An Effective Storage Method During A Sampling of Speech Signals

(음성신호를 표본화할 동안 효율적인 실시간 저장기법)

裵明振*, 李寅燮*, 安秀桔*

(Myungjin Bae, Inseop Lee and Souguil ANN)

要 約

실시간 처리용 프로세서를 갖지 않는 speech analyzer들은 신호를 처리하기 전에, 저장버퍼에 speech sample들을 저장시킬 필요가 있다. 여기서 우리는 음성신호가 analog to digital converter에 의해 sample들로 변환될 때, 그 버퍼를 효율적으로 사용하는 한 알고리즘을 제안하였다.

이 방법을 실시간으로 처리하기 위하여, 그 버퍼를 starting 버퍼와 remain 버퍼로 나누었다. 유성음을 찾을때 까지는 변환된 sample들을 순서적으로 starting버퍼에 저장시키고 나서, 그버퍼를 순환시킨다. 유성음이 찾아진 이후부터는 remain 버퍼에 sample들을 차례로 저장시킨다.

Abstract

It is necessary for the speech samples to be stored in memory buffer before speech analyzers without a real time processor process them. In this paper, we propose an algorithm that uses the buffer efficiently, when the analog speech signal is converted to the digital samples by the analog to digital converter.

In order to implement this method in real time, the buffer is divided into the starting buffer and the remaining buffer. Until a voiced speech is found, the converted samples are sequentially stored in the starting buffer, and then the buffer is shifted. When a voiced speech is found, the next samples are sequentially recorded in the remaining buffer.

1. Introduction

Recently, the development of the speech signal techniques made it possible for the machine to recognize the human speech and to be controlled by this. And the development of semiconductor techniques enabled the speech signals to be processed nearly in real time. But in the case of a system which is not really a real time system, the signal processing must be done with the storing speech sample in the memories. There are some types of method which store the input speech signal obtained through the A/D converter in memory buffer, such as "alarm method" and "tape record method".

The alarm method is one that gives the alarm at the starting and end part in storing the sampled speech. Because of the time difference between the alarm and the speaker's response to it, it is difficult to find precisely the starting point and the end point. Therefore, since the memory contains many unnecessary silence interval signals, this method is not economic in terms of memory usage. And it is difficult to get the natural sound from unskilled speaker

*正會員, 서울大學校 電子工學科

(Dept. of Elec. Eng., Seoul Nat'l Univ.)

接受日字 : 1986年 10月 16日

because the speaker strains himself to synchronize with the alarm. And it has the defect that we should repeatedly try to get the natural sound and find correct signals from them.

On the other hand, the tape recorder method is another one that we can find the alarm point and the end point by regenerating the recorded speech signals in the tape recorder, and then store the signals in the memory. This is a method which finds the points mechanically. Since the finding of exact points is not easy, economical use of memory is almost impossible.

In this paper, we have proposed the starting part finding method of words. For the purpose of real time processing of that, first we divide the memory buffer into the starting buffer and remain buffer. Until the voiced speech is found, the sampled speech is stored circularly in the starting buffer regions. After that, we store it in remain buffer till the end of the word.

In the end, for implementing the starting-partfinding-method, it is necessary in the starting buffer region, to precisely classify voiced-, unvoiced-, and silence in real time.

II. Finding the Speech Signal

Beginning parts of speech signal can be classified as silenced-unvoiced-voiced- and silence-voiced-. To find the starting point of speech signal, silence-unvoiced and voiced must be classified. Unvoiced is the type of colored noise which has high frequency, and lower energy than voiced speech. Thus, silence-unvoiced classification is difficult and unvoiced may or may not texist in the beginning part. But voiced speech has the resonant frequency according to each phoneme, and furthermore has more energy and semiperiodic characteristic. These properties make it easy to classify the voiced speech.

If we find the voiced speech which is easy to detect, we assume that the starting part of speech signals is found. We find the voiced speech by its semiperiodic characteristic. When stable pitches are found more than 3 times, successively we suppose that the speech signals are found.

Generally, voiced speech has lower frequency than unvoiced speech, so average zero crossing rates (Zav) in a frame (=10 msec) of the voiced and the unvoiced speech are very different. And voiced speech is excited by glottal wave which has a semiperiodic fundamental frequency and is resonated by vocal tract, thus it has larger energy than unvoiced speech and is periodic. Therefore, average energy (Eav) or magnitude (Mav) in a frame have a great difference between the voiced and the unvoiced speech. Up to now, as a way to find voiced speech in a frame, the Zav and the Eav (or Mav) have been used as basic parameters.

