Voice Activity Detection with Run-Ratio Parameter Derived from Runs Test Statistic

Kwang-Cheol Oh*

ABSTRACT

This paper describes a new parameter for voice activity detection which serves as a front-end part for automatic speech recognition systems. The new parameter called run-ratio is derived from the runs test statistic which is used in the statistical test for randomness of a given sequence. The run-ratio parameter has the property that the values of the parameter for the random sequence are about 1.

To apply the run-ratio parameter into the voice activity detection method, it is assumed that the samples of an inputted audio signal should be converted to binary sequences of positive and negative values. Then, the silence region in the audio signal can be regarded as random sequences so that their values of the run-ratio would be about 1. The run-ratio for the voiced region has far lower values than 1 and for fricative sounds higher values than 1. Therefore, the parameter can discriminate speech signals from the background sounds by using the newly derived run-ratio parameter.

The proposed voice activity detector outperformed the conventional energy-based detector in the sense of error mean and variance, small deviation from true speech boundaries, and low chance of missing real utterances

Keywords: Voice Activity Detection, Speech Recognition, Runs Test Statistic

1. Introduction

The application of speech recognition technology to a hand-held device requires real-time property as well as its recognition accuracy because of its limited resources. Since hand-held devices such as cellular phone, PDA, and smart phone are generally restricted in using memory and computational resources, the efficient usage of these resources is very important to success in deploying speech recognition technology. To efficiently use resources, it is required that the system is able to process only the speech-present region among the input signal.

Voice activity detection refers to a process which looks for the location of an utterance in an input audio signal prior to the main recognition process. The detector informs the recognizer of the boundary locations whenever it detects a speech signal.

* Human and Computer Interaction Lab., Samsung Advanced Institute of Technology

Then the speech recognition process can be exclusively performed on this specified audio segment. Therefore, portions of speech that are missed by the speech detector will never be exposed to the recognizer. On the other hand, portions of background noise that are exposed to the recognizer are very likely to induce recognition errors by misleading the hidden Markov model (HMM) used for recognition.

Many voice activity detection methods have been proposed with their own detection parameters The most widely used parameters are short-time energy and zero-crossing rate. The short-time energy parameter achieves good performance using signal-to-noise ratio[1]. Some detectors use zero-crossing ratio as their parameters for detecting fricative or consonant sounds[2]. However, they usually fail when the zero-crossing ratio parameter is employed in a noisy environment. A large number of parameters such as pitch[3], teager energy[4], and high order statistics[5] can be used successfully with or without the energy parameter. However, they require high computational load.

In order to design a good voice activity detection algorithm which can be used successfully in hand-held devices and with background noise, we should consider three points: low computational complexity, detection accuracy, and performance consistency. Considering these requirements, the energy parameter may be a strong candidate for the voice activity detection. It is simple, requires low computational load, and achieves admissible performance. However, the energy-based voice activity detector has the severe problem of choosing energy thresholds in a noisy environment. And this problem may cause failure in the requirement of performance consistency.

Now, I propose a new parameter which is called a run-ratio parameter to alleviate the problem of choosing energy thresholds and to reduce computational load. The run-ratio parameter is derived from the runs test statistic. This test statistic shows how much the sequence is random. The run-ratio parameter has consistent values in various background noise conditions. Therefore, it could reduce an effort choosing the appropriate thresholds. Furthermore, the amount of computation is less than in the energy parameter.

2. Runs Test Statistic

The new parameter proposed in this paper is derived from the non-parametric statistic runs test: a test statistic for randomness. As the name implies, the runs test investigates randomness for a sequence of events where each element in the sequence can assume one of two outcomes, success (S) or failure (F). A run is defined to be a maximal sub-sequence of same elements. A very small or very large number of runs in a sequence implies non-randomness. For small values of runs, the sequence seems to have sub-sequences with long consecutive same elements. On the other hand, for large

values the sequence seems to have sub-sequences with alternating S and F.

