2006 IEEE International Conference on
Systems, Man, and Cybernetics
October 8-11, 2006, Taipei, Taiwan

Application of GM(1,1) Model to Voice Activity Detection

Cheng-Hsiung Hsieh and Ting-Yu Feng

Abstract—**In this paper, a novel approach to apply GM(1,1) model in voice activity detection (VAD) is presented. The approach is termed as grey VAD (GVAD). In GVAD, the GM(1,1) model is used to estimate non-stationary noise in noisy speech and therefore signal component where an additive signal model is assumed. By estimated noise and signal, the signal-to-noise ratio (SNR) is calculated. Based on an adaptive threshold, the speech and non-speech segments are determined. The proposed GVAD is performed in the time domain and thus has less computational complexity than those frequency domain approaches. Through simulation, the GVAD is verified by cases with non-stationary noise. The result indicates that the proposed GVAD is able to detect voice activity appropriately.**

I. Introduction

VOICE activity detection (VAD) is a scheme to classify a speech signal into two segments: speech and non-speech. The VAD has been used in many applications such as speech coding and wireless speech communications for better bit rate utilization, bandwidth efficiency, and battery saving. In most of VAD approaches, an additive signal model is assumed where a noisy speech results from a sum of clean speech and additive noise.

Several approaches to the VAD problem have been reported. In [1-3], the likelihood ratio test scheme is proposed, where the input speech is transformed by fast Fourier transformation (FFT). For each frequency component the variance of additive noise is estimated by a recursive formula derived from conditional expectation. Then a composite hypothesis test is employed as a decision rule for the proposed VAD. In [4-5], the approach of subband order statistics filters is presented. In the approach, noise and signal are estimated separately in frequency domain through subband order statistic filters. Then SNR is obtained and speech/non-speech segment is determined by a given threshold. In [6], the voice activity detector is based on Kullback-Leibler divergence measure. In [7], a speech segment is classified through long-term spectral divergence. In [8], a radial basis function neural network is applied to VAD while a genetic programming is used in [9]. Note that all approaches mentioned above are performed in frequency domain except in [8].

In this paper, an approach to VAD problem based on grey model, GM(1,1) model, is proposed. For details, one may consult [10-11]. The approach is performed in the time domain and requires no statistical model as in [1-3]. This paper is organized as follows: Section II describes the signal/noise estimation based on GM(1,1) model. Next, the application of grey signal/noise estimation to VAD problem is given in Section III. Then examples are provided to justify the proposed grey VAD in Section IV. Finally, conclusion is made in Section V.

Cheng-Hsiung Hsieh is with the Department of Computer Science and Information Engineering, Chaoyang University of Technology, Wufong, Taiwan 41349, ROC. (e-mail: chhsieh@cyut.edu.tw).

Ting-Yu Feng is a graduate student in the Department of Computer Science and Information Engineering, Chaoyang University of Technology, Wufong, Taiwan 41349, ROC.

II. Signal/Noise Estimation Based on GM(1,1) Model

In this section, a signal/noise estimation approach based on GM(1,1) model is described. Section A gives a brief review of GM(1,1) model. Then the signal/noise estimation based on GM(1,1) model is described in Section B.

A. Review of GM(1,1) model

The GM(1,1) modeling process is briefly described in the following. Given data sequence $\{x(k) > 0, \text{for} 1 \le k \le K\}$, a new data sequence $x^{(1)}(k)$ is found by 1-AGO (first-order accumulated generating operation) as

$$x^{(1)}(k) = \sum_{n=1}^{k} x(n) \tag{1}$$

for $1 \le k \le K$, where $x^{(1)}(1) = x(1)$. From (1), it is obvious that the original data $x(k)$ can be easily recovered from $x^{(1)}(k)$ as

$$x(k) = x^{(1)}(k) - x^{(1)}(k-1) \tag{2}$$

for $2 \le k \le K$. This operation is called 1-IAGO (first-order inverse accumulated generating operation).

By sequences $x(k)$ and $x^{(1)}(k)$, a grey difference equation is formed as

$$x(k) + az^{(1)}(k) = b \tag{3}$$

where

$$z^{(1)}(k) = 0.5[x^{(1)}(k) + x^{(1)}(k-1)] \tag{4}$$

for $2 \le k \le K$, and parameters a and b are called developing coefficient and grey input, respectively.

