destroyerengine) speech noise (SNR = 0dB) voicing decision, (c1) hybrid engine noise tone pitch frequency detector, (c2) Recursive hybrid engine noise level low frequency signal;

(d) Blending plant noise (factory) speech (SNR = 0dB) voicing decision, (d1) pitch frequency sound mixing plant noise detection, (d2) Recursive hybrid plant noise level low frequency signal;

(e) Loud noise mixed people (babble) speech (SNR = 0dB) voicing decision, (e1) were noisy mixed tone pitch frequency detector noise, (e2) mixed low-frequency signals were loud noise recursive degrees.

Experiment 3: The TIMIT speech database. Here the performance of the new method with the traditional center of the autocorrelation function of clipping method [10] and cepstrum [11] compared the performance and evaluation. Test performance indicators used are as follows:

1) Voicing The accuracy (ASR-Acoup Sur Ratio): the right to determine the existence of fundamental frequency of the number of frames in the voice as a percentage of the total number of frames. The higher the index, then determine whether the cyclical performance of voice, the better.

2) The effective fundamental frequency relative error (VPRE-Valid Pitch Relative Error): In the standard frame fundamental frequency is not zero, the calculation of non-zero value of the fundamental frequency and the reference fundamental frequency divided by the square error between the reference RMS Mean fundamental frequency. The lower the index, the algorithm accuracy as possible.

Table 1 Three ways in different signal to noise ratio (SNR) performance

signal to noise ratio		SNR=5dB			SNR=0dB		
Noise	function	New method	Autocorrelation	Cepstrum	New method	Autocorrelation	Cepstrum
pink	ASR(%)	96.88	95.08	92.58	96.88	93.07	90.29
	VPRE	0.1335	0.2200	0.2510	0.2202	0.2581	0.2871
f16	ASR(%)	96.88	95.02	92.30	97.66	92.76	90.10
	VPRE	0.1253	0.2050	0.2460	0.1667	0.2811	0.3125
factory	ASR(%)	96.09	94.50	91.80	92.97	90.56	79.62
	VPRE	0.0636	0.2310	0.2720	0.1950	0.2987	0.3547
babble	ASR(%)	95.31	94.03	91.09	91.80	89.45	68.00
	VPRE	0.1226	0.2430	0.2750	0.1758	0.3125	0.4526

As can be seen from Table 1, the new method of voicing error rate lower than the traditional autocorrelation and cepstrum, which cepstrum worst. This is mainly because only cepstrum cepstrum or using complex cepstrum and pitch in if there are peaks corresponding to distinguish the voicing sound and estimated pitch period, voiced in some cases, but sometimes not particularly prominent peak point, and in the case of voiceless but there will be some occasional peaks, resulting in larger Voicing misjudged and effective base frequency error; autocorrelation with relatively fixed clipping threshold, half-octave higher frequency phenomena, and thus also affect the effective fundamental frequency error; oval filtered high-frequency filter, empirical mode decomposition (EMD) of the signal filtering, filtered half-frequency harmonic generation SHG, can effectively filter out on the pitch detection is not the necessary information, and signals of different amplitude can be simplified, thus improving the classification rate Voicing, fundamental frequency reduces the effective error.

6. Conclusions and Outlook

With the lower signal to noise ratio, recursive phase space reconstruction analysis forecasting performance of attractor increasingly blurred, leading to the signal sequence complexity of the analysis dropped. Hilbert-Huang transform with empirical mode decomposition (EMD) to obtain a finite order intrinsic mode function (IMF), each one of the IMF components

typically have real physical meaning, respectively, the characteristic scale of signal parameters in a frequency band information The mode function can reflect the signal at any time with the frequency characteristics. Therefore, this combination of empirical mode decomposition of noisy speech signal pre-noise reduction, and then select the high-dimensional phase space reconstruction of attractor description of the characteristics of the signal to achieve the look and accurate pitch extraction of speech start and end point and purpose.

Voice signal is one-dimensional time-domain signal, using empirical mode decomposition (EMD) correlation, recursive analysis and elliptic filter (EF) the combination of process it, the results show that the method can effectively suppress noise, highlighting the signal periodic structure, weakening caused by the resonance frequency peak half-frequency phenomenon. And can accurately distinguish voicing low tone of voice, voicing transition section of the pitch discrimination is more accurate, and the algorithm is simple and fast. Experiments show that this method can resist the noise interference, is robust, being able to accurately extract the pitch of the cycle to achieve the extraction of the signal, keeping purposes detail and noise suppression.

References

[1] CHUN J, SYING J, ZHANG R. TRUES: Tone recognition using extended segments. *ACM Transactions on Asia Language Informationprocessing*. 2008; 7(3).
[2] FERRER C, T0RRES D, HERNANDEz DIAZ ME. *Contours in the Evaluation of Cycle-to-Cycle Pitch Detection Algorithms*. Proceedings of the 13th Iberoamerican congress on Pattern Recognition. 2008.
[3] Gao Quan, Ding Yume, Wide Yonghong. Digital signal processing theory, implementation, and application. Beijing: Electronic Industry Press. 2007: 144-151.
[4] Marwan N, Thiel M, Nowaczyk NR. *Cross recurrence plot based synchronization of time series*. Nonlinear. Proc. Geophys. 2002; 9: 325-331.
[5] Yan Yun-Qiang, Zhu Yi-Sheng, The voice signal endpoint detection method is based on a recursive analysis. *Communications*. 2007; 28(1): 35-39.
[6] Spib Noise data. http://spib.rice.edu/spib/select_noise.html.
[7] RABINE RLR. On the use of autocorrelation analysis for pitch detection. *IEEE Tram ASSR*. 1977; 25(1): 24-33.
[8] KOBAYASHI H, SHIMAMURA T. *A modified cepstmm method for pitch extraction.* IEEE APCCAS. 1998: 299-302