Let the number of runs in the sequence be R, and let us try to find out the probability distribution for $P(R = r)$. Suppose that the complete sequence contains n_1 S elements with y_1 runs and n_2 F elements with y_2 runs. Then given y_1, y_2 can be one of three values: y_1, $y_1 - 1$ or $y_1 + 1$.

Now, we can derive a probability, $p(y_1, y_2)$ by counting relative frequency. The total number of distinguishable arrangements of n_1 S elements and n_2 F elements is $\binom{n1+n2}{n1}$. The inverse of this number may be a probability of a sample point. And the number of ways of observing y_1 S runs and y_2 F runs is $\binom{n_1-1}{y_1-1}\binom{n_2-1}{y_2-1}$. Then, by multiplying this number by the probability per sample point, we can obtain the probability of exactly y_1 S runs and y_2 F runs:

$$p(y_1, y_2) = \frac{\binom{n_1-1}{y_1-1}\binom{n_2-1}{y_2-1}}{\binom{n_1+n_2}{n_1}} \tag{1}$$

The $P(R=r)$ equals the sum of $p(y_1,y_2)$ over all possible values of y_1 and y_2 such that $y_1 + y_2 = r$. If r is an even value, the y_1 and y_2 will be the same value, r/2, with either the S or F elements commencing the sequences. Consequently, $P(R=r) = 2p(r/2, r/2)$. On the other hand, if r is an odd number, the y_1 and y_2 have the values, $\lfloor r/2 \rfloor \pm$ and these points are mutually exclusive. Then,

$$p(R=r) = p\left(\left|\frac{r}{2}\right|+1, \left|\frac{r}{2}\right|-1\right) + p\left(\left|\frac{r}{2}\right|-1, \left|\frac{r}{2}\right|+1\right). \tag{2}$$

The probability distribution for R tends to be normal as n1 and n2 become large. This normality approximation is good when n1 and n2 are both greater than 10. Consequently, we can use the Z statistic in equation (3) as a large-sample test, where E(R) and V(R) are the expected value and variance of R, respectively. The rejection region for a two-trailed test with ■=0.05 is $|z| > 1.96$.

$$Z = \frac{R - E(R)}{\sqrt{V(R)}} \tag{3}$$

where,

$$E(R) = \frac{2n_1n_2}{n_1+n_2}+1, \qquad V(R) = \frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}. \tag{4}$$

The runs test can be applied to detecting non-randomness of a sequence of quantitative measurements over time. These sequences, known as time series, occur in many fields. Departures in randomness in a series, caused either by trends or periodicities, can be detected by examining the deviations of the time series measurements from their average. Negative and positive deviations could be denoted by S and F, respectively, and we could then test this time sequence of deviations for non-randomness.

3. Voice Activity Detection

The property of the runs test statistic discussed in the previous section is utilized in the voice activity detection problem. If the number of samples in a frame is n, then n = n1 + n2. And R may be some value in the range between 1 and n because R=1 when n1 or n2 is zero and R=n when n1=n2=n/2. For computational simplicity, it is assumed that n1 and n2 always have the same values. Then from equation (4), R would be distributed normally with following mean and variance,

$$E(R) = \frac{n}{2}+1, \qquad V(R) = \frac{n(n-2)}{4(n-1)}. \tag{5}$$

Now, the run-ratio parameter is defined as RR, equation (6), related with runs test statistic R.

$$RR = \frac{2(R-1)}{n}. \tag{6}$$

If n is sufficiently large, the run-ratio parameter has values between 0 to 2 because R has the values between 1 and n. And the statistic of the run-ratio parameter is normal with mean and variance below that which can be derived from equation (5),

$$E(RR) = \frac{2(E(R)-1)}{n} = 1, \qquad V(RR) = 4\frac{V(R)}{n^2} = \frac{(n-2)}{n(n-1)}. \tag{7}$$

The voice activity detection algorithm with the new parameter may operate with the block diagram shown in figure 1. The proposed algorithm composed of 5 blocks: pre-processing, whitening, parameter computing, decision, and control parts. The pre-processing part was designed to filter the input signal according to the background signal property. The new algorithm used three kinds of filters: $1 - z^{-1}$, $1 - 2z^{-1} - z^{-2}$, and 1. When the background signal is small, the first filter is selected. And when the signal is high and has heavy low-frequency components, the second filter is used. Finally, if it is white, the original signal is not processed.

