Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 810–819
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

# Energy-based VAD with grey magnitude spectral subtraction

Cheng-Hsiung Hsieh *, Ting-Yu Feng, Po-Chin Huang

*Department of Computer Science and Information Engineering, Chaoyang University of Technology, Wufong 413, Taiwan, ROC*

Received 1 December 2007; received in revised form 11 July 2008; accepted 27 August 2008

**Abstract**

In this paper, we propose a novel voice activity detection (VAD) scheme for low SNR conditions with additive white noise. The proposed approach consists of two parts. First, a grey magnitude spectral subtraction (GMSS) is applied to remove additive noise from a given noisy speech. By this doing, an estimated clean speech is obtained. Second, the enhanced speech by the GMSS is segmented and put into an energy-based VAD to determine whether it is a speech or non-speech segment. The approach presented in this paper is called the GMSS/EVAD. Simulation results indicate that the proposed GMSS/EVAD outperforms VAD in G.729 and GSM AMR for the given low SNR examples. To investigate the performance of the GMSS/EVAD for real-life background noises, the babble and volvo noises in the NOISEX-92 database are under consideration. The simulation results for the given examples indicate that the GMSS/EVAD is able to handle appropriately for the cases of the babble noise with the SNR above 10 dB and the cases of the volvo noise with SNR 15 dB and up.


*Keywords:* Voice activity detection; Grey system; Magnitude spectral subtraction; G.729; GSM AMR

## 1. Introduction

Voice activity detection (VAD) is a very important preprocessing scheme required in many speech systems, such as speech recognition, speech coding, speech communication, and so on. The objective of VAD is to determine if a segment is speech or non-speech. Accurate VAD is able to improve the performance of a speech recognition system in various background noise levels. VAD also can be applied in discontinuous transmission to save battery consumption, to reduce the average bit rate, and to enhance coded speech quality. In a voice over internet protocol (VoIP), VAD is used in speech coding to further reduce the required bandwidth.

During the last decades, many researchers have proposed various approaches to improve the performance of VAD. In (Sohn et al., 1999; Chang and Kim, 2001, 2003), a likelihood ratio test scheme was proposed, where the input speech was transformed by fast Fourier transform. For each frequency component the variance of additive noise was estimated by a recursive formula derived from conditional expectation. Then a composite hypothesis test was employed as a decision rule for the proposed VAD. To improve the VAD performance in (Sohn et al., 1999), a novel scheme called uniformly most powerful test was developed in (Kim et al., 2007) where the decision rule in (Sohn et al., 1999) was modified. By a revised contextual likelihood ratio test, an improved statistical test for VAD was proposed in (Ramirez et al., 2007). In (Longbotham and Bovik, 1989; Ramirez et al., 2005), the approach of subband order statistics filters was presented. In the approach, noise and signal were estimated separately in frequency-domain through subband order statistic filters. Then SNR was obtained and a speech/non-speech segment was determined by a given threshold. In (Tahmasbi and Rezaei, 2007), the generalized autoregressive conditional heteroscedasticity filter was used to model time-varying speeches. Then distribution parameters were estimated. Based on the likelihood ratio test, a segment was

* Corresponding author. Fax: +886 4 23742375.
*E-mail address:* chhsieh@cyut.edu.tw (C.-H. Hsieh).

0167-6393/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2008.08.005

determined as speech or non-speech. Through a set of sub-band log-energies and noise prototypes, a hard-decision clustering scheme was employed to discriminate speech and non-speech segments in (Ramirez et al., 2006). In (Gorriz et al., 2006), the speech/non-speech discrimination was considered as an unsupervised learning problem. Based on a hard-decision clustering scheme, a VAD was devised where the contextual information was used to smooth the decision function and to improve the VAD performance. In (Kim and Park, 2004), a radial basis function neural network was applied to VAD while a genetic programming was used in (Estevez et al., 2005). In (Ramirez et al., 2004), the voice activity detector was developed based on Kullback–Leibler divergence measure. In (Ramirez et al., 2004), a speech segment was classified through long-term spectral divergence. Based on the low-variance spectrum estimation, speech and non-speech segments were determined by SNR measure with an adaptive threshold in (Davis et al., 2006). To extend the Gaussian assumption for most of statistical model based approaches, multiple models were incorporated into the VAD in (Chang et al., 2006).

