IRT Filtering 법에 의한 음성신호의 기본주파수 수출

Inverse Rate Type Filtering for the Pitch Extraction

*배 명 진 (Bae, M. J.)
**안 수 길 (Ann, S. G.)

요 약

음성 신호 처리 분야에서 고속이며 정확히 기본 주파수를 추출하는 방법은 중요하며, 이를 위한 많은 연구가 제안되어 왔다. 이러한 방법들은 보통 성도의 방해를 제거한 후에 기본 주파수를 검출하기 때문에 지금까지는 처리과정이 복잡하다.

우리는 그러한 과정을 간단히 처리할 수 있는 Inverse Rate Type(IRT) Filtering 법을 제안하였다. 제안된 방법은 유한 길이의 정수 계산이고 덧셈과 뺄셈으로 처리될 수 있기 때문에 범용 마이크로 컴퓨터에 의해서도 실시간 처리를 할 수 있게 된다.

ABSTRACT

In the area of speech signal processing, estimating an accurate pitch period in real time is an important problem; and many methods have been presented. Most of these methods require rather complicated operations, since they estimate the pitch period after eliminating the effect of vocal tract.

In this paper, we propose the Inverse Rate Type (IRT) filtering technique which can estimate the pitch period on an efficient way. The method can estimate the pitch period in real time using a general purpose microcomputer, since it employs finite word length arithmetics with only addition and subtraction operations.

*서울대학교 전자공학과 박사과정,
**서울대학교 전자공학과 교수

I. INTRODUCTION

Since vocoder theory was investigated by Dudley in 1939., many methods on the data compression of speech signals have been developed. One of them is the source coding method which proceeds the speech production model by dividing it into the excitation part and the filtering part. This method can compress the data more than any other wave coding methods. To characterize the excitation part in the source coding method such as LPC vocoder, the pitch extraction is needed.

Available pitch extraction methods until now can be essentially divided into two types as follow:

* Direct Detection Methods: Parallel Processing Method(1), Average Magnitude Difference Function(1), Area Comparison Method (4), etc.

* Post Detection Methods: Simplified Inverse Filter Tracking method(6), Center Clipping Autocorrelation Function Method(1), etc.

The latter is to find the pitch period after eliminating the interference of the resonance characteristics of the vocal tract to the fundamental frequency, and the former is to find the pitch period from the unmodified speech samples. In both methods we generally eliminate the high frequency formants by a preceding low pass filter. The formants adjacent to the fundamental frequency, however, are not eliminated. Thus, the direct detection methods in general take rather complex decision algorithm so that the halving or the doubling of the pitch which are caused by the existance of the major formants, could be avoided.

On the other hand, the post detection methods estimate continuously the major formants by an inverse filter, and remove them so that a formant free pitch could be obtained. It is accounted for the fact that the post detection methods take a long computation time and do not adapt higher pitch speakers like children because of the limited number of filter order. (1)

In this paper, we proposed an inverse rate type filter whose gain is inversely proportional to the frequency, instead of the ordinary low pass filter with a sharp cutoff characteristic. The filter can be realized by associating filters with ramp function characteristics with different cutoff frequencies: The ramp function could be approximated by a sinc squared function. The resulting impulse response is a trianglar wave shape in time domain. The number of computation steps for the present speech sample by this method is ten additions and nine subractions which is considerably simple.

II. DERIVATION OF THE CHARACTERISTICS FOR THE IRT FILTERING

Considering a voiced speech in frequency domain we observe that the components of the fundamental frequency are the origin of the fine structure in the resonance curve of their vocal tract. The lowest frequency of those peaks in the curve is called as the first formant, F1. Since the energy of the first formant is higher than the others, the effect of the first formant is observed clearly in the time domain, and since the first formant is most adjacent to the fundamental frequency, it can strongly influence the precision in extracting the fundamental frequency.

As the range of the fundamental frequency being generally 40-400Hz, most pitch extractors before the application of the algorithm, carry

out the pre-filtering, that is, low pass filtering in order to eliminate the effect of formants. However, the first formant being generally close to the fundamental frequency, can not be eliminated by the ordinary LPF.

It is necessary to estimate the first formant and substract it in order to reduce the interference from the original speech wave. For this purpose the pitch extractor uses a coefficient calculator for the vocal tractor parameter calculation and the obtained parameters are applied to the inverse filter for the elimination of the vocal tract characteristics, and then the periodic component(which is the pitch) is detected from the residual signal (7).

