잡음이 섞인 음성에서의 음성/무언의 구별

(Speech / Silence Discrimination of Noisy Speech)

* 은 종 관 (Un, C. K.)
** 김 현 수 (Kim, H. S.)

Abstract

In this paper an investigation for detecting the presence of speech in noisy signal corrupted by white Gaussian noise has been done. Speech/silence discrimination is made based on the energy and the autocorrelation sum of input signal. The threshold of energy is adapted by comparing with the energy and autocorrelation of the incoming noisy speech. Computer simulation has been done with clean speech and noisy speech with signal-to-noise ratios (SNR's) of 20, 10 and 0 dB. The percentage of discrimination errors was about 2 percents when the SNR was 20 dB.

요 약

본 논문에서는 음성이 백색 Gaussian 잡음에 섞여 있을때 음성의 유무를 구별하는 방법에 관해서 연구되었다. 제안된 방법은 음성과 무언이 입력신호의 에너지와 자기상관함수의 합에 의해서 구별된다. 에너지의 threshold치는 입력되는 잡음음성의 에너지와 자기상관함수를 비교함으로써 적응되도록 하였다. 이 방법을 시험하기 위해서 잡음이 없는 음성, 0, 10 및 20 dB의 잡음이 섞인 음성을 사용하여 computer simulation을 하였다. SNR이 20 dB 일때 구별의 오차율은 2 %로 나왔다.

I. INTRODUCTION

Discrimination of speech and silence is required in many areas of speech processing and coding, such as speech interpolation, vocoding and speech recognition.

It is well known that in a two-way telephone conversation, speech activities occur

* 과학기술원 교수
** 과학기술원

only about 40 percent of the time. Accordingly, the use of speech interpolation in long distance telephony can double channel capacity without increasing the facilities of the transmission medium. In a speech interpolation system, accurate detection of silence in conversational speech is essential for the speech interpolator to function properly.

So far, many silence detection algorithms have been proposed for use in the speech interpolation systems. Most of these algorithms have been devised assuming that the input speech is reasonably free from background noise and acoustic distortion. In practical situations, however, one cannot always expect clean and distortion-free speech. Accordingly, this study is addressed to the problem of detecting silence from noisy (and also clean) speech of which result can be used in a digital speech interpolator and other applications.

Among many silence detection algorithms, early silence detectors used signal power as a parameter for decision [1]. An improved version is the detector that uses the envelope signal as a parameter of speech/silence discrimination [2]. Also, the detector proposed by Rabiner and Sambur is based on zero-crossing rate and signal energy [3]. In a different approach, waveform quantizers were also used in silence/speech discrimination. Schafer and Jackson used adaptive differential pulse code modulation (ADPCM) for this purpose [4]. Un and Lee used bit alteration rate of the linear delta modulation (LDM) bit stream and the band-pass filtered output of decoded LDM signal [5].

Although those methods discussed above yield fairly accurate results for clean speech, their effectiveness diminishes when the input speech is noisy. For this reason we propose a new method which discriminates silence from speech based on the absolute sum of autocorrelation and the energy of input speech. This method has been found out to be highly effective for noisy as well as clean speech.

Following this introduction, the speech/silence discrimination algorithm proposed in this work is described in Section II. In Section III computer simulation results are given and discussed. In Section IV hardware design for implementation of the proposed algorithm is discussed. Finally, conclusions are made in Section V.

II. THE SILENCE/SPEECH DISCRIMINATION ALGORITHM

In discriminating silence from noisy speech most errors occur in unvoiced speech. The reason is that since unvoiced sound and background noise are similar in their characteristics, it is difficult to distinguish one from another. In general, autocorrelation of voiced speech is greater than that of noise, but autocorrelation of unvoiced speech is not always greater. Hence, the use of autocorrelation as a parameter in speech/silence discrimination would not be appropriate. Instead, it is proposed to use a modified form of autocorrelation, that is, the autocorrelation sum, which has been found out to be very effective for our purpose.

