Voice Activity Detection Based on Short-Time Energy and Noise Spectrum Adaptation

Dong Enqing[1, 2] (enqdong@mailst.xjtu.edu.cn) Liu Guizhong[2] (Liugz@xjtu.edu.cn)
Zhou Yatong[2] (zytong@mail.com3c.xjtu.edu.cn) Cai Yu[2] (Ramanujancn@yahoo.com.cn)
(1. Department of Communication & Electronic Engineering, Soochow University, Su Zhou 215021, China)
(2. School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China)

Abstract

On the basis of the short-time energy of speech signal and the efficient method of noise statistics adaptation estimation proposed by Sohn et al., a new high robust voice activity detection (VAD) rule for any kinds of environmental noise is proposed in this paper. The accurate recognition rate of the new method is about five percent higher than that of Sohn's method on average, and also has the same merit of tracking the noise spectrum properly as Sohn's method. A great deal of simulation experiments shows that the new method is an efficient and robust voice activity detector.

1. Introduction

In a two-way telephone conversation, one party is active for only about 35 percent of the time[1]. This can be exploited effectively for the reduction of the average bit-rate and co-channel interference in a digital cellular system[2], and also reduces transmitter power consumption in portable equipment. VAD is required in some speech communication applications such as speech recognition, speech coding, hands-free telephony and echo cancellation[3]. For this purpose, various types of VAD algorithms that trade off delay, sensitivity, accuracy and computational cost have been proposed. The earlier algorithms are based on the Itakura LPC distance measure, energy levels, pitch, and zero crossing rates, cepstral feature, adaptive noise modeling of voice signals, periodicity measure etc. Well known examples of VADs include those employed by QCELP speech coders in the IS-95 standard, the EVRC in the IS-127 standard, the VAD adopted for the discontinuous transmission(DTX) model of the GSM standard[4], and Recommendation G.729 Annex B[5] and G.723.1 Annex A in ITU-T[6].

A robust voice activity decision rule proposed by Sohn[7][8], which decision rule and noise statistic estimation algorithm can be optimized respectively by applying a statistical model, is derived from the generalized likelihood ratio test by assuming that the noise statistics are known a prior. The decision rule is very efficient. The advantage of the algorithm is that a novel noise spectrum adaptation algorithm for estimating the time-varying noise statistics can allow for the occasional presence of the speech signal.

It is well known that the short-time energy is an effective and simple classifying characteristic parameter, so many VADs have been designed based on the parameter. For example, the VAD algorithm employed by QCELP speech coder in the IS-95 standard. Since the time-vary noise statistic estimation is very important for VAD in special environment noise, only if the energy level parameter is adopted in a VAD design, the background noise can not be effectively estimated, the designed VAD will be limited in high environment noise. So a new high robust voice activity detection rule for any kinds of environment noise is proposed based on the short-time energy and the adaptation estimation of background noise statistics proposed by Sohn in the paper. A great deal of simulation experiments shows that the new algorithm has about five percent correct decision rate higher than that of Sohn's method on average.

2. Short-Time Energy

Speech is a time-varying and nonstationary signal, but, in a short segment, for example 10~20 ms, the speech signal is nearly stationary. So speech signal can be split many short segments to be processed. Assuming that $s(n)$ is the time series of the input speech, at first, it is split into many frames $f_i(iL+n)$, which length is L, the subscript i describes the serial number of the frame. Each frame is processed by a rectangular window function $w_R(n)$, then the windowed frame is $f_{wi}(n)$, i.e.

$$f_{wi}(n)=f_i(iL+n)\cdot w_R(n), \quad 0\le n\le L-1 \qquad (1)$$

where the rectangular window function $w_R(n)$ is

$$w_R(n)=\begin{cases}1 & (0\le n\le L-1)\\ 0 & (other)\end{cases} \qquad (2)$$

0-7803-7488-6/02/$17.00 © 2002 IEEE.

