에너지 연산자에 기초한 간단한 피치 추적 방법

A Simple Pitch Tracking Algorithm based on the Energy Operator

이 태 호*
Tai-Ho Lee

요 약

유성음의 피치주파수 궤적을 추정할 수 있는 새로운 방법을 제시하였다. 이 방법은 에너지연산자[1]를 두 번 적용하는데 기초하고 있다. Kaiser의 에너지연산자는 정현파의 진폭과 주파수 정보를 추출하는 기능을 가지고 있다. 변조모형에 의하면 유성음은 피치 신호로 변조된 포만트들의 합성으로 파악될 수 있으므로 이 파형의 진폭 포락선을 추출해서 피치 신호와 유사한 파형을 얻는다. 이 파형의 평균 주파수를 검출하여 피치 주파수를 구하는 것이다. 앞부분은 Gopalan의 접근법[9]과 마찬가지이나, 뒷부분의 LPC-스펙트럼 분석등의 과정 대신 또 한번 에너지 연산자를 적용하도록 하여 매우 단순화 되고 온라인 적용이 가능한 알고리듬을 얻었다. 추정 결과는 거친 편이지만 온라인으로 피치 궤적의 일반적 스케치를 얻는데 유용할 것으로 기대된다.

Abstract

A new method for the estimation of pitch-frequency contour of voiced speech is presented. The method is based on the double application of Kaiser's energy operator[1], which has the capabilities of extracting amplitude and frequency of a sinusoidal waveform. According to the modulation model, a vowel can be represented by a combination of damped sinusoids representing formants, modulated by pitch pulses. Therefore, the amplitude envelope of each of the components will give a pitch-like waveform and the pitch can be obtained by averaging the frequencies of this waveform. The first part is the same as Gopalan's approach[9], but by substituting the LPC based spectral analysis with the second application of energy operator, the algorithm becomes very simple and can be processed on-line. Although the estimation is rather coarse, the suggested algorithm can be useful for getting a general sketch of pitch contour on-line.

Key words : Pitch frequency, energy operator.

I. Introduction

Since Kaiser has introduced a nonlinear signal processing function by the name of *energy operator* (EO) [1], there have been various efforts to investigate the characteristics of it and to apply it to practical purposes.[2]-[9] The fundamental importance of this operator is that it reveals time-varying portions of an AM-FM signal. For continuous-time signals the energy operator is defined as following:

$$\Psi[x(t)] \triangleq (\dot{x})^2 - x\ddot{x}\ , \tag{1}$$

where $\dot{x}$ means time derivetive of x. An AM-FM signal can be represented by

$$x(t) = a(t)\cos[\phi(t)], \tag{2}$$

where

$$\dot{\phi}(t) = \omega_i(t) = \omega_c + \omega_m q(t).$$

That is, $x(t)$ is a sinusoidal with time-varying amplitude $a(t)$ and angular frequency $\omega_i(t)$. If the amplitude and frequency varies slowly with respect to the carrier, ω_c, then it can be shown that [2]

$$\Psi[x(t)] \approx a^2(t)\omega_i^2(t) \tag{3}$$

$$\Psi[\dot{x}(t)] \approx a^2(t)\omega_i^4(t)\ . \tag{4}$$

Using (3) and (4), the *amplitude envlope*(AE), $|a(t)|$

*울산대학교 전기전자정보시스템공학부
접수 일자 : 2003. 8. 22 수정 완료 : 2003. 10. 27
논문 번호 : 2003-4-3

and the *instantaeous frequency*(IF), $\omega_i(t)$ can be obtained. These AM-FM detection capability attracted many attempts of their application to speech processing.

Among them, there appear two types of pitch estimation methods: one applies energy operator directly to the speech signals[8], and the other uses the AE and IF values[9]. In the first method the speech signal or its lowpass-filtered one is applied directly to the equation (3) expecting some enhancement may occur to the pitch-pulse features in the speech signals. The output waveform from it is then fed to a sequence of pitch location processes, that is, center-clipping, allocating candidates, eliminating unlikely candidates, etc. Not only the use of energy operator in this method does not seem to improve the process or the result, but also it is not well justified theoretically as will be mentioned later. The second method appears quite promising, especially the one using AE. But the process that follows is a batch job including LPC based spectral analysis and peak picking etc., which cancels out the merits of simplicity and the cheap computational cost of energy operators.