Let the input speech r(t) corrupted by noise be represented by

$$r(t) = s(t) + n(t) \qquad (1)$$

where s(t) and n(t) are clean speech and noise, respectively. The normalized p_{th} order autocorrelation of the m_{th} block noisy speech composed of N samples is given by

$$A_m(p) = [\sum_{i=1}^{N-p} r_m(i) r_m(i+p)]/E_m \qquad (2)$$

where E_m is the energy of the m_{th} block calculated as

$$E_m = \sum_{i=1}^{N} r_m^2(i). \qquad (3)$$

If we assume that there is no correlation between speech and noise, (2) may be written as

$$A_m(p) = A_{m,s}(p) + A_{m,n}(p) \qquad (4)$$

where

$$A_{m,s}(p) = \sum_{i=1}^{N-p} [s_m(i) s_m(i+p)]/E_m \qquad (5)$$

$$A_{m,n}(p) = \sum_{i=1}^{N-p} [n_m(i) n_m(i+p)]/E_m. \qquad (6)$$

One may note from (5) and (6) that as the noise level increases, the normalized p_{th}-order autocorrelation of noisy speech decreases but that of noise remains fairly constant. The p_{th}-order absolute sum of autocorrelations is given by

$$SUMA_m(p) = \sum_{i=1}^{p} | A_m(i) | \qquad (7)$$

In Fig. 1 typical ranges of autocorrelations absolute sums of noise and voiced and unvoiced speech corrupted by noise are shown for different orders of autocorrelation. It is seen that although the ranges of the three different signals overlap each other when the correlation order is low, the gap between speech (voiced and unvoiced) and noise becomes widened as the order increases. Hence, the autocorrelation sum would be an effective parameter in discriminating silence from noisy speech. In the proposed speech/silence discrimination method we use the autocorrelation sum and the energy of input signal as the decision parameters.

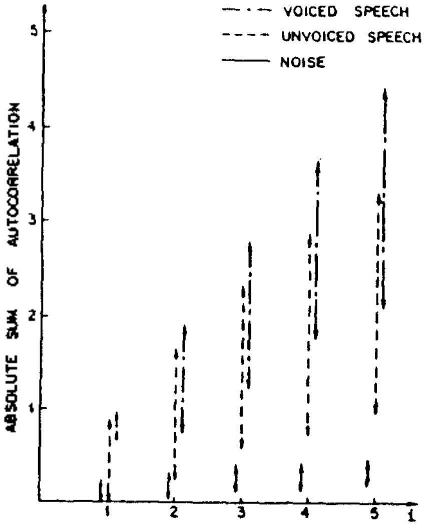

Fig. 1 Range of normalized absolute sum of autocorrelations of 1st to 5th order for noise, unvoiced and voiced sound.

In using the energy of input signal as a decision parameter, we use an adaptive threshold level for which the average noise level is utilized. Initially, the energy of the m_{th} block E_m is compared with a preset threshold value THCL. If we have $E_m \leq$ THCL, the m_{th} block is considered as a "noisy" silence block. The value of THCL may be found empirically by simulation. When L consecutive noisy silence blocks are found, the average noise level $\bar{E}$ is calculated by

$$\bar{E} = \frac{1}{L} \sum_{k=1}^{L} E_k \qquad (8)$$

where E_k is given by (3). Of course, $\bar{E}$ is zero if the input speech is clean. Thereafter, whenever a noisy silence block is detected, the average noise level $\bar{E}$ is updated as

$$\bar{\bar{E}} = \bar{E} \cdot (\alpha - 1)/\alpha + E_m/\alpha \qquad (9)$$

where E is the newly updated noise level and α is an energy adaptation factor. This average noise level will be used as an adaptive threshold level for the input signal energy. Once the autocorrelation sum and the average noise level are determined, the m_{th} block is decided to be as speech if the following conditions are satisfied.

$$\mathrm{SUMA}_n\,(P) \geq \mathrm{TH}, \text{and } E_m \geq K\bar{\bar{E}}.$$

where TH is a threshold value and K is a scale factor. Otherwise, it is considered to be silence.

III. SIMULATION RESULTS

The speech/silence detection algorithm described in the preceding section has been simulated on a computer, and its performance has been compared with the algorithm of Drago et al. that is based on envelope detection and input signal energy. The input speech was low-pass filtered at 3.4 kHz, sampled at 8 kHz, and quantized with 12 bit resolution. Following this preprocessing, speech samples were high-pass filtered at 100 Hz to remove any d.c. or low frequency noise. To generate noisy speech of varying degree (20, 10 and 0 dB), white Gaussian noise was added to clean speech. We used six sentences of male and female speech. The reference data that are to be compared with the algorithm-detected results were obtained using clean speech by the eye detection method. Speech/silence discrimination has been done on a block-by-block basis, each block having 128 samples or 16 ms long.

