Voice Activity Detection Based on GM(1,1) Model

Cheng-Hsiung Hsieh Ting-Yu Feng and Ren-Hsien Huang
Department of Computer Science and Information Engineering
Chaoyang University of Technology
Wufong, Taiwan 413, ROC
E-mail: chhsieh@cyut.edu.tw

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, GM(1,1) model is used to estimate noise in noisy speech and therefore signal where the additive signal model is assumed. By estimated noise and signal, the signal-to-noise ratio (SNR) is calculated. Based on an adaptive threshold, speech and non-speech segments are determined. The proposed GVAD is performed in the time-domain and thus has low computational complexity. In the simulation, GVAD is verified by cases with non-stationary additive white Gaussian noise and is compared with VAD in G.729 and GSM AMR. The results indicate that the proposed GVAD is able to detect voice activity appropriately. In the given examples, the performance of GVAD is better than VAD in G.729 and GSM AMR.

Index Terms—Voice activity detection (VAD), grey model, GM(1,1) model, signal/noise estimation, G.729, GSM AMR

1 Introduction

Voice activity detection (VAD) is a scheme to classify a speech signal into speech and non-speech segments. 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, the 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], a likelihood ratio test scheme is proposed, where the input speech is transformed by fast Fourier transform. 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 a 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]. However, it takes time in training generally.

In this paper, an approach to VAD problem based on grey model, GM(1,1) model, is proposed. 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 described in Section III. Then examples are provided to justify the proposed grey VAD in Section IV where comparisons are made with VAD in G.729 [10] and GSM AMR [11]. Finally, conclusion and further research is described in Section V.

2 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 2.1 gives a brief review of GM(1,1) model. Then the signal/noise estimation based on GM(1,1) model is described in Section 2.2.

2.1 Review of GM(1,1) model

The GM(1,1) modeling process is briefly described in the following. For details, one may consult [12-13]. Given data sequence $\{x(k) > 0, \text{for } 1 \le k \le K\}$, a new data sequence

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.

$x^{(1)}(k)$ 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 $x(k)$ 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

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

where

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

for $2 \le k \le K$, 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

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

and

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

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

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

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

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

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

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

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

2.2 Grey signal/noise estimation

The signal/noise estimation approach based on GM(1,1) model is described here. The approach is called 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 *i*th 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, by (9), as $e_{ij}(k) = x_{ij}(k) - \hat{x}_{ij}(k)$ where $\hat{x}_{ij}(k)$ is the estimate of $x_{ij}(k)$.

Step 3. Note $e_{ij}(k) \ne n_{ij}(k)$ 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 factor 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 $\hat{n}_i(1) = \hat{n}_i(2)$ is assumed.

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

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

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

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

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.

3 Application of GSNE to VAD

In this section, the motivation of GM(1,1) model based voice activity detection (VAD) which is called grey VAD (GVAD) is described. Then the application of GSNE to VAD is described and detail implementation steps of GVAD are given.

3.1 Motivation

The proposed GVAD is motivated by the following three observations. First, for a noisy speech it can be considered as two components: non-speech and speech segments. A non-speech segment is thought as constant signal with additive noise while a speech segment as random exponential signal with additive noise. Second, GSNE based on GM(1,1) model is able to estimate additive Gaussian noise accurately both for constant and random exponential signals as shown previously. It implies that the estimation error of GM(1,1) model can be related to additive noise whose statistics can be estimated appropriately with an appropriate scaling factor α. Third, in the additive signal model statistics of signal can be estimated accurately as well if noise estimation is appropriate. Consequently estimation of signal-to-noise ratio (SNR) may be good through estimated noise and signal. If this is true, an estimated SNR should be a good indication for speech and non-speech segment. Thus, the objective of VAD is achieved.

3.2 The proposed GVAD

Assume that the additive signal model is appropriate for 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 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 overlapped segments of length L and denote $x_i(k) = s_i(k) + n_i(k)$ as the ith speech segment.

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

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

Step 5. By (12), estimate the segmental standard deviation of $\hat{n}_i(k)$, $\sigma_{\hat{n}}(i)$, where index i denotes the ith segment of $x(k)$.

Step 6. Similarly, by (12) estimate the segmental standard deviation of $\hat{s}_i(k)$, $\sigma_{\hat{s}}(i)$.

Step 7. Calculate the ith segmental SNR as

$$SNR(i) = 10 \log \frac{\sigma_{\hat{s}}^2(i)}{\sigma_{\hat{n}}^2(i)} \qquad (13)$$

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 $SNR(i) \geq \eta(i) = \left|\log_{10} \sigma_{\hat{n}}^2(i)\right| - \beta\sigma_{\hat{n}}(i)$ and a non-speech segment otherwise, where 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 9 until all M segmented speech are processed.