Our motivation is to substitute this batch process with the IF extracting algorithm, so that the overall system is composed of two energy operator based stages in cascade. And the system becomes structurally very simple and logically elegant, and what is more, it can be operated on-line.

In II a brief review is given on the discrete energy operator fundamentals, and III is for system structure. The simulation results are given in IV, and the conclusion in V.

II. Discrete Energy Operator Fundamentals

We consider a discrete AM-FM signal of the form:

$$x(n) = a(n)\cos[\Omega_i(n)\, n]\ , \tag{5}$$

where

$$\Omega_i(n) = \Omega_c + \Omega_m q(n)\ .$$

$\Omega_i = \omega_i T$, etc. and T is the sampling period. Then the discrete versions of (1), (3), and (4) are [5]

$$\Psi[x(n)] \triangleq x^2(n) - x(n-1)x(x+1) \tag{6}$$

$$\Psi[x(n)] \approx a^2(n)\sin^2[\Omega_i(n)] \tag{7}$$

$$\begin{aligned}\Psi[x(n) &- x(n-1)] \\ &\approx 4a^2(n)\sin^2[\Omega_i(n)/2]\sin^2[\Omega_i(n)]\end{aligned} \tag{8}$$

From these equations several versions of energy separation algorithms have been generated.[2][5] One of those is given below.

$$\Omega_i(n) \approx \cos^{-1}[G(n)] \tag{9}$$

$$|a(n)| \approx \sqrt{\frac{\Psi[x(n)]}{1-G^2(n)}} \tag{10}$$

where $G(n) = 1 - (\Psi[y(n)] + \Psi[y(n+1)])/4\Psi[x(n)]$, and $y(n) = x(n) - x(n-1)$. From (9) and (10) we get instantaneous frequency(IF), $\Omega_i(n)$ and AM envelope(AE), $|a(n)|$.

III. Algorithm and Implementation

A) Pitch extraction algorithm

Overall system is shown in Fig. 1, and the flow of the processing is as follows:

Step 1. Speech signal is bandpass filtered to get an AM signal of the form of (5).

Step 2. The amplitude envelope AE1= $|a(n)|$ is obtained applying (10) to the result of step 1.

Step 3. AE1 wave form is smoothed by a lowpass filter.

Step 4. IF2, the instantaneous frequency of AE1, is obtained applying (9) to result of step 3.

Step 5. The estimated pitch frequency is obtained by averaging the past values of IF2.

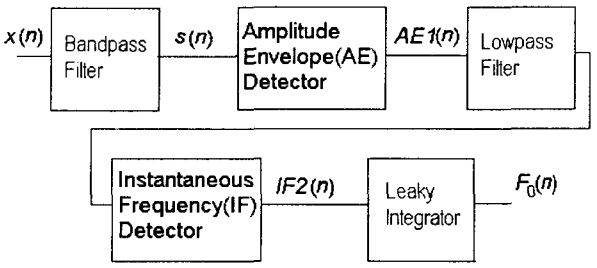

Figure 1. Block diagram of overall system

B) Simulation design

1. Bandpass Filter: According to the modulation model[10], a vowel can be represented by a combination of damped sinusoids representing formants, modulated by pitch pulses. This can be represented by

$$s(n) = \sum_{k=1}^{M} a_k(n)\cos(\Omega_k n)\ . \tag{11}$$

By applying a narrow bandpass filter, a signal with the form of (5) is extracted. The best choice for center frequency will be the formant frequencies, but similar

effect can be expected at other frequencies. To minimize the side lobes a Garbor filter is used,[6] i.e.,

$$h(n)=\exp(-b^2n^2)\cos(\Omega_c n) \quad -N \le n \le N, \quad (12)$$

where Ω_c is center frequency and b controls bandwidth. The half length of filter is chosen by $N=2.5/b$.