Since the computation time depends on the number of the filter order, the signal is preprocessed by a LPF with the cutoff frequency of about 900Hz in order to eliminate the higher frequency formancs and the method be able to use a simple inverse filter, and by consequence simple calculation. But as this method is available upto no more than 250Hz pitch frequency. It is impossible to apply this method to high pitch speakers namely women and children. Also the calculation process is still too complicated for a microcomputer to finish the calculation in real time(1).

To resolve those difficulties, it will be better to execute the calculations of the preprocessing filter and the inverse filter at the same time. We have devised an inverse rate type filter with the characteristic shown in Fig. 2-a in which the gain is inversely proportional to the frequency. The spectrum of a typical voiced speech show in Fig. 1 demonstrates the known fact that the first formant is higher in frequency than the fundamental frequency and there is almost no correlation between them. Therefore, it can be suppressed an elements of the first formant in comparison to the fundamental frequency, as an error signals of inverse filter, when the voiced speech signals are passed in the IRT filter. And since the IRT filter has a cutoff frequency as an preprocessing filter, the upper side components of the cutoff frequency in voiced speech signals can be removed.

As the first formant frequencies generally are higher than 250Hz, we have adjusted the center point of the inverse rate filter around 230Hz. And as the frequency range of the pitch being between 40Hz and 400Hz, the cutoff frequency of the filter is fixed to 460Hz. To avoid the boosting of DC component, the gain under 40Hz is kept constant.

The inverse rate type filter, we propose, is the combination of three ramp filters of which the characteristics are shown in the Fig. 2-a. These ramp characteristics can be approximated by sinc squared functions as shown in Fig. 2-b. The sinc(.) function in frequency domain express a step function in time domain; i.e.,

$$g(n) = u(n) - u(n-N) \quad \langle - - \rangle \quad H(jwT) = \mathrm{sinc}\,(NwT/2) \tag{1}$$

Since the first zero point of the sinc(.) function is to take in the variable as pi, if the pi-point of sinc(.) is 615 Hz, then the length, N, of the step in time domain is

$$pi = 615\,Hz = 2q/N \quad \text{or} \quad N = 8\,KHz/615Hz = 13 \tag{2}$$

Where q is nyquist frequency, 4KHz. Using Eq. (1), it is easily shown that the mathematical representation for convolution is given as

$$A(n) = g(n) * s(n) = \sum_{k=-\infty}^{\infty} g(k) \cdot s(n-k) = \sum_{k=0}^{N-1} s(n-k) \tag{3}$$

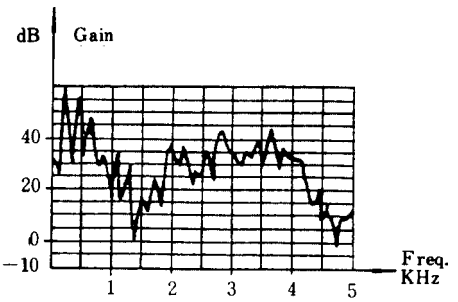


fig. 1 Spectrum of the voiced speech "i"