From (3), parameters a and b can be obtained as

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\boldsymbol{B}^T \boldsymbol{B})^{-1} \boldsymbol{B}^T \boldsymbol{y} \tag{5}$$

where

1-4244-0100-3/06/$20.00 ©2006 IEEE

$$\boldsymbol{B}=\begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(K) & 1 \end{bmatrix} \quad (6)$$

and

$$\boldsymbol{y}=\begin{bmatrix} x(2) \\ x(3) \\ \vdots \\ x(K) \end{bmatrix} \quad (7)$$

It can be shown that the solution of $x^{(1)}(k)$ is given as

$$x^{(1)}(k)=[x(1)-\frac{b}{a}]e^{-a(k-1)}+\frac{b}{a} \quad (8)$$

where parameters a and b are found in (5). By 1-IAGO, the estimate of $x(k)$, $\hat{x}(k)$, is obtained as

$$\hat{x}(k)=x^{(1)}(k)-x^{(1)}(k-1) \quad (9)$$

where $\hat{x}(1)=x^{(1)}(1)=x(1)$. The estimation error for $x(k)$ is given as

$$e(k)=x(k)-\hat{x}(k) \quad (10)$$

which will be used to estimate additive noise in the GVAD.

B. Grey signal/noise estimation

The proposed signal/noise estimation approach based on GM(1,1) model is described here. The approach is termed as the grey signal/noise estimation (GSNE) hereafter. Assume the available noisy signal $x(k)$ has the additive signal model $x(k)=s(k)+n(k)$ where $s(k)$ and $n(k)$ are the clean signal and the additive noise in $x(k)$, respectively. Denote the ith segment of noisy signal as $\{x_i(k), \text{for } 1\le k\le L\}$ where $L=1+N_{ss}(K-1)$ is the total number of samples. Notation K is the number of samples used in GM(1,1) modeling and $N_{ss}=\lfloor L/(K-1)\rfloor$ is the number of subsets with one sample overlapped. The proposed GSNE is implemented by the following steps.

Step 1. Divide $\{x_i(k), \text{for } 1\le k\le L\}$ into N_{ss} subsets of K samples as $\{x_{ij}(k), \text{for } 1\le j\le N_{ss}, 1\le k\le K\}$. The way to divide $x_i(k)$ into subsets for the case $K=4$ is depicted in Figure 1 where the square indicates the sample overlapped.

Step 2. For subset j, find the estimation error of GM(1,1) model as $e_{ij}(k)=x_{ij}(k)-\hat{x}_{ij}(k)$ where $\hat{x}_{ij}(k)$ is obtained from (10).

Step 3. Note $e_{ij}(k)\neq n_{ij}(k)$ but related to $n_{ij}(k)$. The additive noise $n_{ij}(k)$ is estimated as $\hat{n}_{ij}(k)=\alpha e_{ij}(k)$ where $\alpha>0$ is a user-defined scaling parameter and is determined by experiences.

Step 4. Concatenate all $\hat{n}_{ij}(k)$ for $1\le j\le N_{ss}, 1\le k\le K$, to form $\hat{n}_i(k)$ for $1\le k\le L$ where $\hat{n}(1)=0$.

Step 5. Estimate mean μ of additive noise $n(k)$ as

$$\hat{\mu}=\frac{1}{N_{ss}(K-1)}\sum_{i=1}^{N_{ss}}\sum_{k=2+(i-1)(K-1)}^{1+i(K-1)}\hat{n}_i(k) \quad (11)$$

Since $x_i(k)$ is of one sample overlapped, thus only $\hat{n}(1)=0$ is excluded in (11).