2. AE1 detection: Some careful arrangements are needed for the application of (9) and (10) because these are not well-behaving functions. Firstly, (6) can result a negative value, which makes the word 'energy' meaningless. Secondly, in (9) and (10) $G(n)$ should be kept between 0 and 1. To be safe from these problems the input signal should be a single AM-FM signal moderately modulated[3], which is a condition that a speech signal can rarely meet. In [8], (6) is applied directly to the speech signal (original or lowpass filtered). In such a situation there exists no concept of carrier and we can not tell the meaning of the results which is supposed to be an amplitude envelope.

3. Lowpass filter: It is to shape the AE signal more like a sinusoid. A butterworth type is used.

4. IF2 detection: Same precaution applies as in the case of AE detection.

5. Updating pitch frequency: IF2 is a candidate of pitch frequency. However, since IF2 values are not stable, we have tried a few kinds of time averaged value, one of which can be represented by the equation:

$$F_0(n)=\frac{1-r}{1-r^n}\sum_{k=1}^{n} r^k \, \mathrm{IF2}(k) \quad (13)$$

where $F_0(n)$ is the pitch frequence, and r is the forgetting factor. With the value of r ranging (0, 1), the summation part in (13) performs a leaky integration.

IV. Simulation Results

A simulation result is shown in Figure 2, where speech data are taken from Childers' f0625s[11], representing a female voice for 'we were away for a year ago'. The center frequency and the bandwidth of Gabor filter used were 800 Hz and 400 Hz, respectively. In Figure 2(c), resultant pitch frequency contour is given with a reference, which is obtained by the speech analysis toolbox by Childers. A few 'bad' portions are seen in the figure, which correspond to: ① silent(and unvoiced), or ② transient portions of speech. The portions corresponding ② is not obvious in the raw data shown in Figure 2(a), but AE1 values in (b) become much smaller and irregular. This means that the bandpass filtered signal can be much weaker and more unreliable at certain part of the original signal than other parts.

The choice of center frequency of Gabor filter can be critical for some cases. Figure 3 compares the effects of two different center frequencies, 800 Hz and 1 kHz. Although the difference of center frequency is only 200 Hz, the estimated pitch contours show obvious difference.

In Figure 4 and 5, the effect of pitch frequency of speech signal on the processed waveforms are compared. The case of high pitched voice(f0625s, $F_0 \approx 200\,\mathrm{Hz}$) is shown in Figure 4, where the waveform of first AM detection(AE1) is fairly close to a sinusoid, and IF2, the instantaneous frequencies of AE1 show small fluctuations within each pitch period. Figure 5 shows the same variables as Figure 4 for a low pitched voice(m0125s, $F_0 \approx 100Hz$). In this figure we can see that AE1 differs very much from the sinusoidal waveform resulting in large fluctuations in IF2. This fluctuation may result in a large error for very low-pitched voices since the equation (13) for Step 5 is an averaging function of a simple RC Lowpass filter type.

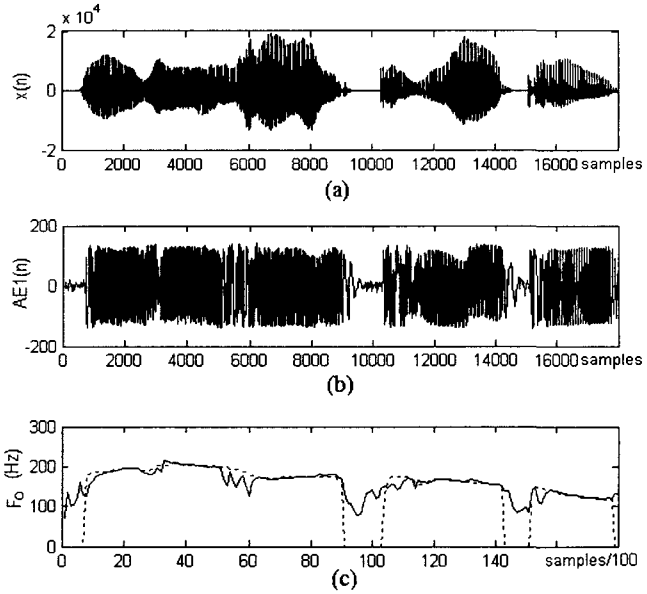

Figure 2. Pitch tracking simulation for f0625.
(a) Original speech signal. (b) Amplitude envelope, AE1. (c) Pitch contours: Simulation result(solid line) and reference(dotted line)[11]

