DOI: 10.2298/CSIS100906032Z

Voice Activity Detection Method Based on Multi-valued Coarse-graining Lempel-Ziv Complexity

Huan Zhao[1], Gangjin Wang[1], Cheng Xu[1], and Fei Yu[2]

[1] School of Information Science and Engineering, Hunan University,
410082 Changsha, P. R. China
hzhao@hnu.edu.cn, wgangjin@163.com, chengxu@public.cs.hn.cn
[2] Jiangsu Provincial Key Laboratory of Computer Information Processing Technology,
215000 Suzhou, P. R. China
hunanyufei@126.com

Abstract. One of the key issues in practical speech processing is to locate precisely endpoints of the input utterance to be free of non-speech regions. Although lots of studies have been performed to solve this problem, the operation of existing voice activity detection (VAD) algorithms is still far away from ideal. This paper proposes a novel robust feature for VAD method that is based on multi-valued coarse-graining Lempel-Ziv Complexity (MLZC), which is an improved algorithm of the binary coarse-graining Lempel-Ziv Complexity (BLZC). In addition, we use fuzzy *c*-Means clustering algorithm and the Bayesian information criterion algorithm to estimate the thresholds of the MLZC characteristic, and adopt the dual-thresholds method for VAD. Experimental results on the TIMIT continuous speech database show that at low SNR environments, the detection performance of the proposed MLZC method is superior to the VAD in GSM ARM, G.729 and BLZC method.

Keywords: speech processing, voice activity detection, Lempel-Ziv complexity, multi-valued coarse-graining, fuzzy *c*-Means clustering algorithm, Bayesian information criterion algorithm.

1. Introduction

Voice activity detection (VAD) is used to distinguish speech from noise and is required in many speech applications, such as speech recognition [1], speech enhancement [2], voice biometrics [3], and speech coding [4]. The VAD process is demonstrated in Fig. 1 [5]. Effective VAD of speech signals can not only reduce the amount of speech signals processing operations, but also improve system performance effectively.

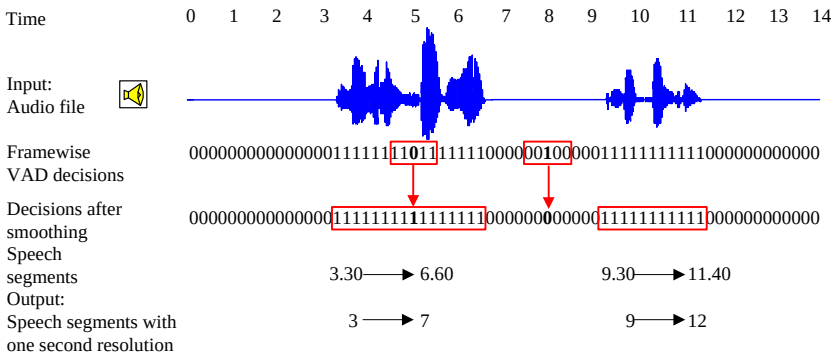

Fig. 1. Illustration of VAD process [5]

At present, various VAD algorithms have been proposed, such as Itakura LPC distance measure [6], cepstral features [7], energy levels [8], the difference of energy and zero-crossing rate [9], spectral entropy [10], energy-spectral entropy [11] and distance entropy [12]. In the condition of high SNR, the differences between voices and background noises are rather distinct so that many algorithms have good performances in VAD. However, the present VAD algorithms have the problem of detection performance in low SNR environments, especially in the presence of non-stationary noise.

From a physics and mathematics viewpoint, many studies have shown that voice signal has non-linear and non-stationary features [13]. Relevant aspects of current research applied in VAD are: permutation entropy [14], approximate entropy [15], C_0 complexity [16], Lempel-Ziv complexity (LZC) [17] and so on. Among them, Ref. [17] proposed a novel VAD algorithm based on binary coarse-graining Lempel-Ziv complexity (BLZC) in the white Gaussian noise environment with a good detection result, but its detection performance declines dramatically in the low SNR non-Gaussian and non-stationary noise environments (such as: factory noise, babble noise, etc.). Considering that the binary coarse-graining method may lose some important information on dynamical systems, we present a novel VAD method based on multi-valued coarse-graining LZC (MLZC). Besides, we use fuzzy *c*-Means clustering (FCMC) [18, 19] and the Bayesian information criterion (BIC) algorithm [19, 20] to estimate the thresholds of MLZC and via dual-thresholds method for VAD. Experimental results show that in a variety of noisy environments, MLZC has a better detection performance than the VAD in GSM ARM, G.729 and BLZC method.

