TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol. 12, No. 12, December 2014, pp. 8205 ~ 8211
DOI: 10.11591/telkomnika.v12i12.6482

Pitch Detection Based on EMD and the Second Spectrum

Jingfang Wang
School of Information Science and Engineering, Hunan International Economics University, Changsha, China, postcode: 410205
email: matlab_bysj@126.com

Abstract

A new method for pitch detection of secondary spectrum is designed in the paper, the noisy speech oval (Elliptic Filter, EF) band-pass filter is designed first in this method, and then the experience mode Decomposition(EMD)of Hilbert-Huang transform (HHT) is used to decompose the signal into a finite number of intrinsic mode functions (IMF), and IMF components of different scales are associated with the decomposition of the signal before calculation, the maximum of two modes associated (IMF) synthetic pitch signal detection is taken. Experimental results show that the method could be better than the traditional autocorrelation method, and cepstrum method has better results, especially with voicing obvious segment features, there is better performance of pitch detection in noisy speech, signal to noise ratio(SNR) also has good robustness in the lower sound environment.

Keywords*: empirical mode decomposition (EMD), elliptic filter (EF), intrinsic mode function (IMF), secondary spectrum, 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 accurate 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.

Pitch estimates usually known as pitch detection. The current pitch detection is mainly based on the traditional voice model, can be divided into time domain, frequency domain and time-frequency domain mixing. Among them, the most representative is the time-domain autocorrelation (ACF) method and the average magnitude difference (AMDF) method [3], but the ACF method is easy to have a "multiplier" and "half frequency" error, AMDF method can not effectively track voice rapid changes in frequency, so when rapid changes in voice frequency and amplitude, the pitch detection accuracy decreased significantly. Commonly used frequency cepstrum [4], the introduction of the number of operations, the calculation of the digital signal processing increased significantly, and so vulnerable to the effects of noise, pitch detection accuracy decreased. Combination of time-frequency wavelet transform domain is also highly vulnerable to the impact of noise, with the voice signal to noise ratio decreased, increasing the error detection.

In 1998, Nordeng E. Huang proposed a non-stationary signal adaptive decomposition of the program: first signal EMD (Empirical Mode Decomposition, referred to as EMD) [5], screening a series of single-frequency narrow-band mode component (available the way they are constructed by linear superposition of the original signal), then every single frequency band component of the Hilbert transform mode, the instantaneous received signal when time-

Received July 14, 2014; Revised October 17, 2014; Accepted November 4, 2014

frequency distribution and distribution. The program is called the Hilbert-Huang Transform. The method there is a prominent problem, if the signal exists in singularity, the EMD filter out the first mode component will contain the singular point, while the normal signal component will be pushed to the next mode after the component level by level of so that the physical meaning of the original with a complete signal components are filtered to different components of the mode. The absence of a unified solution to the problem, according to the specific characteristics of the analysis signal processing. Many problems for the above problem, we use elliptic band-pass filter (Elliptic Filter) [6] 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. And then combining with the second spectrum of pitch detection. 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. Eelliptic Filter

Elliptic filter (Elliptic filter) [6], 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.

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 60 Hz, and calculated to the first N0 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 (x(i), i = 1,2, ..., L) the average energy ($\frac{1}{M}\sum_{i=1}^{L} x^2(i)$) of the double threshold for the frame voicing decision, pitch set the frame voiceless Zero, or determine the standard deviation of the initial noise segment Q0 size. If Q0 <α (eg α = 0.15), directly calculated from the secondary pitch frequency spectrum, or quasi-variance calculations voiced frame 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 Synthesized signal seek a second spectrum calculating pitch.

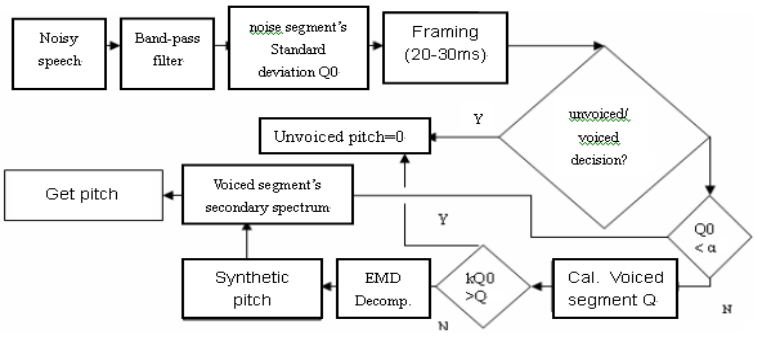

Figure 1. pitch detection process

Voiced frame of a second frequency signal spectrum calculation: fx = IFFT (|FFT (x) | ^{2}). $n_0 = [\frac{f_s}{f_0}]$, fs is the sampling frequency, f0 upper frequency limit for the pitch, [x] said that the greatest integer not exceeding z, find the max (fx (i> n0)) 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)