We have found that the accuracy of the discrimination algorithm depends strongly on the threshold level TH of the autocorrelation sum and the scale factor K of the average noise energy. The optimum values of TH and K that give least errors regardless of noise level were 1.0 and 1.7, respectively, when the autocorrelation order was 5. The energy adaptation factor α [see Eq. 9] used in our simulation was 10. As one can expect, the accuracy of discrimination improved (particularly for very noisy (0 dB) speech) as the order p of autocorrelation increased. In our algorithm we have used p=5.

Figures 2 through 5 show the results of discrimination for clean and noisy speech of 20, 10 and 0 dB, respectively. In the reference data both voiced (v) and unvoiced (uv) speech as well as silence (s) are shown. In Table 1 the percentages of errors of the proposed algorithm are shown for different noise levels. Also shown in this figure is the result obtained by the method of Drago et al. [2]. For the Drago's algorithm we optimized the threshold value at different noise levels, and used the hangover time of 4 ms. Comparing the performances of the two methods, one can see that the proposed method yields significantly less number of errors.

Finally it is worthwhile to note that although we have tried to remove additive noise by Wiener filtering before speech/

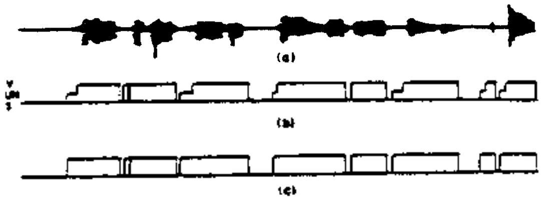

Fig. 2 Speech/silence discrimination of clean speech

a) Clean speech

b) Decision result by eye detection

c) Decision result by algorithm detection

* V: Voiced, UN: Unvoiced, S: silence

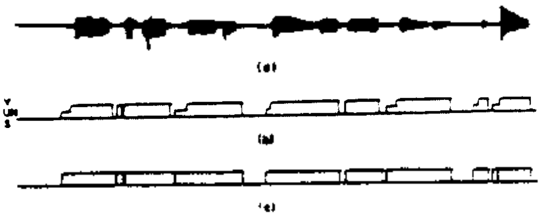

Fig. 3 Speech/silence discrimination of 20 noisy speech

a) 20 dB noisy speech

b) Reference data

c) Discrimination result with 5th-order autocorrelation (p=5)

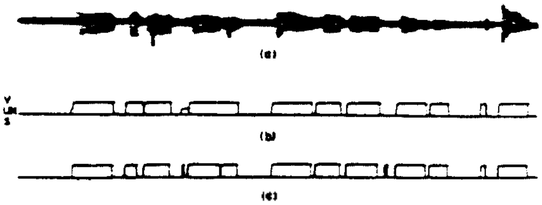

Fig. 4 Speech/silence discrimination of 10 dB noisy speech

a) 10 dB noisy speech

b) Reference data

c) Discrimination result with 5th-order autocorrelation (p=5)

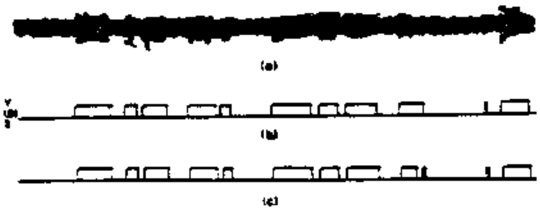

Fig. 5 Speech/silence discrimination of 0 dB noisy speech

a) 0 dB noisy speech

b) Reference data

c) Discrimination result with 5th-order autocorrelation (p=5)

Table 1. Percentage of discrimination errors of the proposed method and the method of Drago et al.

Method / Noise level	Error Rate (%)	
	Proposed Method	Method of Drago et al.
Clean	0	4.4
20 dB	2.1	5.6
10 dB	3.5	7.1
0 dB	5.3	10.5

silence discrimination, we were not successful in reducing errors. In fact, the prefiltering process increased errors in some portions of speech. A possible explanation might be that, contrary to the assumption that the additive noise is white, it is not in reality. This aspect needs for further study.