However, ZCR(Zero Crossing Rate) and the ENG (Energy) are very sensitive to the changes of ground position (=dc offset) for they are calculated based on ground position (=zero level) and the changes of sampling rate of A/D converter. [8] Moreover, they are generall dealt with by a unit of frame to make the difference between voiced and unvoiced speech wide. Since the calculating process is made in the end part of each frame, it is difficult to process in real time.

In this paper, to solve these problems, we use the area corresponding to zero crossing interval (ZCI) and semiperiodic property of voiced speech as basic parameters.

In a given speech waveform, s(t), the area of zero crossing interval in $n1 \leqq t \leq n2$ is

$$A(n2) = \sum_{t=n1}^{n2} |S(t)|$$

$$= \left(\frac{1}{n2-n1} \sum_{t=n1}^{n2} |S(t)|\right) \cdot (n2-n1)$$

$$= Mav / Zav \qquad (1)$$

From eq.(1), it is obvious that the area is proportional to average magnitude (Mav) and is inversely proportional to average zero crossing rate (Zav). Since the voiced speech has lower first formant frequency (about three times) than the unvoiced speech as shown in Fig 1, the 1/ZCR is large and since it has higher energy (about 10 times greater than) as shown in Fig 1, its Mav is large, too. The production of these two factors is the area value (about 30 times greater than) given in eq. (1). Therefore, if we use this value in the voiced-unvoiced speech classification, we can easily classify them by a unit of peak of the waveform, not by a frame unit. [4]

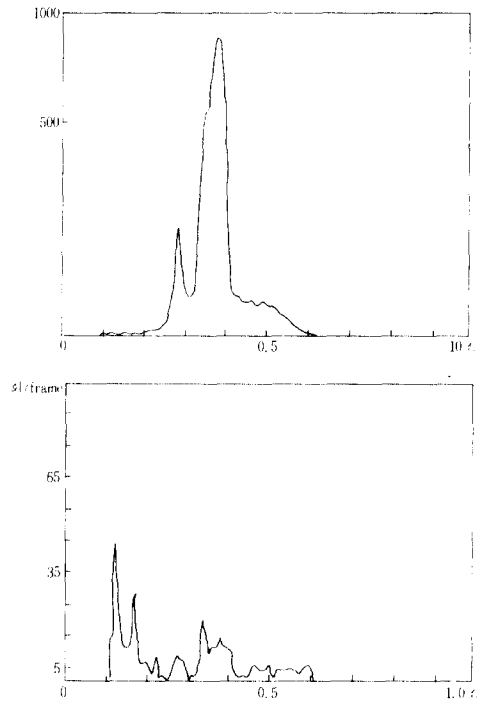


Fig. 1. Comparison Between the Average Zero Crossing Rate (Zcr) and the Average Magnitude (Mav) for a Speech Signal "SAM".

Considering the spectrum of the voiced speech, its effect is dominant because the first formant has the highest energy Then, approximately, the vocal tract impulse response can be modelled as follows. [4]

$$V(t) = g(B1,t)*\sin 2\pi F1 t \qquad (2)$$

where F1 is first formant frequency and B1 is bandwidth.

In this equation, the gain g (.) decrease as a function of B1 and the vocal tract impulse response V(t) is a decaying oscillation with frequency F1. [7]

And the excitation source of voiced speech is impulse which has a fundamental frequency F0 and is affected by glottal characteristic in going through the glottis as follows. [2]

$$\begin{aligned} G(t) &= \tfrac{1}{2}\{1-\cos(\frac{\pi t}{Ni})\}, \quad 0 \le t \le N_1 \\ &= \cos\{\pi(n-N_1)/2N_2\}, \quad N_1 \le t \le N_1+N_2 \\ &= 0, \quad \text{otherwise} \end{aligned} \qquad (3)$$

To generate actual speech signal, eq. (2) and eq. (3) should be convolved. In eq. (2), speech signal is generally modelled in relation with N1 > N2 and magnitude also increases as t increase in n0 < t < n1 interval. [2] The relation, 1/f1 < N1 results from relating eq. (2) to eq. (3). [8] Hence, after the two equations have been convolved, the ZCI (Zero Crossing Interval) of first positive peak in a pitch period is enlarged and more emphasized than others.