This paper is organized as follows. In Sec. 2, the multi-valued coarse-graining Lempel-Ziv complexity (MLZC) feature is described. Next, the FCMC algorithm and BIC algorithm are applied to estimate the thresholds of the MLZC feature and dual-thresholds VAD method are given in Sec. 3. In Sec. 4, simulations are provided to verify the MLZC approach whose results are compared with the VAD in GSM ARM, G.729 and BLZC method. Finally, the conclusion and further researches are given in Sec. 5.

2. Multi-valued Coarse-graining Lempel-Ziv Complexity

With the development of science, especially of nonlinear science, a common viewpoint has been formed, that is, the speech signal is a complex time series and acts as an unstable strange attractor in a chaotic system rather than a random signal. There have been many definitions of complexity measure, for example, Kolmogorov Complexity (KC). Lempel and Ziv introduced an easy mathematical method to calculate the measure of Kolmogorov Complexity which is defined as Lempel-Ziv complexity (LZC) [21]. LZC analysis is based on a coarse-graining of measurements, i.e. the signal to be analyzed is transformed into a sequence whose elements are only a few symbols. The most widely calculation of LZC is based on the binary sequence which generated by the mean value or zero of the input signal, but the binary sequence cannot well characterize speech signal and may lose some important speech information easily. Therefore, we present a novel VAD method based on multi-valued coarse-graining LZC (MLZC). In the following section we present a detailed study of the MLZC for VAD.

2.1. Binary Coarse-graining Method

In nonlinear time series, the traditional computation of complexity is based on binary sequence, i.e. given a dynamic system time sequence $X=\{x_i|i=1,2,\ldots,n\}$, and then the average for the time series is

$$X_{ave}=\frac{1}{n}\sum_{i=1}^{n}x_i\ , \tag{1}$$

so, the binary sequence $s_i(1\le i\le n)$ can be obtained by

$$s_i=\begin{cases}1, & \text{if} \quad x_i>X_{ave}\\ 0, & \text{else}\end{cases}. \tag{2}$$

2.2. Multi-valued Coarse-graining Method

As the time series produced by the binary coarse-graining method is likely to be missing some important information on dynamical systems, we use multi-valued coarse-graining method reconstruct the time series, which is defined as follows [22]:

Let $X=\{x_i|i=1,2,\ldots,n\}$ be a set of time sequences, and let $x_{\max}$ be the maximum value and $x_{\min}$ be the minimum value of the set. Besides, let $L(L>2)$ be the data coarse-graining segment number of the set, then we define

$$d=\left(x_{\max}-x_{\min}\right)/L\ . \tag{3}$$

Let $\{y_j|j=1,2,\ldots,L\}$be a set of symbol sequences, where each value of y_j is different. Let $s_i(1\le i\le n)$ be the symbol string which is the L-valued coarse-graining result of the set X, defined by

$$s_i = \begin{cases} y_j, \ x_{\min} + (j-1)d \le x_i \le x_{\min} + jd \ \ (j = 1,2,\cdots,L) \\ y_L, \ x_i = x_{\max} \end{cases}. \tag{4}$$

2.3. Lempel-Ziv Complexity

In the previous section we got the symbol string $P=\{s_i|i=1,2,\ldots,n\}$, and the basic idea of Lempel-Ziv complexity analysis is as follows [23, 24]:

Let S and Q denote, respectively, subsequence of the sequence P and SQ be the concatenation of S and Q, while sequence SQv is derived from SQ after its last character is deleted (v means the operation to delete the last character in the sequence). Let $V(SQv)$ denote the vocabulary of all different subsequences of SQv. First, let $c(n)=1$, $S=s_1$, $Q=s_2$, therefore, $SQv=s_1$. Now, suppose $S=\{s_1,s_2,...,s_r\}$, $Q=s_{r+1}$. If $Q\in V(SQv)$, then s_{r+1} is a subsequence of $\{s_1,s_2,\ldots,s_r\}$. At this point, S needn't change and Q update to be $Q=\{s_{r+1},s_{r+2}\}$, then judge whether Q belongs to SQv or not (meanwhile, S needn't change, Q updated and SQv should also update), and continue until $Q\notin V(SQv)$. Now, suppose $Q=\{s_{r+1},s_{r+2},\ldots,s_{r+l}\}$, then $\{s_{r+1},s_{r+2},\ldots,s_{r+l}\}$ is not a subsequence of $\{s_1,s_2,\ldots,s_r,s_{r+1},\ldots,s_{r+l-1}\}$, so increase $c(n)$ by one. Afterwards, combine S with Q and S is renewed to be $S=\{s_1,s_2,\ldots,s_r,s_{r+1},s_{r+2},s_{r+3},\ldots,s_{r+l}\}$, by this time take Q as $Q=s_{r+l+1}$. Repeat these procedures until Q is the last character. At this time, the number of different subsequences is $c(n)$. If the length of the symbol sequence is n, the upper bound of $c(n)$ is given by

$$c(n) < \frac{n}{(1-\varepsilon_n)\log(n)}, \tag{5}$$

where, n is a small quantity and $\varepsilon_n\to 0$ ($n\to 0$). Therefore, in general $n/\log(n)$ is upper bound of $c(n)$, i.e.,

$$\lim_{n\to\infty} c(n) = b(n) \equiv \frac{n}{\log(n)}, \tag{6}$$

so, $c(n)$ is the asymptotic behavior of the random sequence, and $c(n)$ can be normalized via this limit

$$C(n) = \frac{c(n)}{b(n)}. \tag{7}$$

Example: In order to make the calculation of the $c(n)$ easily understood, Fig. 2 shows how to transform a segment of a speech signal series into a ternary sequence by three-valued coarse-graining method (i.e. $L=3$) and the result of complexity analysis on the ternary sequence. After three-valued coarse-graining, the resulting P=000001000012111122221211 (length n=25), and the complex counter $c(n)$ of the sequence P is calculated by complexity

analysis as follows. Symbol "•" denotes the end of each different subsequence, and the number of "•" is equal to the value of $c(n)$.

1) Fist character (i.e. in this case 0) is always a novel one. Therefore, the first subsequence is →0•, i.e. $c(n)$=1.

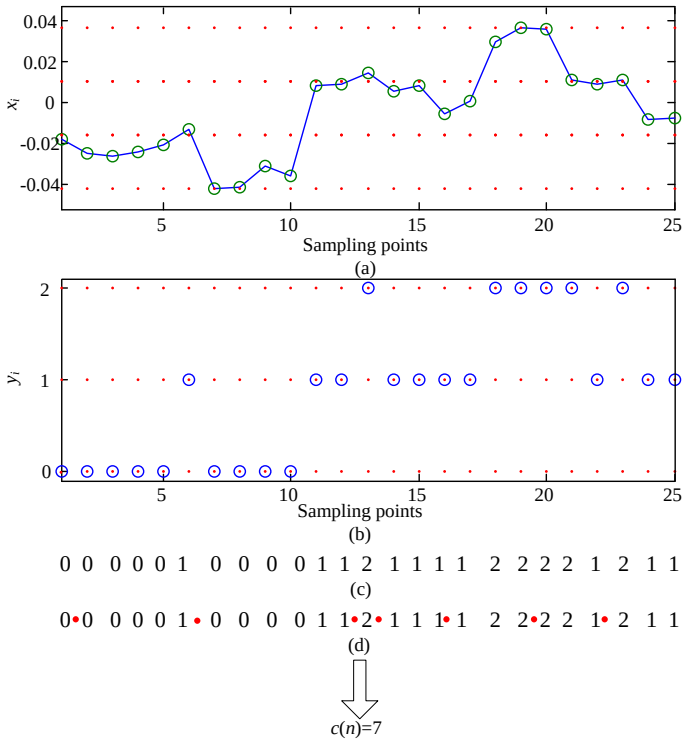

Fig. 2. An illustration showing how to calculate the $c(n)$ (L=3). (a) A segments of speech signal series, (b) Transform the speech signal series into a ternary sequence by three-valued coarse-graining method, (c) Ternary sequences, (d) Complexity analysis result