So that the short-time energy for the ith frame can be defined as following:

$$E_i = \sum_{n=0}^{L-1} f_{wi}^2(n) \tag{3}$$

3. Voice Activity Decision Rule of Sohn Method

The decision rule of a VAD can be formulated by a decision statistic. We use the statistical model in which the speech and noise signals are Gaussian random processes that are independent of each other, then the discrete Fourier transform(DFT) coefficients of each process are asymptotically independent Gaussian random variables. The L dimensional coefficient vectors of speech, noise, and noisy speech are denoted as $\mathbf{S}, \mathbf{N}$, and $\mathbf{X}$, with their kth elements S_k, N_k, X_k, respectively. In this statistical model, the variances of N_k and S_k are given by.

$$\lambda_N(k) = S_N(2\pi k / L) \tag{4}$$

$$\lambda_S(k) = S_S(2\pi k / L) \tag{5}$$

where $S_N(\omega)$ and $S_S(\omega)$ denote the true power spectra of noise and speech, respectively. The variance of X_k is given by:

$$\sigma_X^2(k) = \lambda_N(k) + \lambda_S(k) \tag{6}$$

The noise statistics $\lambda_N(k)$ are assumed to be known a priori. Then, the two hypotheses of the voice activity detection problem are as follows:

H_0: speech absent: $\mathbf{X} = \mathbf{N}$

H_1: speech present: $\mathbf{X} = \mathbf{N} + \mathbf{S}$

where H_1 is a composite hypothesis with a set of L unknown parameters, $\Theta = \{\lambda_S(k) : k = 0, \cdots, L-1\}$. The joint probability density functions conditioned on H_0 and on H_1 and Θ are given by:

$$p(\mathbf{X} \mid H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi \lambda_N(k)} \exp\left\{ -\frac{|X_k|^2}{\lambda_N(k)} \right\} \tag{7}$$

$$p(\mathbf{X} \mid \Theta, H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \exp\left\{ -\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)} \right\} \tag{8}$$

The generalized LRT replace Θ with its maximum likelihood estimate, $\hat{\Theta} = \{\hat{\lambda}_S(k) : k = 0, \cdots, L-1\}$, where $\hat{\lambda}_S(k)s$ are obtained by the power subtraction method, i.e.,

$$\hat{\lambda}_S(k) = |X_k|^2 - \lambda_N(k) \tag{9}$$

and the corresponding decision rule using the log likelihood ratio is obtained by substituting Eq. (9) into Eq. (8) as follows:

$$\Lambda_g = \frac{1}{L} \log \frac{p(\mathbf{X} \mid \hat{\Theta}, H_1)}{p(\mathbf{X} \mid H_0)} = \frac{1}{L} \sum_{k=0}^{L-1} \left\{ \frac{|X_k|^2}{\lambda_N(k)} - \log \frac{|X_k|^2}{\lambda_N(k)} - 1 \right\} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{10}$$

4. Spectrum Adaptation Estimation of Background Noise

The optimal estimate of the variance of the background noise Fourier expansion coefficients $\lambda_N(k)$ in terms of the minimum mean square error is given by:

$$\begin{aligned} \hat{\lambda}_N(k) &= E(\lambda_N(k) \mid X_k) \\ &= E(\lambda_N(k) \mid H_0) P(H_0 \mid X_k) + E(\lambda_N(k) \mid H_1) P(H_1 \mid X_k) \end{aligned} \tag{11}$$

Using Bayes rule:

$$P(H_0 \mid X_k) = \frac{p(X_k \mid H_0) P(H_0)}{p(X_k \mid H_0) P(H_0) + p(X_k \mid H_1) P(H_1)} = \frac{1}{1 + \varepsilon \Lambda(k)} \tag{12}$$

where $\varepsilon = P(H_1)/P(H_0)$ and $\Lambda(k) = p(X_k \mid H_1) / p(X_k \mid H_0)$. Similarly, the following holds:

$$P(H_1 \mid X_k) = \frac{\varepsilon \Lambda(k)}{1 + \varepsilon \Lambda(k)} \tag{13}$$

Substituting Eq. (12) and Eq. (13) into Eq. (11) yields:

$$E(\lambda_N(k) \mid X_k) = \frac{1}{1 + \varepsilon \Lambda(k)} E(\lambda_N(k) \mid H_0) + \frac{\varepsilon \Lambda(k)}{1 + \varepsilon \Lambda(k)} E(\lambda_N(k) \mid H_1) \tag{14}$$