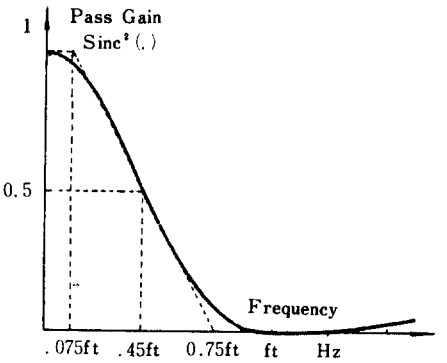


Fig.2-b Equivalent Function, to the Ramp Function: $Sin^2c(.)$

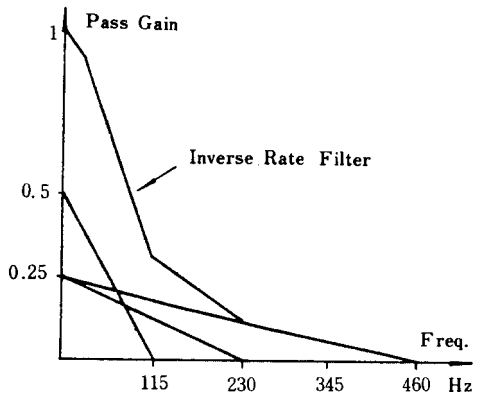


Fig.2-a Modelling LPF for the pitch extraction

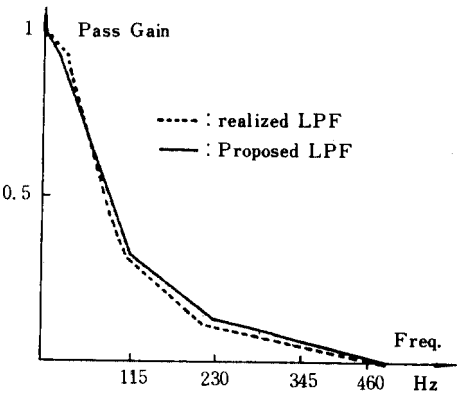


Fig.2-c Comparison the proposed LPF with the realized LPF

and

$$y(n) = g(n) * A(n)$$

$$= \sum_{i=-\infty}^{\infty} g(i) \left[\sum_{K=0}^{N-1} s(n-k-i) \right]$$

$$= \sum_{i=0}^{N-1} \sum_{K=0}^{N-1} s(n-k-i) \qquad (4)$$

or

$$y(n) = s(n) * g(n) * g(n) \longleftrightarrow Y(jwT)$$

$$= S(jwT) \{ sinc(jwT) \cdot sinc(jwT) \} \qquad (5)$$

where s(n) is the speech sample at time n. To utilize the linear part of the sinc square function, the 3/4pi-point of the variable of the $sin^2c(.)$ must be adjusted to the cutoff frequency of the filter. The combined characteristics of three filters with each different pi-point is shown in the Fig. 2-c with the ideal inverse rate type filter which is in the Fig. 2-a.

III. ALGORITHM OF THE IRT FITERING

To obtain the sinc square filtered signal at time n+1, we use Eq.(4). Figure 3 illustrates the operations implied by Eq.(4) for a value of n+1. that is

$$y(n-1)=\sum_{i=0}^{N-1}\sum_{k=0}^{N-1}s(n-k-i-1)$$

$$=\sum_{i=1}^{N}\sum_{k=0}^{N-1}s(n-k-i)$$

$$=\sum_{i=1}^{N}A(n-i)$$

$$=\sum_{i=0}^{N-1}A(n-i)-A(n)+A(n-N)$$

$$=y(n)-A(n)+A(n-N) \quad (6)$$

Here, the previously computed y(n) is the sinc square filtered signal at foregoing speech sample and the area of triangular window as continued curve in Fig.3. Also, A(n) is a content of temporary addition buffer as

$$A(n)=\sum_{k=-\infty}^{\infty}\delta(k)*s(n-k)=\sum_{k=0}^{N-1}s(n-k)$$

$$=\sum_{k=-1}^{N-2}s(n-k)-s(n+1)+s(n-N+1)$$

$$=A(n+1)-s(n+1)+s(n-N+1) \quad (7)$$

where the A(n+1) is already computed area by step function for the foregoing sample. If we add -s(n+1) and s(n-N+1) to A(n+1), it will give A(n). Similarly, the A(n-N) is

$$A(n-N)=A(n-N+1)-s(n-N+1)$$

$$+s(n-2N+1) \quad (8)$$

Thus, to calculate the filtered signal for a present speech sample, we calculate A(n) and A(n-N). Those values are subtracted and added respectively from and to the triangle area, y(n) which was calculated one sample earlier. Thus, the result for the computation procedure of a sinc squared filter is that we need only three additions and substractions each sample.

To realize the inverse rate type filter of Fig. 2-a, we have composed three parallel branches as shown in Fig.4 with N-value 13,26 and 52 each in expression (6). The pi-point of a sinc squared function which express the cutoff frequency (615, 307, and 153 Hz) of filter, will vary in accordance with the N-value (13, 26, and 52). The obtained gain according to the N-values is an area value (=N*N) of triangular window.