Step 6. Estimate standard deviation σ of $n(k)$ as

$$\hat{\sigma}=[\frac{1}{N_{ss}(K-1)}\sum_{i=1}^{N_{ss}}\sum_{k=2+(i-1)(K-1)}^{1+i(K-1)}(\hat{n}_i(k)-\hat{\mu})^2]^{1/2} \quad (12)$$

III. Application of GSNE to VAD

The application of GSNE to VAD is described here. The proposed VAD based on GSNE is called grey VAD (GVAD). Assume that the additive signal model is appropriate for the noisy speech $x(k)$, that is, $x(k)=s(k)+n(k)$ where $s(k)$ denotes the clean speech and $n(k)$ as additive noise. The clean speech signal is assumed in the wave file format whose range is within $(-1,1)$. The implementation steps of GVAD are described as follows:

Step 1. Shift up the level of $x(k)$ by a positive constant C, $x(k)\leftarrow x(k)+C$, such that $x(k)>0$.

Step 2. Divide $x(k)$ into M non-overlapped segments of length L and denote $x_i(k)=s_i(k)+n_i(k)$ as the ith speech segment of length L. Without loss of generality, the length of $x(k)$ is assumed a multiple of L.

Step 3. Estimate additive noise $n_i(k)$ as $\hat{n}_i(k)$ based on GSNE where $\hat{n}_i(1)=\hat{n}_i(2)$.

Step 4. Estimate the signal $s_i(k)$ as $\hat{s}_i(k)=x_i(k)-\hat{n}_i(k)$ and $\hat{s}_i(1)=\hat{s}_i(2)$.

Step 5. Estimate the segmental standard deviation of $\hat{n}_i(k)$, $\sigma_n(i)$ as in (12), where index i denotes the ith segment of $x(k)$.

Step 6. Estimate the segmental standard deviation of $\hat{s}_i(k)$, $\sigma_s(i)$, by (12) similarly.

Step 7. Calculate the ith segmental SNR

$$SNR(i)=10\log\frac{\sigma_s^2(i)}{\sigma_n^2(i)} \quad (13)$$

Step 8. Determine if $x_i(k)$ is a speech or non-speech segment as follows. Given threshold $\eta(i)$ for the ith speech segment, mark $x_i(k)$ as a speech segment if $SNR(i)\ge\eta(i)-\beta\sigma_n(i)$ and a non-speech segment otherwise, where $\eta(i)=\left|\log\sigma_n^2(i)\right|$ and β is a scaling factor.

Step 9. Shift down the level of $x_i(k)$,

$x_i(k) \leftarrow x_i(k) - C$.

Step 10. Continue Steps 3 to 10 until all M segmented speech are processed.

The flowchart of GVAD is depicted in Figure 2.

IV. Simulation Results

In this section, the proposed GVAD is justified. The examples used in the simulation are speech files b.wav, f0125s.wav, and f0126s.wav in [12] which are, respectively, male speech 'b', and female oral reading sentences: "We were away a year ago." and "Should we chase those cowboys?" For more details, one may consult in the Appendix 4 of [12]. These speech files are considered as clean speeches whose sampling rate is 10 KHz. The additive non-stationary noise is generated artificially. In the simulation, all speech files are level-shifted by 5, i.e., $C = 5$ and the segment length is set to 240. The number of samples used in GM(1,1) modeling is 4, i.e., $K = 4$ where the parameter $\alpha = 1.7$ is employed. The parameter β in the GVAD is set to 7.5. The simulation results are given in Figures 3 to 5. In the figures, the clean speeches, additive noise, estimated noise by GSNE, and the noisy speech with GVAD output are shown from the top subplot down to the bottom. The SNRs for cases b.wav, f0125s.wav, f0126s.wav, are 2.89dB, 7.66dB, 3.75dB, respectively. As shown in Figures 3 to 5, the proposed GVAD is able to distinguish speech and non-speech segments appropriately for cases with non-stationary noise.

V. Conclusion

In this paper, a novel grey VAD based on GM(1,1) model is proposed. The GM(1,1) model is applied to estimate signal/noise. The estimation is called as GSNE. Based on GSNE, a novel scheme to VAD problem is presented and is termed as GVAD where an adaptive threshold is employed. The proposed GVAD is performed in the time domain and thus no transformation is needed. Moreover, statistical assumption is not needed in the GVAD. Consequently, the proposed approach has advantages of computational complexity over those in [1-9] and no statistical assumption is required when compared with [1-3]. Examples are given to verify the GVAD. The results show that the proposed GVAD approach works well in the examples.