Since this positive peak is a first peak of eq. (2) which has damped oscillation in a pitch interval, its magnitude (MAG) is larger than others, ZCI is emphasized by eq. (3), and its 1/Zcr is greater than others. Thus, if we calculate the area per every peak by eq. (1), the first positive peak area in a pitch is distinguished clearly from others. [4]

This property is repeated everytime the glottal peak which is the excitation source is generated, and the emphasized positive peak appears in every period of pitch. In other words, if we extract only the first positive peaks, they represent the pitch period. [4]

To find a pitch, we put the threshold level based on emperical result as follows:

$$Pth = 3/4*A(i) \qquad (4)$$

In this equation, the A(i) means the first positive peak area value of the latest true pitch.

III. Necessity of the Starting Buffer

To store only a part of speech signals in memory buffer, the exact classification of the silence and the speech for each speech sample. must be done previously. Especially, the unvoiced speech and silence is too complex. Thus we use the finding of voiced speech in rather than unvoiced speech in finding the first part of speech signals . Because the voiced speech has more dominant semiperiodic characteristic, we think that voiced speech is found if we find the stable periodic pitch 3 times. However, if we save this part in memory the first parts of speech signals which may be the unvoiced speech can not be stored.

Hence, we can catch the beginning part by storing the speech signal circularly in a certain length buffer till as Fig. 1. the stable voiced speech is found as in Fig.2. Because this buffer

contains the part which is the starting of the speech signals, we call this "starting buffer". The length of starting buffer is assigned to a possible unvoiced speech for 48 msec, and a transient voiced speech for 48 msec, so total assignment is 96 msec. We got these results from various experiments and we think it is necessary and sufficient length to cover starting part of all speech sound. Sampling the speech signal at fs = 8kHz, memory size of starting buffer needs.

Ms=fs*96= 768 words

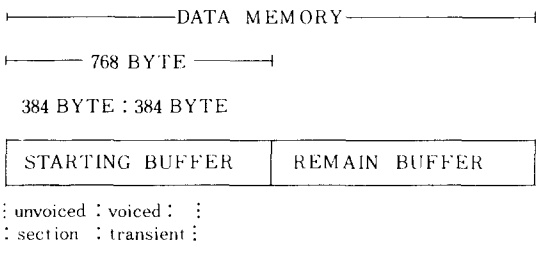


Fig. 2. Memory Assignment.

The support of hardware is needed to accomplish the shift of all starting buffer within 100 microseconds for each speech data input.

If we use micro-processor to implement this, we need about 20 clocks for shifting of 1 data. 5 microsec (20*250 nsec) per 1 data shift is needed, if we use a micro-processor which has 4MHz clock frequency (e.g. 8086 micro-processor). Thus it needs 3.8 msec (768*5 microsec) totally. It is impossible to do that in real time, because the operating time must be less than sampling rate. Hence it needs some other hardwares (as an example DMA) to shift the data. The alternative method that can be replaced is circulating the pointer that indicates the present address of input sample value till voiced speech is found. This method doesn't need hardware, but pointer must be rearranged to indicate the first address of buffer after inputing a word unit of speech data.

IV. Process of Starting Buffer

When evaluating the positive peak through eq. (1), we extract the pitch from that result. If the period of extracted pitch (number of samples) satisfies the following condition:

$$\text{the latest pitch}-5 < \text{extracted-pitch} < \text{the latest pitch}+5 \qquad (7)$$

and it is found sucessively 3 times, we can assume the voiced speech found. After that, the data are stored in the remaining buffer. The processing in the starting buffer is presented in Fig.3 as flow chart. In the starting buffer region, one cycle of starting buffer can be processed within 100 microseconds which is less than 8 kH sampling time (125 microseconds).

Unvoiced speech has a major formant in 2000-3000 Hz and its energy is lower than that of voiced speech with 3 formant and it has a characteristic similiarity to the colored noise. [7][8] If unvoiced speech is taken in 8 kHz sampling rate, the range of period that waveform takes up in time domain is (8000/2000 = 4) – (8000/3000 = 2). Therefore, when the zero crossing interval corresponds to that of unvoiced speech, it take up 2–4 samples and everytime the waveform with this pitch appears.