$\lambda_N^{(m)}(k)$ denotes $\lambda_N(k)$ at the mth frame. To obtain a feasible estimator of $\lambda_N(k)$ rather than Eq. (14), we use the current frame measurement $|X_k^{(m)}|$ instead of $E(\lambda_N^{(m)}(k) \mid H_0)$ when speech is absent. When speech is present, for the observed information $|X_k^{(m)}|$ not to be reflected on $\hat{\lambda}_N^{(m)}(k)$, we replace $E(\lambda_N^{(m)}(k) \mid H_1)$ by the estimate of the previous frame, $\hat{\lambda}_N^{(m-1)}(k)$, and a recursive formula for $\hat{\lambda}_N(k)$ is obtained as follows:

$$\hat{\lambda}_N^{(m)}(k) = \frac{1}{1 + \varepsilon \Lambda^{(m)}(k)} |X_k^{(m)}|^2 + \frac{\varepsilon \Lambda^{(m)}(k)}{1 + \varepsilon \Lambda^{(m)}(k)} \hat{\lambda}_N^{(m-1)}(k) \tag{15}$$

As we have no estimate of the speech parameter set Θ, it seems reasonable to use the generalized likelihood ratio instead of $\Lambda^{(m)}(k)$ in Eq. (15), which is defined as:

$$\Lambda_g^{(m)}(k) = \frac{p(X_k^{(m)} \mid \hat{\lambda}_S^{(m)}(k), H_1)}{p(X_k^{(m)} \mid H_0)} \tag{16}$$

Since the decision is not made for each frequency band k, but made once by observing all the frequency bands, we replace the $\Lambda_g(k)s$ with their geometric mean Λ_g in Eq. (10), as follows:

$$\hat{\lambda}_N^{(m)}(k) = \frac{1}{1 + \varepsilon \Lambda_g^{(m)}} |X_k^{(m)}|^2 + \frac{\varepsilon \Lambda_g^{(m)}}{1 + \varepsilon \Lambda_g^{(m)}} \hat{\lambda}_N^{(m-1)}(k) \tag{17}$$

If $\Lambda^{(m)}$ were fixed for frame index m, Eq. (17) would be an estimator of the time-varying spectrum with an exponentially weighted averaging.

5. Decision Rule via Short-Time Energy and Adaptation Spectrum Estimation of Background Noise

The short-time energy is an effective and simple classifying characteristic parameter in a VAD design. For improving the robustness of distinguishing speech from background noise, a new method of combining the short-time energy with the adaptation estimation of background noise statistics $\hat{\lambda}_N^{(m)}(k)$ from Eq. (17) is proposed to design a sensitive, robust decision rule for any kinds of environment noise. The design of the new decision formula is illumined by the SNR formula, so that the formula of the new decision rule that is similar to SNR formula is given by:

$$\xi(i) = 10\log 10(E_i / N_i)\begin{cases} H_1 \\ \gtrless \; \mu \\ H_0 \end{cases} \tag{18}$$

where $N_i = \sum_{k=1}^{L} \hat{\lambda}_N^{(m)}(k)^2$, E_i is the energy using Eq. (3). $\xi(i)$ denotes the characteristic parameter for the ith frame decision, μ denotes a specified threshold, the threshold can be updated adaptively according to the decision results of the current region.

6. Simulation Experimental Results

Before making VAD decision, the speech signal has been normalized, so the waveform amplitudes of the speech signal mentioned in the following experimental results are not the factual levels. The frame length is 10 ms in the experiment, i.e., there are 80 samples for each frame. To verify the robustness of the decision rule designed by combining the short-time energy with the noise statistics adaptation estimation for any kinds of environmental noise, comparison between Sohn method and the new method is adopted to validate the performance.

Sohn method does not directly give the results, neither does the proposed new method after each frame decision was performed. To obtain less delay for the new VAD method, the correct result of the current frame using the contextual correction is given after the second frame behind the current frame was obtained. So the delay is only 20 ms.

Fig. 1(a) is the waveform curve of speech "China" without noise in English, in the experiment, the performance comparison between the two methods is performed when the speech is mixed with many kinds of higher background Gaussian white noise. It is noticeable that the phonetic symbol [n] of the word is obviously lower energy than that of other ones, and it is easy to be submerged by a higher background noise. A Gaussian white noise with −15dB segment SNR is mixed. It can be seen from the Fig. 1(b) that the [n] pronunciation is undetected by Sohn method, however, the pronunciation region can be detected accurately by the new method in Fig.1 (c).

To verify the robustness of the two methods for mobile environment noise, many real car noises are adopted in the experiment. The Fig. 2(a) is the waveform curve of the clear Chinese speech "Na-Mi". These car noises are mixed with the clear speech (see Fig. 2(b)). The Fig. 2(c) is the decision result of Sohn method, the car noises in 750 ms and 1600ms are falsely detected as speech. Fig. 2 (d) is the decision result of the new method, there are no these false decisions.