The flowchart of GVAD is depicted in Figure 2. At least three advantages can be found in GVAD. First, the computational complexity is low since GVAD is a time-domain approach with simple mathematical operations. Though in the GM(1,1) modeling, matrix inversion is required in (5). However, it is cheap in computation when $K = 4$ which is the case for GVAD. Second, no recalculation of noise estimation is required for overlapped segments. This is generally not the case for VAD approaches in frequency-domain. Third, an adaptive threshold is used for estimation of a wide range of noise. Thus, the GVAD is able to apply in non-stationary additive noise cases. This will be justified in the following section.

4 Simulation Results and Discussion

In this section, the proposed GVAD is justified. The examples used in the simulation are speech files b.wav, and f0125s.wav in [14] which are, respectively, male speech 'b' and female oral reading "We were away a year ago." For more details, one may consult in the Appendix 4 of [14]. These speech files are considered as clean speeches. The non-stationary additive white Gaussian noise (AWGN) is artificially generated. In the simulation, all speech files are level-shifted by 5, i.e., $C = 5$ and the segment length L is set to 240. For adjacent segments, 160 samples are overlapped. This is same as G.729. The number of samples used in GM(1,1) modeling is 4, i.e., $K = 4$. The parameter $\alpha = 1.7$ is employed in GSNE and the parameter β in GVAD is set to 7.5.

First, GVAD is verified through the examples. The

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.

simulation results are shown in Figures 3 and 4 where subplots, from top to bottom, depict the clean speech, additive noise, estimated noise, and noisy speech with GVAD output, respectively. The SNR in Figures 3 to 5 are 2.89 dB and 7.66 dB, respectively. Obviously, the VAD outputs given in Figures 3 and 4 show that the proposed GVAD is able to distinguish speech and non-speech segments appropriately. In GVAD, SNR is used as a measure to discriminate speech and non-speech segments. Thus, the performance of GVAD heavily depends on appropriate estimation of additive noise. Since additive noise can be estimated accurately by GSNE, it results in the promising performance in GVAD.

Second, GVAD is compared with VAD in G.729 and GSM ARM. The segment length in GSM AMR is 160 samples. In the simulation, SNR is 5 dB for all cases. Simulation results are depicted in Figures 5 and 6. In Figures 5 and 6, the clean speech, noisy speech, segmental SNR with adaptive threshold, and VAD outputs from GVAD, G.729, GSM AMR from the top subplot down to the bottom one. The simulation results suggest that the output of GVAD is more accurate than those in G.729 and GSM AMR. It seems that VAD both in G.729 and GSM AMR suffer from transient state in the beginning, especially for the case of f0125s.wav. The proposed GVAD is free from it because the signal/noise estimation in GSNE is on segment-by-segment basis. In other words, the estimation in the current speech segment has nothing to do with other speech segments.

To sum up, the proposed GVAD, as shown from Figures 3 to 6, is able to distinguish speech and non-speech segments appropriately. In the given examples, GVAD performs better than VAD in G.729 and GSM ARM with much less computational complexity.

5 Conclusion and Further Research

In this paper, a novel VAD based on GM(1,1) model is proposed where GM(1,1) model is applied to estimate signal/noise. The estimation is called GSNE. Based on GSNE, an approach 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 computational complexity in GSNE is low. Moreover, statistical assumption is not needed in GVAD. Consequently, the proposed GVAD has advantage of low computational complexity over those in [1-9] and requires no statistical assumption when compared with [1-3]. Examples for non-stationary AWGN are given to verify the GVAD and then the performance of GVAD is compared with VAD in G.729 and GSM AMR. The results show that the proposed GVAD approach works well in the examples and has better performance than VAD in G.729 and GSM ARM.

As shown in simulation results, the proposed GVAD works well for AWGN examples. In further research, GVAD will be extended to more general cases where additive noise can be babble, pink, and so on.

6 Acknowledgment

This work is supported by National Science Council of Republic of China under grant NSC 95-2221-E-324-040.

7 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.

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.

[10] A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, ITU, ITU-T Rec. G.729-Annex B, 1996.
[11] Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels, ETSI, ETS1 EN 301 708 Recommendation, 1999.
[12] J.L. Deng, "Control Problems of Grey System," *System and Control Letters*, pp. 288-294, 1982.
[13] J. Deng, "Introduction to Grey System Theory," *Journal of Grey System*, Vol. 1, pp. 1-24, 1989.
[14] D.G. Childers, *Speech Processing and Synthesis Toolboxes*, John Wiley & Sons, Inc., 1999.