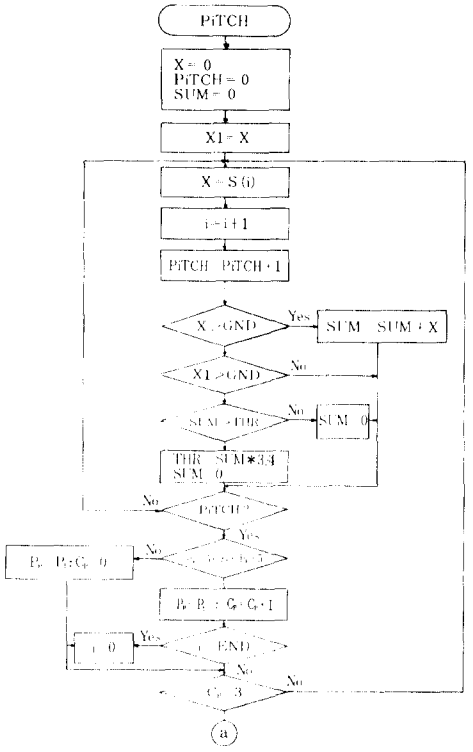
The voiced speech has the formants of high frequencies apart from the first formant, by which the pitch period error of 2 or 4 samples may happen. To remove this error, the preprocess filter is needed. And the filter can be realized by hardware or software. [6] However, in this paper, we can dispense with preprocess filter for the real time process by allowing the possible pitch error in ed. (7).

To reduce the effect of DC offset of an A/D converter, we can modify the variable nl and n2 of the interval of eq. (1). The modified interval is between the points that cross a certain level (offset value = d) above the zero crossing point of waveform as follows:

$$A(n2) = \sum_{t=n1}^{n2} (S(t)-d) \qquad (8)$$

This case corresponds to that if peak is below the offset value, the area is not calculated, with the area ratio of unvoiced and voiced sustained.

When the sampled speech is circularly stored in starting buffer, we think that the speech is found if we find 3 quasi-stable pitches in eq. (7). Hence, we store the rest of speech in remaining buffer.



X1 : Post Speech Sample
X : Present Speech Sample
SUM : Positive Area Buffer
C_P : Pitch Counter
P_P : Post Pitch Period
P_i : Present "
i : Starting Buffer Address

Fig. 3. Flow Chart.

V. Simulation and Result

In this paper, we used the 16 bit personal computer, IBM PC/XT for simulation. We used, as the speech signal data,

"인수네 꼬마는 천재 소년을 좋아한다."
;DATA1
"서울대 전자공학과 음성 신호 처리팀이다."
;DATA2
"예수님께서는 천지창조의 교훈을 말씀하셨다."
;DATA3
"역사에 관한 교훈 남자 여자"
;DATA4

for simulation.

The speech signal is obtained from 4 different speakers using 8 bit A/D converter. Each speech signal sampled at 8 kHz, has 3 second duration. We use the ACM (Area Comparison Method) for pitch extraction. If the period of the extracted pitch (number of samples) satisfies eq. (7) and if it is found 3 times successively, we assume the voiced speech is found and set the starting point.

To find the decision rule for stable pitch, we vary the deviation degree of pitch from ±2 samples to ±5 samples. Since it does not demand narrow deviation degree of pitch for stable pitch we decided that ±5 samples were suitable deviation degree.

In Fig. 4. we represent the starting point and starting buffer of DATA 1. Since starting part is the voiced ('1') in this data, it needs the classification of silence and voiced. We can see in Fig. 4. that the first part of starting buffer region is filled with silence signal and its length is about 60 msec. This part has no meaning in DATA 1 for it is the space to cover unvoiced speech. However, we can see in Fig. 5 that starting buffer region contains the unvoiced speech ('ㅅ') and the transient part of voiced speech ('ㅓ'). In Fig. 6 and Fig. 7, we represent the results of DATA 3 and DATA 4.

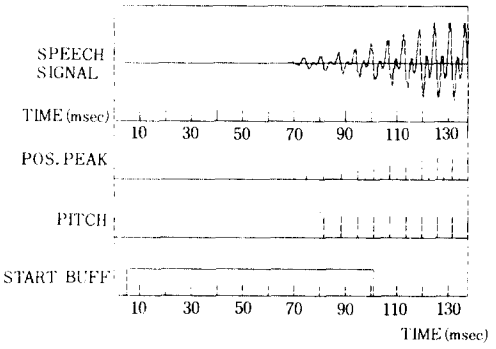


Fig. 4. DATA 1.