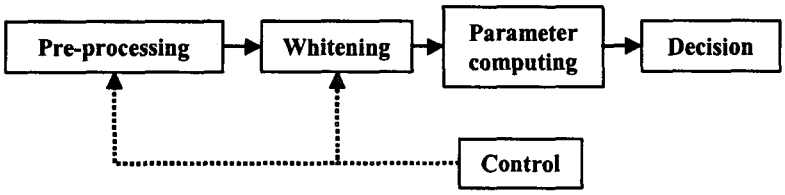

Figure 1. Structure of new speech detector.

The whitening part was adopted for strengthening the randomness of the background signal. This part was devised to reduce the effect of the characteristic of the background signal. If the background signal is not white, then the run-ratio parameter will stray from 1. Therefore, the run-ratio parameter values are required to be 1 for every kind of background signal. This procedure is regarded as the whitening process and implemented by simply adding white noise to the original signal.

Parameter computing and decision parts are similar to those in energy-based voice activity detector. The run-ratio parameter was calculated by the equation (6) for every 10 msec frame. The frames are not overlapped in computing run length for simplicity. The decision rule is simpler than ordinary energy-based voice activity detector because the new parameter does not depend on the amount of the noise power. Two thresholds are used. One threshold is for fricative signals and the other is for voiced sounds instead of low and high thresholds in energy based detector. Finally, the control part decides the pre-processing filter and the amount of added white noise.

4. Experiments and Discussion

Experiments were initially carried out to simulate the possibility of the run-ratio parameter. The run-ratio parameters for voiced speech signal and unvoiced fricative signal are calculated and summarized as histograms. The voiced sounds lead to low run-ratio values and their main peak is at near 0.2 as shown in Figure 2. And the parameter distribution for fricative sounds has a peak at 1.35. These values are different from those for the white random signal, which has normal distribution with mean 1 and variance 0.0123 according to equation (7) when n=80. Therefore, the parameter can easily distinguish voiced and fricative sounds with a background signal when the background is white.

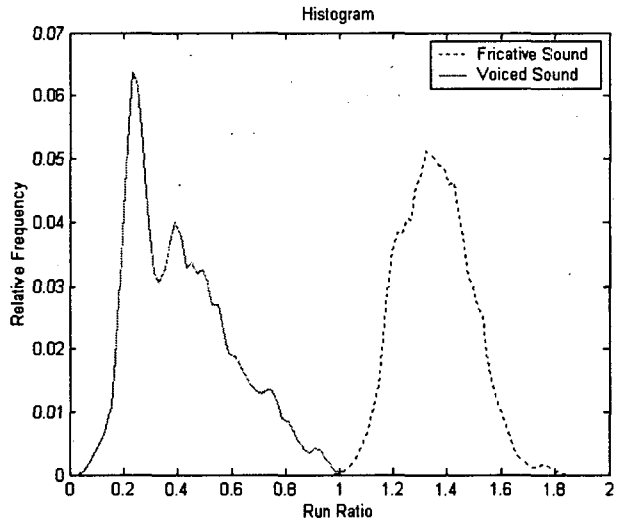

Figure 2. Histogram of run-ratio parameter.

Next, the run-ratio parameter was applied to the voice activity detection problem. For this detection experiment, 600 utterances were obtained using a hand-held device(PDA) which sampled input signal with 8 kHz sampling rate and 16 bit resolution. Twelve speakers, 6 male and 6 female, were involved in the recording process and each speaker spoke 50 isolated words. Each word was uttered somewhere within a 5-second session. For experiments in noisy environment, white and car noises from NOIEX-92 DB[6] were added to the recorded signal. The amount of the added noise was calculated according to the SNR level 20 and 10 dB.