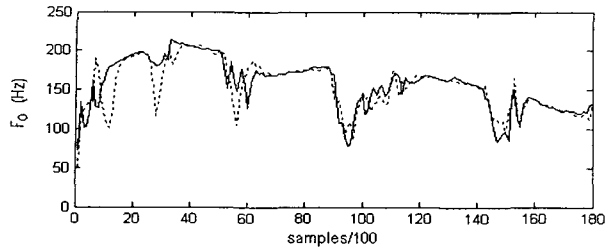

Figure 3. Pitch contours obtained at different center frequencies: Solid line is for 800 Hz, and dotted line is for 1 kHz.

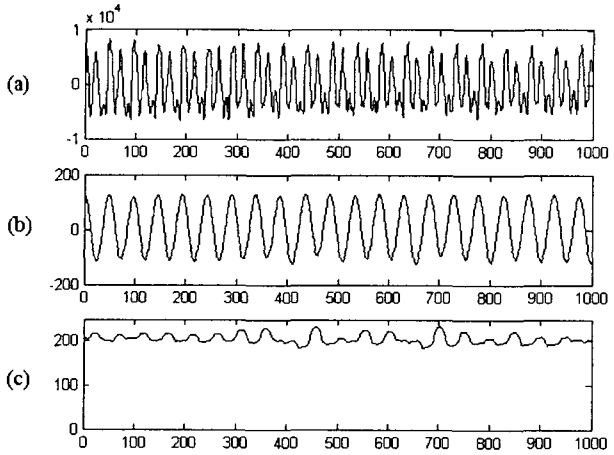

Figure 4. Wave forms for a high pitched signal. (a) Original signal, (b) Amplitude envelope(AE1) (c) Instantaneous frequency(IF2) in Hz.

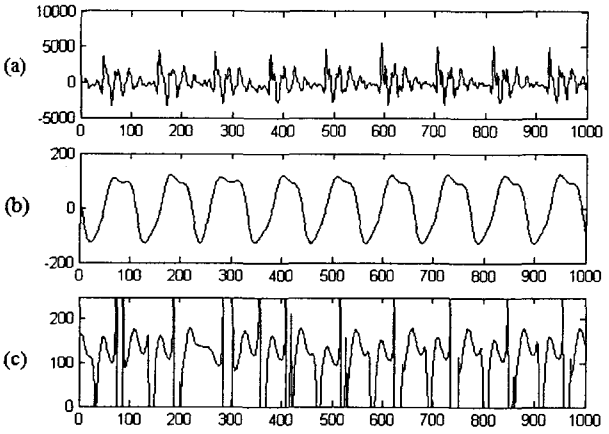

Figure 5. Wave forms for a low pitched signal. (a) Original signal, (b) Amplitude envelope(AE1) (c) Instantaneous frequency(IF2) in Hz.

V. Conclusion

We have presented a simple algorithm which can give rough sketch of pitch contour on-line. The algorithm can not lead to exact pitch contour, and can be unreliable under certain conditions, but its simplicity and on-line applicability make it attractive for certain applications. For practical application some elaborations are needed. Some of these problems are discussed below:

1) The estimated pitch contour is sensitive to the choice of center frequency of the bandpass filter. If the frequency range of the Gabor filter falls on the valley of speech spectra, resulting outputs become very weak and unstable. Because the speech is time-varying signal, it is hard to select a center frequency to match all through the sentence. A multi-frequency approach[12] can be used to solve this problem, that is, several Gabor filters with different center frequencies are applied concurrently, and then, one of the outputs are chosen by an appropriate automatic selection rule.