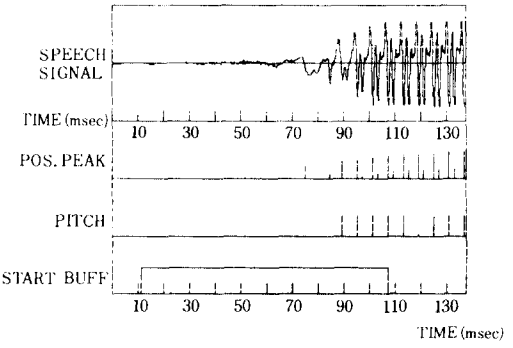


Fig. 5. DATA 2.

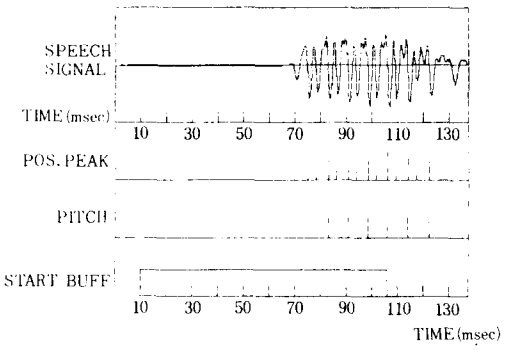


Fig. 6. DATA 3.

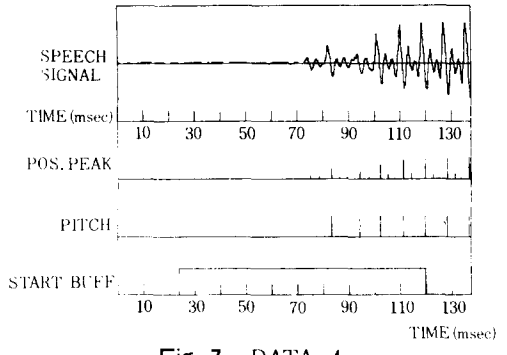


Fig. 7. DATA 4.

As a result, we can see it is performed efficiently to find the starting part of the speech signals and to use the starting buffer. Also, as in Table 1, we save the memory space as much as about 700–3000 samples.

Table 1. Silence Size Comparison

	ALARM METHOD	PROPOSED METHOD
DATA1	3715 SAMPLES	520 SAMPLES
DATA2	2031 SAMPLES	160 SAMPLES
DATA3	2951 SAMPLES	480 SAMPLES
DATA4	1132 SAMPLES	400 SAMPLES

V. Conclusion

In this paper, we have proposed the starting part finding method of the word which efficiently stores the speech signals in memory buffer. Alarm method and tape recorder method are not suitable for the efficient use of memory.

We divide the buffer into the starting buffer and the remain buffer to implement the proposed method in real time. The sampled data are circularly stored in the starting buffer until voiced speech, the rest data are circularly saved in the remain buffer.

When this method is used, the memory to store speech samples can be economically used, and although the inexperienced speaker tells an utterance, the utterance as natural sound can be stored. Also, it can be implemented by a general purpose microcomputers without any high speed numeric coprocessor.

If we can make decision of voiced– unvoiced –silence in real time, the realization of word boundary finding method is possible.

References

[1] B.S. Atal and L.R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition", *IEEE, Trans.*, Acoustic Speech Signal Processing, vol., ASSP–24, pp.201–212, June 1976.

[2] A.E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels", *J. A. S. A.*, vol.49, pp.538–590, 1971

[3] L.R. Rabiner and M.R. Sambur, "An Algorithm for determining the endpoints of isolated utterences", *B. S. T. J.*, vol.54, No.2, pp 297–315, Feb.

[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.13–17, March 1985

[5] Myungjin BAE and Songuil ANN, "The voiced–unvoiced–silence classification by emphasized spectrum of speech signals", *J. A. S. K.*, 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, 1987.

[7] J.D. Markel and A.H. Gray, "Linear prediction of speech", *Springer–Verlag*, 1976.

[8] L.R. Rabiner and R.W. Schafer, "Digital processing of speech signals", *Prentice–Hall*, Englewood Cliffs, New Jersey, 1978. *