Note that all approaches mentioned above are performed in frequency-domain except in (Kim and Park, 2004) and that the noisy speech is directly put into the VAD without preprocessing. In this paper, an approach to the VAD problem without any statistical assumption is presented where a grey magnitude spectral subtraction (GMSS) is employed to deal with additive white noise for low SNR cases. Consider the enhanced speech by the GMSS as an estimate of clean speech. Then an energy-based VAD (EVAD) is used to discriminate speech and non-speech segments. The proposed approach is called the GMSS/EVAD.

This paper is organized as follows. In Section 2, the grey magnitude spectral subtraction (GMSS) based on the grey noise estimation (GNE) is described. Next, the proposed energy-based VAD (EVAD) with the GMSS is given in Section 3. In Section 4, simulations are provided to verify the GMSS/EVAD approach whose results are compared with those from G.729 VAD and GSM AMR VAD. Finally, conclusion and further researches are given in Section 5.

## 2. Grey magnitude spectral subtraction

In this section, the grey magnitude spectral subtraction (GMSS) is introduced. Section 2.1 gives a brief review of the GM(1,1) model. In Section 2.2, the GNE based on the GM(1,1) model is described. Then the magnitude spectral subtraction based on the GNE is given in Section 2.3.

### 2.1. Review of the GM(1,1) model

In this section, the GM(1,1) model is briefly reviewed which is then applied to the GNE in Section 2.2. For details about the GM(1,1) model, one may consult (Deng, 1982, 1989). The GM(1,1) modeling process is described in the following. Given non-negative data sequence $\boldsymbol{x} = \{x(1), x(2), \ldots, x(K) \text{ for } 1 \leqslant k \leqslant K\}$, a new data sequence $x^{(1)}(k)$ is found by the first-order accumulated generating operation (1-AGO) as

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

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

$$x(k) + az^{(1)}(k) = b \tag{2}$$

for $2 \leqslant k \leqslant K$, where parameters $a$ and $b$ are called the developing coefficient and the grey input, respectively. The $z^{(1)}(k)$ in (2), called the background value, is defined as

$$z^{(1)}(k) = 0.5[x^{(1)}(k) + x^{(1)}(k-1)]. \tag{3}$$

Let

$$\boldsymbol{y} = \begin{bmatrix} x(2) \\ x(3) \\ \vdots \\ x(K) \end{bmatrix} \tag{4}$$

and

$$\boldsymbol{B} = \begin{bmatrix} -z^{(1)}(2) & 1 \\ -z^{(1)}(3) & 1 \\ \vdots & \vdots \\ -z^{(1)}(K) & 1 \end{bmatrix}. \tag{5}$$

Then (2) can be written as

$$\boldsymbol{y} = \boldsymbol{B}\begin{bmatrix} a \\ b \end{bmatrix}, \tag{6}$$

where parameters $a$ and $b$ are found by

$$\begin{bmatrix} a \\ b \end{bmatrix} = (\boldsymbol{B}^{\mathrm{T}}\boldsymbol{B})^{-1}\boldsymbol{B}^{\mathrm{T}}\boldsymbol{y}. \tag{7}$$

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

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

By the first-order inverse accumulated generating operation (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. The minimum number of data required in the GM(1,1) modeling process is as few as four, i.e. $K = 4$. The GM(1,1) modeling is depicted in Fig. 1.

To justify the GM(1,1) model is able to estimate speech signal appropriately, the clean speech b.wav (male speech