We have used an 8-bit A/D converter. When we have taken 52 as maximum N-value, the gain become 2704 ($=N*N < 2^{12}$). By consequence to obtain the sinc square characteristics, the integer operation of maximum 20(= 8+12) bits will be necessary. And it enables

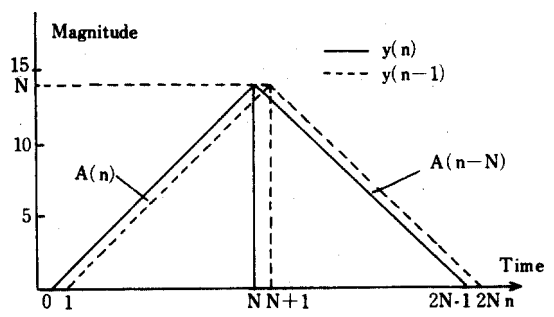


Fig. 3 Impulse Response for the $\sin^2 c(.)$ Spectrum.

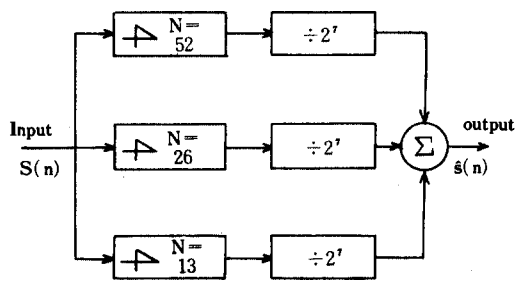


Fig. 4 Realization of the proposed LPF.

the calculation with Finite Word Length (FWL) calculation by an ordinary microcomputer. And before to combine the 20 bit values which are calculated in order to construct the IRT filted signal, we divide them by 2^7= 128 which will enable us to pursue the LPF calculation with 16 bit arithmetics.

Finally, we conclude that the processing for each speech sample is sufficient with ten additions and nine subtractions by means of the proposed method.

IV. EXPERIMENTATION AND RESULTS

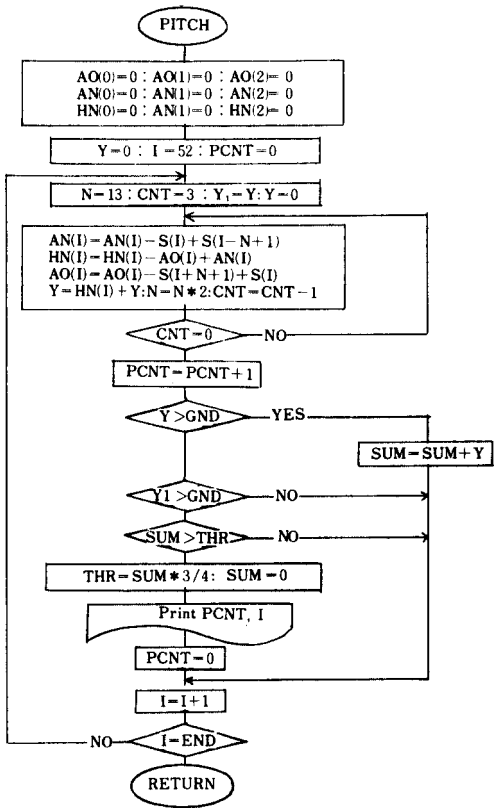
In order to extract the pitch in speech signals, we carried out the procedure of the flow-chart as fig.5. First, the Eq.(6) is carried out three times to process the IRT filtering for a speech sample (CNT=3 at the flow-chart), and then the results are combined (Y=HN(I)+Y at the flow-chart). As a result of the processing a speech signal with the proposed LPF, the glottal part of speech signal is emphasized. Inspecting it visually in time domain, we see that the first positive peak in a pitch period intensified.

Next, in order to detect the periodicity of voiced speech, we calculate the area of the positive peaks of the filtered samples (SUM= SUM+Y at the flow-chart), and locate the area values at the end of the positive peaks:i.e.,

$$P(n2) = \sum_{n=n1}^{n2} Y(n) \qquad (9)$$

here, the positive area, P(n2) exists at time domain for n1<n<n2, and Y(n) is a filtered signal at time n.

By searching the calculated areas for the positive peaks, the peaks with a dominant large value appears each pitch period in voiced speech.



AO(.) = step area buffers
AN(.) = step area buffers
HN(.) = triangular area buffers
Y1 = foregoing filtered value
Y = IRT filtered value
CNT = counter of sinc filter
PCNT = pitch period counter
SUM = area buffer of positive peak
THR = threshold level for detecting a pitch
I = address of speech sample stored in memory

Fig. 5 Flow-chart for pitch extraction using the IRT filtering.