2) The second character of P is 0 and this is the first subsequence. In this situation, old subsequence S=0, the current subsequence Q=0, concatenated subsequence SQ=00 and previous subsequence SQv=0. Therefore, $Q \in SQv$, so Q is not a new subsequence →0•0, i.e. $c(n)$=1.

3) The third character of P is still 0. The old subsequence (before "•") S=0, the current subsequence Q=00, concatenated subsequence SQ=000 and previous subsequence SQv=00. Therefore, $Q \in SQv$, so Q is not a new subsequence →0•00, i.e. $c(n)$=1.

4) Results of the fourth and fifth character are the same as the third one. When came to the sixth character of P is 1, the old subsequence S=0, the current subsequence Q=00001, concatenated subsequence SQ=000001 and

previous subsequence *SQv*=00000. Therefore, $Q \notin SQv$, so *Q* is a new subsequence →0•00001•, i.e. *c*(*n*)=2.

5) The seventh character of *P* is 0. The old subsequence *S*=000001, the current subsequence *Q*=0, concatenated subsequence *SQ*=0000010 and previous subsequence *SQv*=000001. Therefore, $Q \in SQv$, so *Q* is not a new subsequence →0•00001•0, i.e. *c*(*n*)=2.

6) Before the 12th character of *P*, the new subsequence has not appeared. When came to the 12th character of *P* is 1, the old subsequence *S*=000001, the current subsequence *Q*=000011, concatenated subsequence *SQ*=000001000011 and previous subsequence *SQv*=00000100001. Thus, $Q \notin SQv$, so *Q* is a new subsequence →0•00001•000011•, i.e. *c*(*n*)=3.

7) The 13th character of *P* is 2. The old subsequence *S*=000001000011, the current subsequence *Q*=2, concatenated subsequence *SQ*=00000100001 12 and previous subsequence *SQv*=000001000011. Therefore, $Q \notin SQv$, so *Q* is a new subsequence →0•00001•000011•2•, i.e. *c*(*n*)=4.

8) The 14th character of *P* is 1. The old subsequence *S*=0000010000112, the current subsequence *Q*=1, concatenated subsequence *SQ*=00000100001 121 and previous subsequence *SQv*=0000010000112. Therefore, $Q \in SQv$, so *Q* is not a new subsequence →0•00001•000011•2•1, i.e. *c*(*n*)=4.

9) When came to the 16th character of *P* is 1. The old subsequence *S*=0000010000112, the current subsequence *Q*=111, concatenated subsequence *SQ*=0000010000112111 and previous subsequence *SQv*=000001000011211. Therefore, $Q \notin SQv$, so *Q* is a new subsequence →0•00001•000011•2•111•, i.e. *c*(*n*)=5.

By this process, the sequence *P* is scanned and partitioned as follows:

P=0•00001•000011•2•111•122•221•211

The number of symbol "•" in *P* is seven and this is the value of complexity counter *c*(*n*).

2.4. Algorithm Validation

To verify the effectiveness of LZC in detecting the nonlinear signal, the Logistic model was adopted as a verification object. The Logistic map is a simple mathematical model that describes how the quantity changes of insects over time, which is the best known of the nonlinear dynamic system. This is a one-dimensional Logistic map defined by [25]

$$x_{n+1} = \lambda x_n \left(1 - x_n\right), \tag{8}$$

where λ is an external parameter, $1 \le \lambda \le 4$, and the range of x_n is changed from a circle to the interval [0,1]. Fig. 3(a) shows the evolution of Logistic map bifurcation diagram in the range $3.5 < \lambda < 4$. It is known that there is a stable fixed point $x_n=0$ in the range $0 \le \lambda < 1$, and another stable fixed point $x_n = 1-1/\lambda$ in the range $1 \le \lambda < 3$, we call this periodic be the 1-cycle; when $3 \le \lambda < 1+\sqrt{6}$, x_n always oscillates between two values, and the two values are dependent on λ, we call this periodic be the 2-cycle; when $1+\sqrt{6} \le \lambda < 3.545$, x_n always