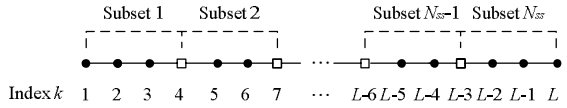

Fig. 1 One sample overlapped subsets for GSNE (K=4)

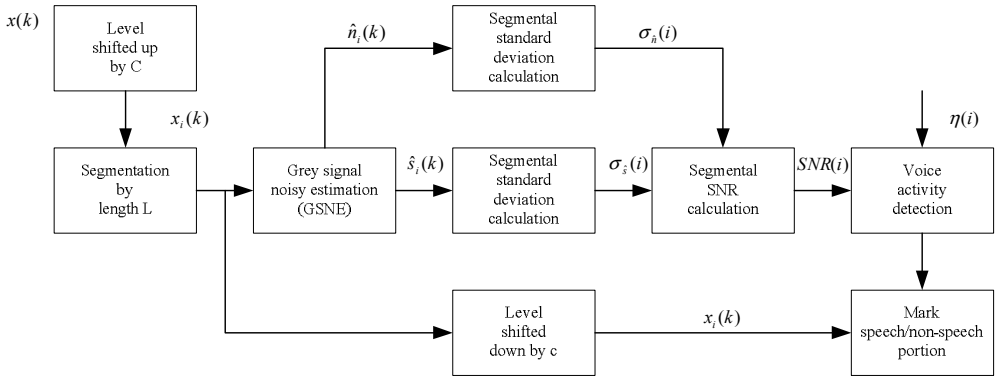

Fig. 2 The flowchart of the proposed GVAD algorithm

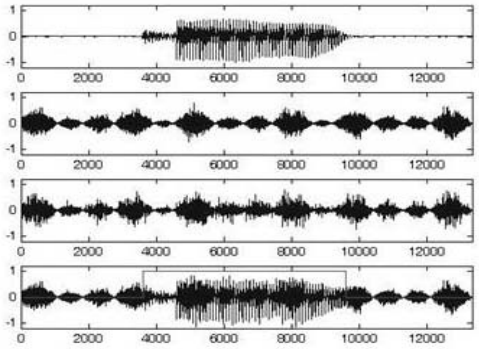

Fig. 3 The clean b.wav, additive noise, estimated noise, and noisy b.wav with GVAD output

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.

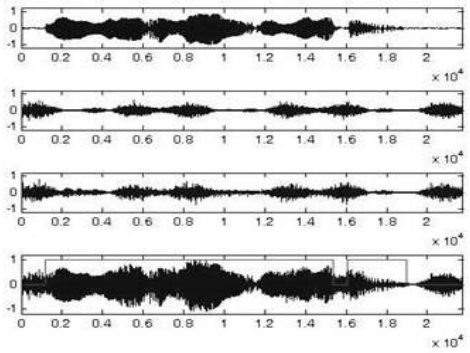

Fig. 4 The clean f0125s.wav, additive noise, estimated noise, and noisy f0125s.wav with GVAD output

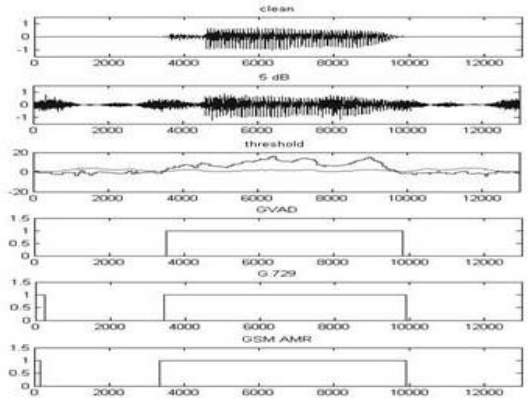

Fig. 5 The clean b.wav, noisy b.wav, threshold, and GVAD, G.729, GSM AMR VAD outputs

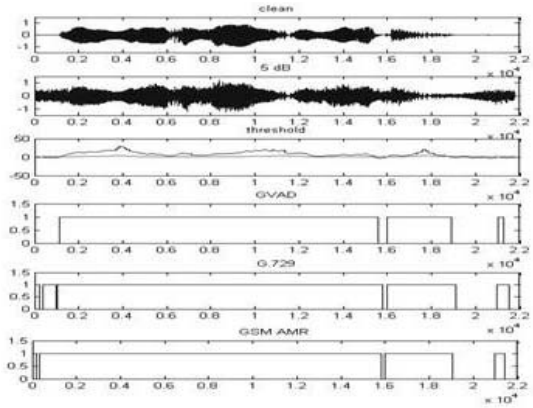

Fig. 6 The clean f0125s.wav, noisy f0125s.wav, threshold, and GVAD, G.729, GSM AMR VAD outputs

6th IEEE/ACIS International Conference on Computer and Information Science (ICIS 2007)
0-7695-2841-4/07 $25.00 © 2007 IEEE

IEEE COMPUTER SOCIETY

Authorized licensed use limited to: CHAOYANG UNIVERSITY OF TECHNOLOGY. Downloaded on April 22,2010 at 06:36:47 UTC from IEEE Xplore. Restrictions apply.