References

[1] J. Sohn, N. S. Kim, and W. Sung, "A Statistical Model-Based Voice Activity Detection," *IEEE Signal Processing Letters*, Vol. 6, No. 1, pp. 1-3, January 1999.

[2] J.-H. Chang and N. S. Kim, "Voice Activity Detection Based on Complex Laplacian Model," *Electronics Letters*, Vol. 39, No. 7, pp. 632-634, April 2003.

[3] J.-H. Chang and N. S. Kim, "Speech Enhancement: New Approaches to Soft Decision," *IEICE Trans. on Information and Systems*, Vol. E84-D, No. 9, pp. 1231, September 2001.

[4] Harold Gene Longbotham and Alan Conrad Bovik, "Theory of Order Statistic Filter and Their Relationship to Linear FIR Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 2, pp. 275-287, Feb. 1989.

[5] J. Ramirez, J.C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "An Effective Subband OSF-Based VAD with Noise Reduction for Robust Speech Recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 13, No. 6, pp. 1119-1129, Nov. 2005.

[6] Javier Ramirez, Jose C. Segura, Carmen Benitez, Angel de la Torre, and Antonio J. Rubio, "A New Kullback-Leibler VAD for Speech Recognition in Noise," *IEEE Signal Processing Letters*, Vol. 11, No. 2, pp. 266-269, Feb. 2004.

[7] Javier Ramirez, Jose C. Segura, Carmen Benitez, Angel de la Torre, "Effective Voice Activity Detection Algorithms Using Long-Term Speech Information, " *Speech Communication*, Vol. 42, pp. 271-287, 2004.

[8] K.-I. Kim and S.-K. Park, "Voice Activity Detection Algorithm Using Radial Basis Function Network," *Electronic Letters*, Vol. 40, No. 22, pp. 1454-1455, Oct. 2004.

[9] P.A. Estevez, N. Becerra-Yoma, N. Boric and J.A. Ramirez, "Genetic Programming-Based Voice Activity Detection," *Electronic Letters*, Vol. 41, No. 20, Sep. 2005.

[10] J.L. Deng, "Control Problems of Grey System," *System and Control Letters*, pp. 288-294, 1982.

[11] J. Deng, "Introduction to Grey System Theory," *Journal of Grey System*, Vol. 1, pp. 1-24, 1989.

[12] D.G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, Inc., 1999.

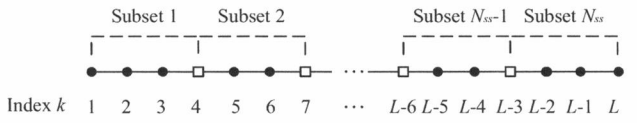

Fig. 1 One sample overlapped subsets for GSNE (K=4)