oscillates between four values, the 2-cycle is repelling, but a 4-cycle; when $3.545 \leq \lambda < 3.56995$, x_n oscillates between 8 values, then 16, 32…, i.e. the 8-cycle, 16-cycle, 32-cycle…; when $3.56995 \leq \lambda < 4$, the time series undergo the four different evolution stages, i.e. fixed point, unstable fixed point, periodic, and chaotic, until the chaos phenomena. Fig. 3 (b)~(c) shows the change of BLZC and MLZC under the Logistic map evolution, respectively.

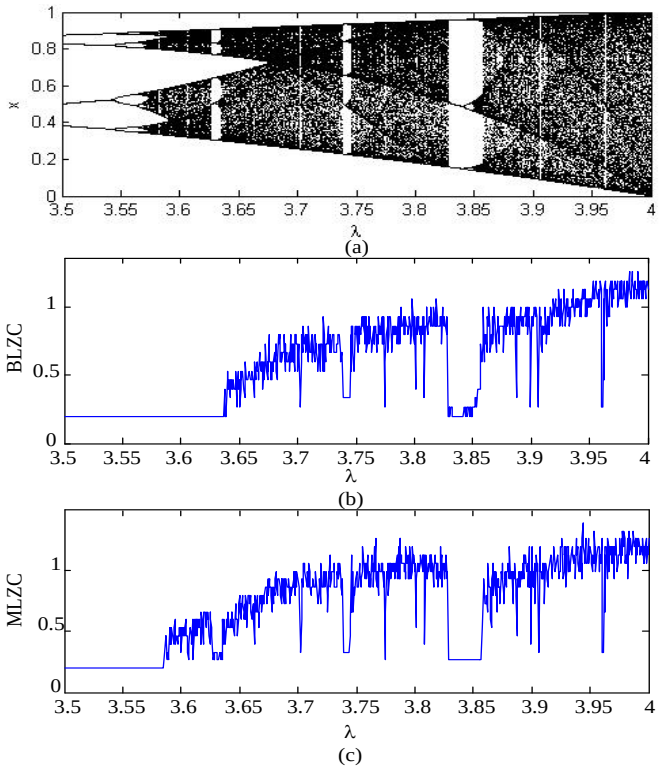

Fig.3. Logistic equations and LZCs for varying control parameter. (a) Bifurcation diagram, (b) BLZC, (c) MLZC

As shown in Fig. 3 (c), when the time series bifurcate evolutes from one state to another state, the MLZC changes obviously, with the same change paces of the Logistic map evolution, i.e. in the range $3.56 < \lambda < 4$. However, Fig. 3 (b) shows that the changes of the BLZC occurs in the range $3.64 < \lambda < 4$. Therefore, the MLZC is superior to the BLZC, which can detect and amplify small changes in the time series and can be used to detect mutations in the signal. Figure 4 displays the LZC of the clean speech under the different coarse-graining methods. We see that the BLZC feature is difficult to distinguish between voice and silence, while $L>2$, different LZC under the L can accurately characterize voice and silence. Without loss of generality, we take $L=3$ in the following discussion.

3. Thresholds Estimation and Algorithm

In this paper, FCMC [18, 19] and BIC [19, 20] algorithms are used to estimate the thresholds of the MLZC feature for VAD. Besides, we use dual-thresholds method for VAD. It can maintain fast tracking speed of environment change