EMD is to decompose the signal to be the time delay between adjacent peaks is defined as the time scale, nonlinear, non-stationary signal screening, broken down into different time scales contain a limited number of intrinsic mode function (Intrinsic Mode Function, IMF) component of and, decomposition of the order IMF components are stationary narrow-band signals. IMF component must meet the following two conditions: (1) For a list of characteristics of the data from the global point of view, the number of extreme points have an equal number or a maximum difference of zero point; (2) at a certain local point by point and the maximum The definition of the minimum point of the two zero mean envelope, that envelope up and down on the timeline of local symmetry. EMD decomposition of the above two conditions is the end of the convergence criteria, each one of the IMF components can be considered as an intrinsic signal mode function.

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:

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

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:

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

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:

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

Figure 2 shows the ellipse of a voice filtered through EMD decomposition (n = 7) received the first 4 IMF components and synthetic (IMF2 + IMF3) pitch signal. Figure 2 shows, the signal layers from high to low frequency filters, each one of the IMF showed a component of the modal scales, and no modal overlap. The EMD decomposition of a synthetic pitch frequency signal filtered with a half-frequency harmonics, dynamic screen automatically combined.

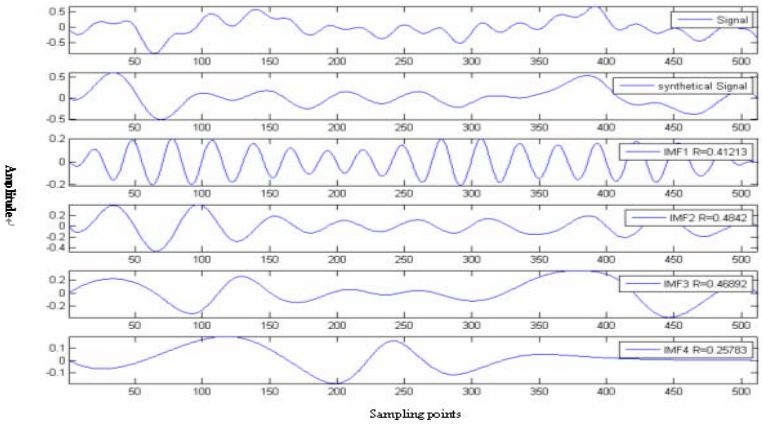

Figure 2. EMD decomposition of speech signal synthesis and pitch

4. Experimental Evaluation

Background noise taken from Noisex-92 database [7], 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 M = [0.025fs] point, frame shift $\frac{M}{2}$.

***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 3, 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) filtered signal with the average energy ellipse. 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 the average signal energy and dual threshold discriminant dividing line.

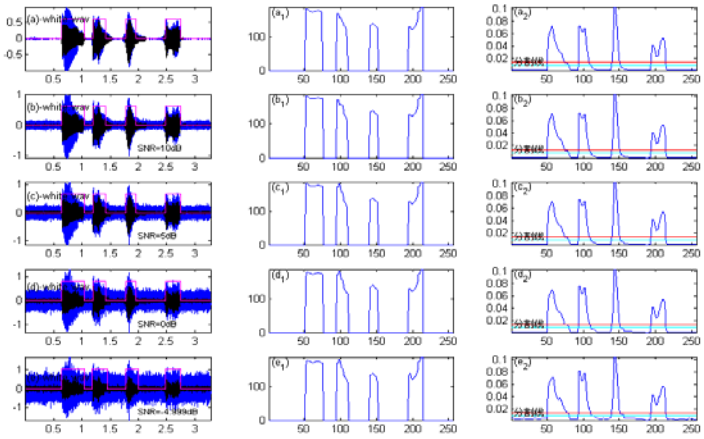

Figure 3. The original audio mix of white noise (white) with different SNR Comparison of Fundamental Frequency Detection algorithm

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 loud noise (babble), respectively, the method used in the signal to noise ratio (SRN) pitch detection under the 0db were shown in Figure 4, the legend above.