In this voice activity detection experiment, the proposed detector with the new parameter is compared with energy-based method. Since the energy-based detector should update its thresholds depending on the background noise signal, the whole signal

of 5 seconds was used for the experiment. And the energy-based detection experiments are performed in two different ways: with or without a noise suppression process, which is a spectral subtraction method.

The errors were defined as frame differences between detected boundaries and hand labeled speech boundaries. In the voice activity detection experiments, the frame differences between true and detected boundaries are considered in performance comparison as well as the error mean and variance. Small differences may not affect speech recognition accuracy because most of HMM-based speech recognition systems have the silence model. However, large differences can influence recognition accuracy because the different quality of the background signal mismatches the silence HMM model.

The results of the voice activity detection experiments are summarized in Tables 1, 2, 3 where NS stands for noise suppression.. Table 1 shows the error mean and variance in unit of frame, a 10 msec duration. For the proposed method, the error mean and variance were much smaller than those of other two detectors. Many errors for the two energy-based methods in a clean environment seem to be caused by the lip noise at beginning and ending of the speech signal. And this phenomenon leads to reduced errors in 20 dB white noise because the added noise hides the lip noise. This result implies that our method has lip-noise immunity. The reason for this result is that the proposed parameter has different kinds of values for the fricative and voiced signals in comparison with the energy-based method.

Table 1. Error mean and variance in voice activity detection experiments.

		Run-ratio	Energy without NS	Energy with NS
Clean	mean	-0.2	1.7	2.0
	variance	2.0	8.5	6.6
White noise 20 dB	mean	1.2	1.7	1.0
	variance	3.4	7.3	6.2
White noise 10 dB	mean	3.2	14.6	2.2
	variance	4.8	49.6	5.8
Car noise 20 dB	mean	0.4	2.7	2.3
	variance	3.7	7.3	7.4
Car noise 10 dB	mean	3.2	4.3	2.5
	variance	5.5	12.4	7.2

The experimental results for the voice activity detection are further analyzed by investigating the number of errors within 5 frames, errors more than 5 frames and less than 10 frames. and errors more than 10 frames and less than 15 frames as shown in

Table 2. This analysis is meaningful because the small frame error can cause a small difference in recognition score with correct speech boundary. Most errors in the proposed method are within 5 frames although the detection capability of the energy-based method with the noise suppression is similar to that of the proposed method. Therefore, the run-ratio based detector can give more chance to correctly recognize detected speech signals with small errors in contrast to the energy-based detector with noise suppression.

Table 2. Error analysis in voice activity detection experiments.

		Run-ratio	Energy without NS	Energy with NS
Clean	$\|E\| \le 5$	14	78	72
	$5 < \|E\| \le 10$	1	62	44
	$10 < \|E\| \le 15$	1	52	39
White noise 20 dB	$\|E\| \le 5$	77	88	41
	$5 < \|E\| \le 10$	15	69	25
	$10 < \|E\| \le 15$	5	49	24
White noise 10 dB	$\|E\| \le 5$	109	276	98
	$5 < \|E\| \le 10$	20	247	38
	$10 < \|E\| \le 15$	4	209	28
Car noise 20 dB	$\|E\| \le 5$	39	53	68
	$5 < \|E\| \le 10$	19	39	45
	$10 < \|E\| \le 15$	18	32	38
Car noise 10 dB	$\|E\| \le 5$	88	81	78
	$5 < \|E\| \le 10$	38	68	51
	$10 < \|E\| \le 15$	16	61	43

Finally, Table 3 shows the missing errors which are defined as the number of cases when the detector does not detect speech in the input signal. The detector with the proposed parameter achieves small missing errors compared with energy-based detector. Therefore, the proposed algorithm hardly fails to catch the utterances in spite of heavy noise environment. This ability is slightly better than that of energy-based detector with noise suppression. Consequently, the speech recognition errors can be reduced by carefully re-investigating neighboring frames near the detected boundaries in heavy noise environment.

Table 3. Missing errors in voice activity detection experiments.