Thus, if we pick out the large areas, the pitchs of voiced speech will be found. In order to detect the large areas, we chose a threshold as 3/4 of first positive area of the foregoing pitch (Thr= SUM*3/4 at flow-chart):i.e.,

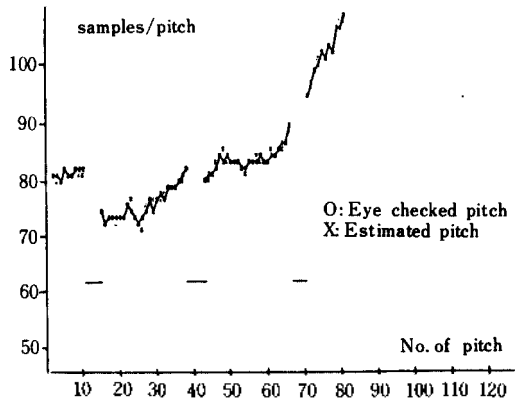


Fig.6-a Result for speech "insunekomada": 28-years sold man speaker.

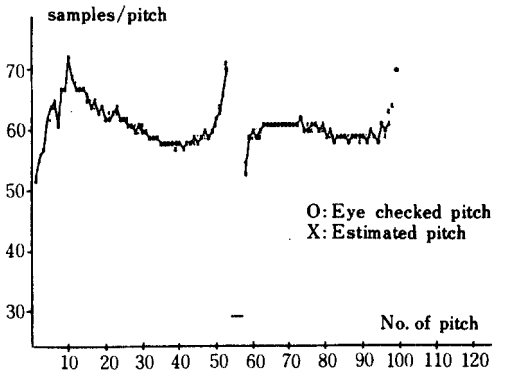


Fig.6-b Result for speech "annyunghaseyo": 25-years old man speaker.

$$Thr(n-1) = 3/4 \cdot SUM(n) \quad (10)$$

where the SUM(n)is a positive area value calculated in the foreoging pitch. Since an area differences between the first positive peak and the side peaks in a pitch interval exhibit experimentally a value more than 1/2, we decide the multipling constant for threshold as 3/4.

In consequence of this method, as the contours of the eye checked pitch and of estimated pitch are proven for a speech "insunekomada" and for a speech "annyunghaseyo" in Fig. 6.

V. CONCLUSION

We have proposed an inverse rate type filter which has the characteristics of inversely proportional gain to the frequency and which we can use in stead of ordinary sharp cutoff LPF when processing a voiced speech signal. Thus, the noxious major formants proximate to the pitch are attenuated to have almost the effect of the post pitch detection method without the complexity.

As this method is implement, in the time domain with three triangular shapes, a calculation of ten additions and nine subtractions to have the desired response for the sampled speech signal. This process can be treated with a finite word length integer arithmetics and requires only addition and subtraction which enables a real time process by an ordinary microprocessor.

As this method does the double process of LPF, filtering and major formants reduction, required process time is considerably short and the period calculation is simpler as the periodicity is ameriorated by the filtering: Both permit the real time pitch calculation with a microporcessor.

REFERENCES

1. L.R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", Prentice-Hall, Englewood Cliffs, New Jersey, 1978.
2. J.D. Markel and A.H. Gray, "Linear Prediction of Speech", Springer-Verlag, Berlin, 1976.

3. Myungjin BAE, "A Study on the Fundamental frequency Extraction of Speech Signals using Second Order Rundown Method", Seoul National University, MA Paper, Jan. 1983.
4. Myungjin BAE and Souguil Ann, "The High Speed Pitch Extraction of Speech Signals using the Area Comparison Method", KIEE, Vol. 22, No. 2, pp.101-105, feb. 1985.
5. Myungjin BAE and Souguil ANN, "The Voiced-Unvoiced-Silence Classification By Emphasized Spectrum of Speech Signals", JASK, Vol. 4, No. 1, pp.9-15, June, 1985.
6. Myungjin BAE and Souguil ANN, "Low Pass Filtering on the High Speed Pitch Extraction", KIEE, to be published, 1986.
7. A.D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation", IEEE Trans. on Audio and Electroacoustics, Vol. Au-20, No. 5, pp.367-377, December, 1972.
8. L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. Mc. Gonegal, "A Comparative Performance Study of Several Pitch Detection Algorithms", IEEE Trans, Vol. ASSP-24, No.5, pp.399-418, Oct. 1976.