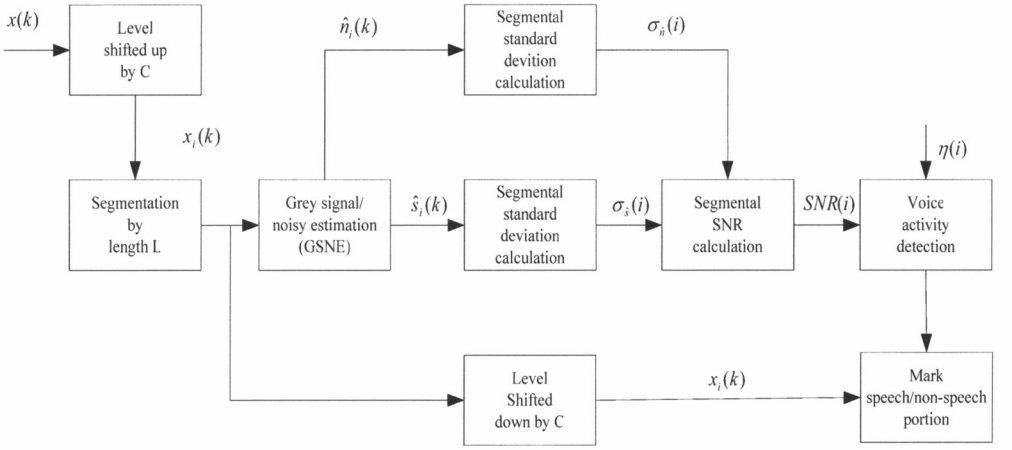

Fig. 2 The flowchart of the proposed GVAD algorithm

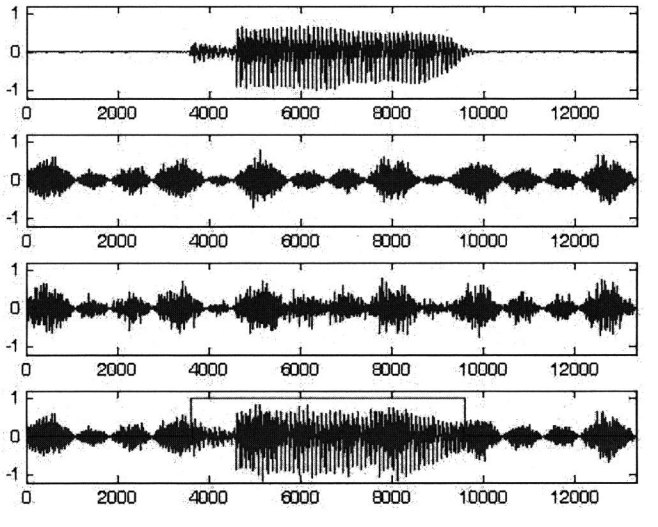

Fig. 3 The clean b.wav, additive noise, estimated noise, and noisy b.wav with GVAD output

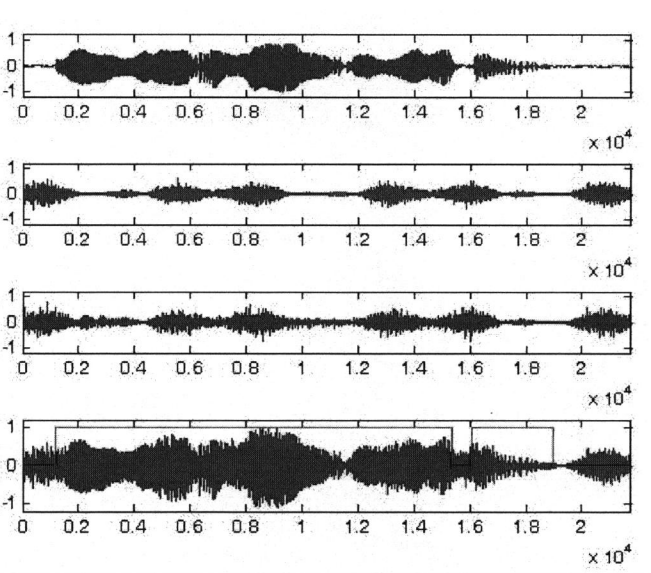

Fig. 4 The clean f0125s.wav, additive noise, estimated noise, and noisy f0125s.wav with GVAD output

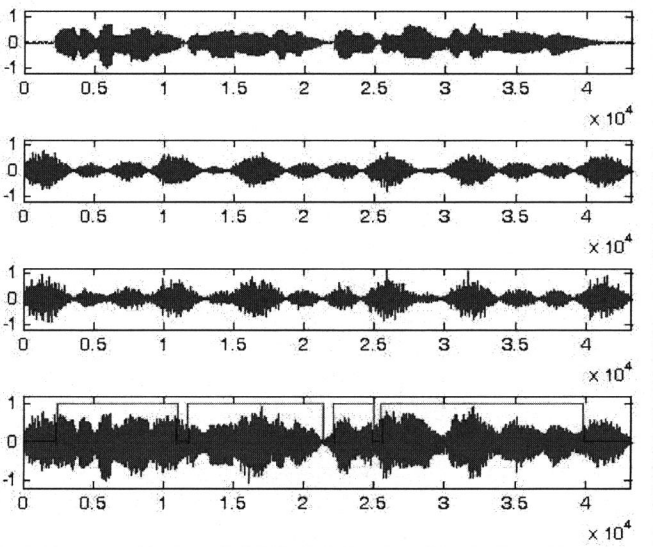

Fig. 5 The clean f0126s.wav, additive noise, estimated noise, and noisy f0126s.wav with GVAD output