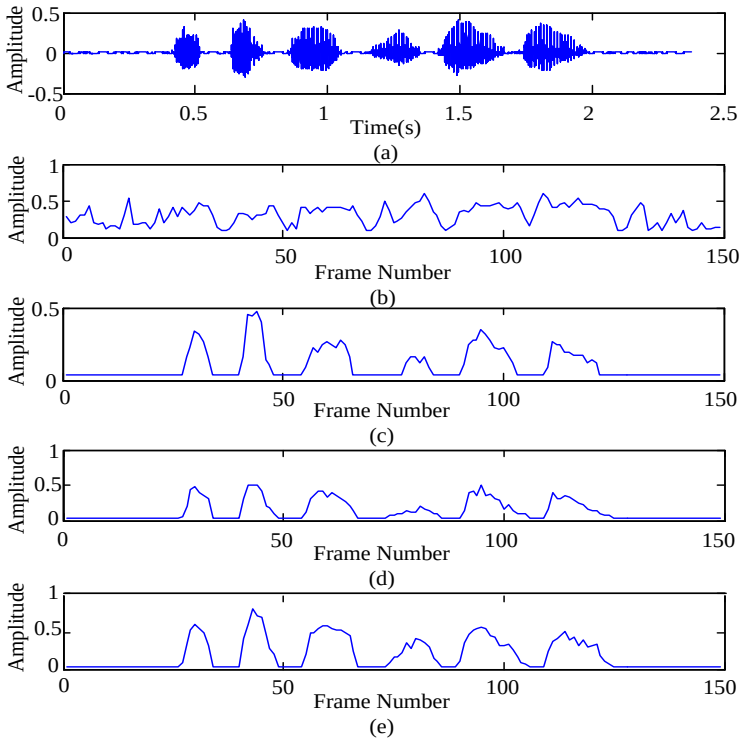

Fig.4. LZC under different coarse-graining methods. (a) Clean speech waveform, (b) BLZC for noisy speech, (c)~(e) MLZC for noisy speech, L=3, 4, 5, respectively

when combined with online update. The brief introduction to the algorithms is given as follows.

3.1. Fuzzy *c*-Means Clustering Algorithm

Assume that the unlabeled object data $X=\{x_i|i=1,2,\ldots,N\}$, C is the expected cluster number and $\{m_j|j=1,2,\ldots,C\}$ are the center of the clusters. The most widely used objective function model for fuzzy c-Means clustering (FCMC) in X is the weighted within groups sum of squared errors objective function J_b, which is used to define the constrained optimization problem:

$$\min\left\{ J_b = \sum_{j=1}^{C}\sum_{i=1}^{N}\left(\mu_j(x_i)\right)^b \left\|x_i - m_j\right\|^2 \right\}, \quad s.t. \sum_{j=1}^{C}\mu_j(x_i) = 1, i = 1,2,\cdots,N \tag{9}$$

where b>1 is the fuzzifier parameter, and $\mu_j(x_i)$ is the grade of membership of x_i in the j-th cluster and subjects to the constrains.

Minimization of J_b subjects to constrains, leads to the following function:

$$m_j = \frac{\sum_{i=1}^{N}\left(\mu_j(x_i)\right)^b x_i}{\sum_{i=1}^{N}\left(\mu_j(x_i)\right)^b},\ j = 1,2,\cdots,C, \tag{10}$$

$$\mu_j(x_i) = \frac{\left(1/\left\|x_i - m_j\right\|^2\right)^{1/(b-1)}}{\sum_{k=1}^{C}\left(1/\left\|x_i - m_k\right\|^2\right)^{1/(b-1)}},\ i = 1,2,\cdots,N,\ j = 1,2,\cdots,C. \tag{11}$$

Using iterative method for solving (10) and (11), we get the fuzzy c-Means clustering algorithm.

3.2. Bayesian Information Criterion Algorithm

For a speech signal, we need to determine whether it contains a clean voice only, or also includes the background noise. In this paper, we use the Bayesian information criterion (BIC) algorithm to determine the best cluster number [19, 20].

According to BIC, the best model number is the one with maximized BIC value. If voice and the background noise are modeled as a multi-variance Gaussian distribution $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ is the sample mean vector and $\boldsymbol{\Sigma}_i$ is the sample covariance matrix, the BIC value is [19]

$$\mathrm{BIC}(C) = \sum_{i=1}^{C}\left\{-\frac{1}{2}N_i \log\left|\boldsymbol{\Sigma}_i\right|\right\} - \frac{\log(N)}{2}\lambda_p C\left[d + \frac{d(d+1)}{2}\right], \tag{12}$$

where N is the total sample number, N_i is the number of sample in the i-th cluster, λ_p is the penalty weight, and d is the dimension of the feature space.