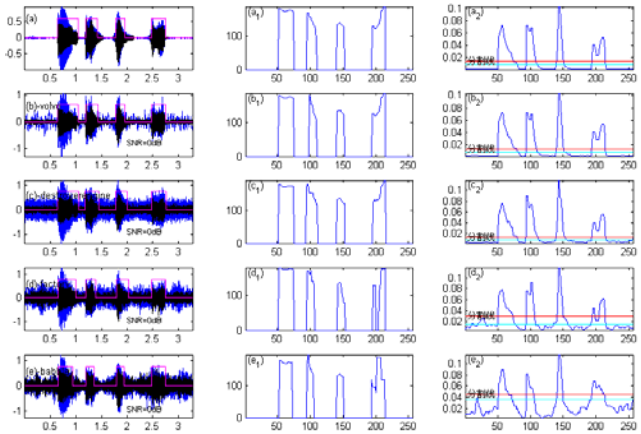

Figure 4. The original speech mixed with different noise (SNR = 0dB) algorithm under the Fundamental Frequency Detection with Comparison

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

(b) Hybrid car noise (volvo) speech (SNR = 0dB) voicing decision, (b1) hybrid vehicle noise tone pitch frequency detector, (b2) hybrid car noise, the average energy of low-frequency signal;

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

(d) Blending plant noise (factory) speech (SNR = 0dB) voicing decision, (d1) Mixed plant noise tone pitch frequency detector, (d2) the average mixed-signal low-frequency noise power plants;

(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 average energy.

Experiment 3: The TIMIT speech database. Here the performance of the new method with the traditional center of the autocorrelation function of clipping method [8] and cepstrum [9] to compare and evaluate the performance. 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.

5. Conclusions and Outlook

Voice signal is a one-dimensional time-domain signal, empirical mode decomposition (EMD) associated method is combined with the elliptic filter (EF) to process signal, the research results show that this method can effectively suppress noise, prominent signal periodic structure, it can weaken doubling phenomenon which is caused by the formant. And voicing and low tone voice can accurately distinguished, pitch discrimination of Voicing transition section is more accurate, and the algorithm is simple and fast. Experimental results show that the method can resist the interference noise, there is better robustness, the pitch period can be more

accurately extracted, signal extraction is achieved, signal detail is maintained and noise is suppressed.

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_DIAZM E. *Contours in the Evaluation of Cycle-to-Cycle Pitch Detection Algorithms.* Proceedings of the 13th Iberoamerican congress on Pattern Recognition. 2008.
[3] LI H, DAI B-GI, LU W_A Pitch Detection algorithm based on Amdf and Acf. *Digital Object Identifier*. 2006; 1: 14-19.
[4] KOBAYASHI H, SHIMAMURA. A modified cepstmm method for pitch extraction. *IEEE APCCAS.* 1998. 299-302.
[5] Huang N E, Shen Z, Long S R, et al. *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-station time series analysis*. Proceeding of the Royal Society of London, 454(A). 1998; 903-995.
[6] Gao Xiquan, Ding Yumei, Kou Yonghong etc. *Digital signal processing – principles, implementation and application.* Publishing House of Electronics Industry; Beijing, China. 2007; 144-151.
[7] Spib Noise data[EB/OL], http://spib.rice.edu/spib/select_noise.html
[8] RABINERLR. On the use of autocorrelation analysis for pitch detection. *IEEE Tram ASSR* 1977; 25(1): 24-33.
[9] KOBAYASHI H, SHIMAMURA T. A modified cepstmm method for pitch extraction. *IEEE APCCAS.* 1998; 299-302
[10] Mohd R Jamaludin, Sheikh HS Salleh, Tan T Swee, Kartini Ahmad, Ahmad KA. Ibrahim, Kamarulafizam Ismail. An Improved Time Domain Pitch Detection Algorithm for Pathological Voice. *American Journal of Applied Sciences*. 2012; 9(1): 93-102.
[11] He Ba, Na Yang. Ilker Demirkol and Wendi Heinzelman. *BaNa: A Hybrid Approach for Noise Resilient Pitch Detection*. IEEE Statistical Signal Processing (SSP), Ann Arbor, Michigan. 2012.
[12] J Bartošek. A Pitch Detection Algorithm for Continuous Speech Signals Using Viterbi Traceback with Temporal Forgetting. *Acta Polytechnica*. 2011; 51: 5,8-13.