	Run-ratio	Energy without NS	Energy with NS
Clean	0	5	2
White noise 20 dB	0	12	2
White noise 10 dB	0	40	0
Car noise 20 dB	3	13	2
Car noise 10 dB	5	20	4

In short, the proposed voice activity detection method with the run-ratio parameter revealed several advantages over the energy-based method with noise suppression as well as without noise suppression. First, the error variance for the proposed method is much smaller than the counterpart. Second, the error distribution for the proposed method is concentrated on the small differences, within 5 frames, compared with the energy-based ones. For white background noise, the proposed method can detect almost all utterances.

Now, the computational complexity of the proposed method can be compared with the energy-based one without noise suppression. It is assumed here that the computational complexities of the two methods are equal except for parameter computation. The amount of computation for the proposed parameter is estimated from the equation (6). If the 8 kHz sampling is assumed, the number of computations can be summarized as shown in Table 4. The energy parameter is usually calculated on 30 msec duration in order to smooth its values. However, the number of computations for the 10 msec energy is also shown in the table for the purpose of the comparison with the proposed parameter. We can see the number of multiplications in energy computation is similar to the number of the comparisons in run-ratio computation. Therefore, as long as the computational complexity for comparison is much lower than that for multiplication, the proposed parameter can be obtained with much lower complexity.

Table 4. Comparison of the number of computation.

	Addition/Count	Multiplication	Comparison
Run-ratio	< 81	2	80
Energy (10msec)	80	80	-
Energy (30msec)	240	240	-

Finally, the run-ratio parameter seems to be similar to zero-crossing rate, but it is different in computing their parameters. The zero-crossing rate is usually computed by clipping some positive and negative values from the input signal. The proposed parameter is obtained from the signal which is veiled by the added white noise instead of

clipping it. Furthermore, the meaningful values of the run-ratio parameter can be obtained for the voiced sound region although the zero-crossing rates for that region are discarded.

5. Conclusion

This paper described a new parameter for the voice activity detection problem. The parameter was derived from the runs test statistic which tests randomness from a sequence with two elements. The runs test is modified to apply it to voice activity detection and the modified parameter is called run-ratio. The voice activity detection method with the run-ratio parameter has several advantages such as good error mean and variance, small deviation from the correct speech boundaries, and low chance of missing speech over the energy-based parameter.

The error variance for the proposed method is about half in comparison with the energy-based method with noise suppression in 20 dB SNR white noise environment. And 82% of the error for the proposed method is concentrated within 5 frames, compared with 60% for the energy-based ones.

The proposed method can be applied directly to speech recognition in a hand-held device because of its consistent detection accuracy and low computational burden. However, other kinds of noise environment except white and car noises have to be considered before practical application. Furthermore, a combined method with the energy parameter can be considered in the future study.

References

[1] Mauuary, L. & J. Monn. 1993. "Speech/non-speech detection for voice response systems." *Eurospeech '93*, Berlin, Germany, 1097-1100.

[2] Savoji, M. H. 1989. "A robust algorithm for accurate endpointing of speech." *Speech Communication*, 8, 45-60.

[3] Hanada, M., Y. Takizawa & T. Norimatsu. 1990. "A noise robust speech recognition." *Proc. ICASSP*, 893-896.

[4] Ying, G. S., C. D. Mitchell & L. H. Jamieson. 1993. "Endpoint decision of isolated utterances based on a modified teager energy measure." *Proc. ICASSP*, 732-735.

[5] Martin, A., L. Karray & A. Gilloire. 2000. "High order statistics for robust speech/non-speech detection." *Eusipco*, Tampere, Finland, 469-472.

[6] Varga, A., H. J. M. Steenneken, M. Tomilson & D. Jones. 1992. "The NOISEX-92 study on the effect of additive noise on automatic speech recognition." *Documentation on the NOISEX-92 CD-ROMs.*

Received: January 28, 2003
Accepted: March 4, 2003

▲ Kwang-Cheol Oh
Human and Computer Interaction Lab., Samsung Advanced Institute of Technology
San 14-1, Nongseo-ri, Giheung-eup, Yongin-si, Gyeonggi-do, 449-712, KOREA
Tel: +82-31-280-1714 Fax: +82-31-280-9257
E-mail: okcheol@orgio.net