We applied the BIC criterion to determine the best cluster number C_{best} for VAD can be present as

$$C_{best} = \begin{cases} 1, & \text{if} \quad \mathrm{BIC}(1) > \mathrm{BIC}(2) \\ 2, & \text{else} \end{cases}. \tag{13}$$

3.3. Thresholds Estimation

In this section, how to use the FCMC and BIC algorithms to ascertain the thresholds for VAD is illustrated. Before making thresholds estimation, we need to pass framing, adding window and other pretreatment to speech signals. The algorithm steps are as follows [19]:

Step 1 Calculate the MLZC for each frame by (7).

Step 2 Given the cluster number C=2, making FCMC on the MLZC of frames.

Step 3 Use (13) to determine the best cluster number C_{best}.

Step 4 IF C_{best} =1

Step 2 obtained the cluster center m_{11}, and then the thresholds formula of MLZC is:

$$\mathrm{TH}_{high} = m_{11} + \alpha_{high}, \tag{14}$$

$$\mathrm{TH}_{low} = m_{11} + \alpha_{low}, \tag{15}$$

where TH_{high} and TH_{low} are the higher and lower thresholds respectively, and α_{high}, α_{low} are empirical constants.

ELSE

Step 2 obtained the cluster centers m_{21} and m_{22}, then the mean of MLZC of the voice and background noise are given by

$$m_{speech} = \max\{m_{21}, m_{22}\}, \tag{16}$$

$$m_{noise} = \min\{m_{21}, m_{22}\}. \tag{17}$$

So the threshold formula of MLZC is:

$$\mathrm{TH}_{high} = m_{noise} + (m_{speech} - m_{noise})\beta_{high}, \tag{18}$$

$$\mathrm{TH}_{low} = m_{noise} + (m_{speech} - m_{noise})\beta_{low}, \tag{19}$$

where, β_{high}, β_{low} are empirical constants.

END

3.4. Dual-thresholds Method

After obtaining the thresholds of MLZC by the above steps, we use the dual-thresholds method for VAD. The dual-thresholds arithmetic is first introduced by Lawrence Rabiner [26]. The improved algorithm we use for VAD can be present as follows [12, 27].

Step 1 As shown in Fig. 5, the higher threshold TH_{high} and the lower threshold TH_{low} are got in Sec. 3.3.

Step 2 Comparing the current MLZC with the TH_{high}, if MLZC>TH_{high}, the current frame affirmatively belongs to voice signal segment. As a result, we can obtain the two approximate endpoints N_1 and N_2.

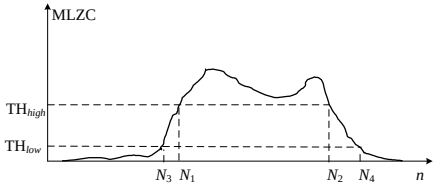

Fig. 5. Dual-thresholds for VAD

Step 3 Searching forward from N_1, if in the first frame MLZC<TH_{low}, then we recorded the frame as N_3. Besides, searching backward from N_2, if in the first frame MLZC<TH_{low}, then we recorded the frame as N_4. Therefore, we can initially get the starting endpoint N_3 and the ending endpoint N_4. N_3N_4 is defined as a segment of voice signal.

Step 4 If the voice segment is less than 4 frames, take it as the result of mutation of the background noise, and it should be omitted.

Step 5 If the interval between adjacent voice segments is less than 0.2s, merge the two adjacent voice segments.

4. Experiments

4.1. Experimental setup

The original speeches used for simulation and test are taken from the DARPA TIMIT Acoustic-Phonetic Speech Corpus [28]. 450 English sentences are selected. All of them are sampled at 16 kHz, and quantized into 16 bits. Different background noises with different time-frequency distributions were taken from the NOISEX-92 database [29]. The tested noisy environments include White noise, Babble noise, Factory noise, Volvo (car) noise. Noise has been added to the clean speech signal with 5 SNRs (0, 5, 10, 15, 20dB).