In addition, the speech with time-varying babble noise is utilized to compare the performance between the two methods. The experiments validate that the new method has higher robust than Sohn method in various degrees. To quantificationally demonstrate the difference of the correct decision rate between the two methods, comparisons and tests are performed between the two methods by using a great deal of real speech data in many kinds of environment noise levels, correct decision rate comparison curves are obtained (Fig.3). The correct decision rate is the ratio of the number of correct decision frames to the total number of frames. As a whole, the correct decision rate of the new method is about 5 percent higher than that of Sohn method. Although only 5 percent proportion, but it is very significant in practical wireless cellular communication.

7. Conclusions

A great deal of simulation experiments indicates that the new proposed VAD decision rule designed by using short-time energy and noise statistics adaptation estimation is more efficient than Sohn method. The accurate recognition rate of the new method is about five percent higher than that of Sohn method on average, and also has the same merit of tracking the noise spectrum properly as Sohn's method, especially for time-varying noise such as babble noise.

References

[1] Brady P.T., "A technique for investigating on-off patterns of speech," Bell. Syst. Tech. J., 44:1-22, 1965.

[2] Gersho A. and Paksoy E., "An Overview of Variable Rate Speech Coding for Cellular Networks," IEEE Conf. Selected on Topics Wireless Commun. p.172-175, Vancouver, June 1992.

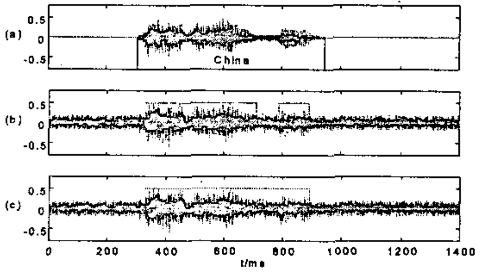

Fig. 1. Comparison chart of VAD with Gaussian white noise

(a) speech signal without background noise

(b) decision results of Sohn method

(c) decision results of the new method

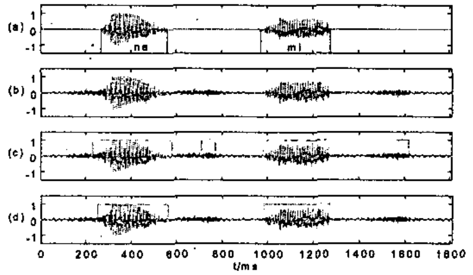

Fig. 2. Comparison chart of VAD with car noise

(a) speech signal without noise (b) speech with car noise

(c) decision results of Sohn method

(d) decision results of the new method

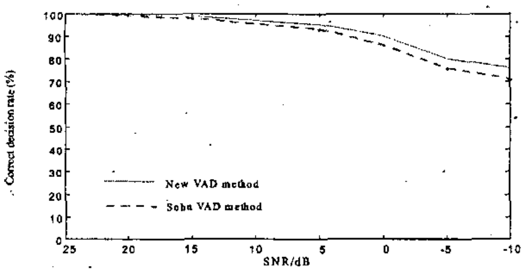

Fig. 3. Comparison curve of correct decision rate

[3] Tanrikulu O., Baykal B., Constantinides A. G., and Chambers J. A., "Residual echo signal in critically sampled subband acoustic echo cancellers based on IIR and FIR filter banks," *IEEE Trans. Signal Proccessing*, 45(4):901–912, 1997.

[4] Theodore S. R., Wireless Communications Principles and Practice. Prentice-Hall, 1996.

[5] Benyassine A., Shlomot E., Su H. Y., Massaloux D., Lamblin C., and Petit J. P., "ITU Recommendation G729 annex B: A silence compression scheme for use with G729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35(9): 64–73, 1997.

[6] ITU-T Rec. G723.1, Annex B, "Dual rate speech coder for multimedia communications transmitting at 5.3 and 6.3 kbit/s, alternative specification based on floating point arithmetic", 1996.

[7] Sohn J. and Sung W., " A Voice Activity Detector Emplying Soft Decision Based Noise Spectrum Adaption," IEEE International Conference on Acoustics, Speech and Signal Processing, vol. (1):365-368., May 12-15, 1998.

[8] Sohn J., Nam Soo Kim and Sung W., "A Statistical Model_Based Voice Activity Detection," IEEE Signal Processing Letters, vol. 6 (1):1-3, 1999.