2) It is mentioned in IV, that the fluctuation in IF2 (Figure 4,c and 5,c) may give rise to a large error in low-pitched case. The knowledge of the glottal waveform can be used for the refinement. Since waveforms of AE1 (Figure 4,b and 5,b) are related to the glottal waveform, a correction table may be designed based on the statistical study of average frequency of AE1 versus actual pitch frequency. This correction will need extra delay for final decision of pitch estimate.

3) Silent/Voiced/Unvoiced detection may be incorporated to improve the situation around the segment boundaries.

There have been two types of successful pitch tracking algorithms. One is time-domain processing type based on the inverse filtering and autocorrelation.[13] This type of estimators can be highly reliable, and extensively used for analysis and coding. The other is frequency-domain processing type based on harmonic sine-wave model[14,15], which gives average pitch value for each analysis frame.

Our EO-based algorithm shows inferior results to those of existing ones. But much refinement can be expected when above suggestions are carried out. Also, if the harmonic sine-wave model is combined with EO-based algorithm, another reliable off-line pitch estimator can be obtained.

References

[1] J. F. Kaiser, "On a simple algorithm to calculate the

'energy' of a signal" *Proc. IEEE ICASSP 90*, Albuquerque, NM, Apr. 1990, pp.381-384

[2] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "On separating amplitude from frequency modulations using energy operators", *Proc. IEEE ICASSP 92*, San Francisco, CA, Mar. 1992, pp.II-1-4

[3] A. C. Bovik and P. Maragos, "Conditions for positivity of an energy operator", *IEEE Trans. on Signal Processing*, vol. 42, no. 2, Feb, 1994, pp.469-471

[4] D. Dimitriadis and P. Maragos, "An improved demodultion algorithm using splines", *Proc. IEEE ICASSP 01*, Salt Lake City, UT, May 2001, pp. 3481-3484

[5] P. Magros, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis" *IEEE Trans. on Signal Processing*, vol. 41, no. 10, Oct. 1993, pp.3024-3051

[6] -----,"Speech nonlinearity, modulation, and energy operators", *Proc. IEEE ICASSP 91*, Toronto, Canada, May 1991, pp.421-424

[7] T. F. Quatieri, C. R. Jankowski, Jr., and D. A. Reynolds, "Energy onset times for speaker identification", *IEEE Signal Processing Letters*, vol. 1, no. 11, Nov. 1994, pp. 160-162

[8] R. K. Whitman and D. M. Etter, "An investigation of estimating pitch period using a non-linear differential operator", *Record of 28th Asilomar Conference on Sinals, Systems and Computers*, Pacific Grove, CA, Oct.-Nov. 1994, pp.1237-1241

[9] K. Gopalan, "Pitch estimation using a modulation model of speech", *Proc. WCCC-ICSP2000*, Beijing, Aug. 2000, pp.786-791

[10] A. B. Finberg, R. J. Mammone, and J. L. Flanagan, "Application of the modulation model to speech recognition", *Proc. IEEE ICASSP 92*, San Fransico, Mar. 1992, pp. 541-544

[11] D. G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, Inc. 2000

[12] A. C. Bovik, P. Maragos, and, T. F. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators", *IEEE Trans. on Signal Processing*, vol. 41, Dec. 1993, pp. 3245-3265

[13] K. A. Oh and C. K. Un, "A performance comparison of pitch extraction algorithms for noisy speech" *Proc. IEEE ICASSP 84*, Mar. 1984, vol. 9, pp. 85-88

[14] S. Seneff, " Real-time harmonic pitch detector", *IEEE Trans. on Aucoustics, Speech, and Signal Processing*,vol. ASSP-26, no. 4, Aug. 1978, pp. 358-365

[15] R. J. McAulay and T. F. Quatieri, " Pitch estimation and voicing detection based on a sinusoidal speech model", *Proc. IEEE ICASSP 90*, Albuquerque, NM, Apr. 1990, vol. 1, pp. 249-252

이 태 호 (Tai-Ho Lee)
正會員
1966 한양대학교
1969 서울대학교(공학석사)
1975 연세대학교(공학박사)
1969-1972 한국과학기술연구소
1973-현재 울산대학교 전기전자정 보시스템공학부 교수
관심분야 : 음성신호처리, 인공신경망, 통신시스템