IV. CONSIDERATION OF HARDWARE DESIGN

We now consider the implementation of the proposed speech detector algorithm. A block diagram of the proposed speech detector is shown in Fig. 6. Speech/silence decision is done on a block-by-block basis, each block having 128 samples. The input samples are stored in an input buffer. Since the input signal must be delayed at least one block, two input buffers of 128 bytes are needed. The input and output of the two buffers A and B are controlled by a controller. If the buffer A is full, input samples are stored in the buffer B. Then, calculation of absolute sum of autocorrelation and energy of the samples in the buffer A is done. At the same time, the samples that are stored in the buffer B are transferred as the output signal. Speech/silence decision of a block is done by comparing the absolute sum of autocorrelation and the energy of input signal with the threshold values.

In the discriminator, calculation of autocorrelation and energy requires a major portion of processing time. The numbers of computations for calculation of energy and autocorrelation sum are 753 multiplications and additions per one block. If we use a

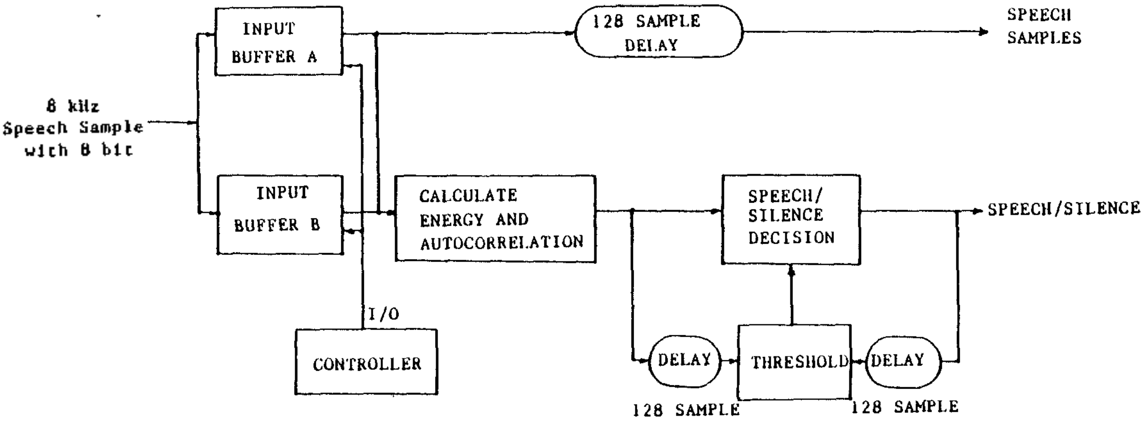

Fig. 6 Block diagram of the proposed speech detector

single-chip microprocessor such as 8751 for the detector, the computation time of energy and absolute sum of autocorrelations of input signal required would be less than 5.5 ms. Therefore, real time speech/silence discrimination is possible.

V. CONCLUSION

In this paper a speech detection algorithm that use the absolute sum of autocorrelation and the energy of input noisy speech have been presented. With this algorithm the silence portions that are longer than 16 ms can be detected accurately. When the input speech was clean, it yielded no decision error. When SNR's of input speech were 20, 10 and 0 dB, the percentages of errors were 2.07, 3.45 and 5.34, respectively.

REFERENCES

1. H. Miedema and M. C. Schachtman, "TASI quality-effect of speech detectors and interpolation," B.S.T.J., pp. 1455-1473, July 1962.
2. P. G. Drago, A. M. Molinari, and F. C. Vagliani, "Digital dynamic speech detectors," IEEE Trans. Commun., vol. COM-26, No. 1, Jan. 1978.
3. L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," B.S.T.J., pp. 297-315, Feb. 1975.
4. R. W. Schafer, K. Jackson, J. J. Dubnowski, and L. R. Rabiner, "Detecting the presence of speech using ADPCM coding," IEEE Trans. Commun, vol. COM-24, No. 5, pp. 563-567, May 1976.
5. C. K. Un and H. H. Lee, "Voiced/unvoiced/silence discrimination of speech by delta modulation," IEEE Trans. Acoustics, Speech and Signal Processing, vol. ASSP-28, No. 4, Aug. 1980.