Based on the above experimental speech environments, we set the values of each parameter as follows: the speech frame length is 512 (32ms); frame shift is 256 (16ms); window function is hamming window; and experiments show that the threshold estimated parameters are: α_{high}=5.4, α_{low}=-0.24, β_{high} =0.15, β_{low}=-0.042.

4.2. Experimental Results of VAD

In this section, we carried out a series of experiments to evaluate the effectiveness of the VAD algorithm. As shown in Fig. 6~Fig. 9, the VAD outputs for the MLZC are investigated in the given speeches, whose results

are compared with the baseline algorithms from GSM AMR VAD [30], G.729 VAD [31] and BLZC [17]. We can see that, as the SNR dropped, GSM AMR's VAD detection performance became stably, and can only detect part of the voice; G.729's VAD detection performance declined sharply, especially when SNR = 5dB (in Fig. 9(c)), the whole noisy speech was detected as the speech sound; BLZC's detection performance has also declined but still been able to find all of the voice, yet falsely detected the three speech segments as one voice segment; However, MLZC's detection performance showed good robustness, and accurately detected the speech starting and ending positions. In order to better distinguish BLZC and MLZC detection performance, Fig. 10 and Fig. 11 respectively show the VAD results of the BLZC and MLZC in different noise environments (SNR = 5dB). As can be seen from the figures, the voice truncated errors (the voice misclassified as the noise) of the BLZC method is very common (such as, in babble noise and factory noise environments) and its extended errors (the noise misclassified as the voice) exit in the 4 noises environments. Fortunately, the MLZC only h-

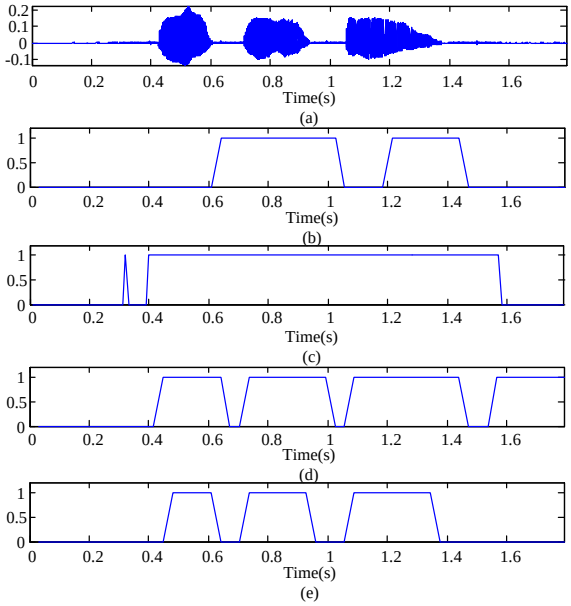

Fig. 6. (a) Clean speech waveform, (b) GSM AMR VAD results, (c) G.729 VAD results, (d) BLZC VAD results, (e) MLZC VAD results

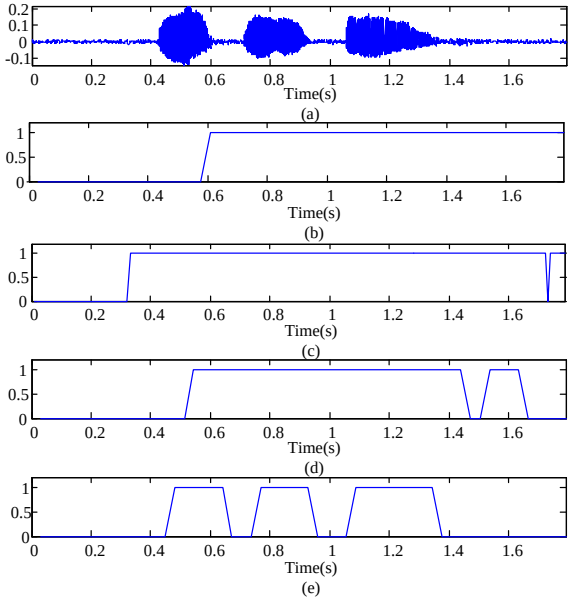

Fig. 7. (a) Noisy speech waveform (Babble noise, SNR=15dB), (b) GSM AMR VAD results, (c) G.729 VAD results, (d) BLZC VAD results, (e) MLZC VAD results

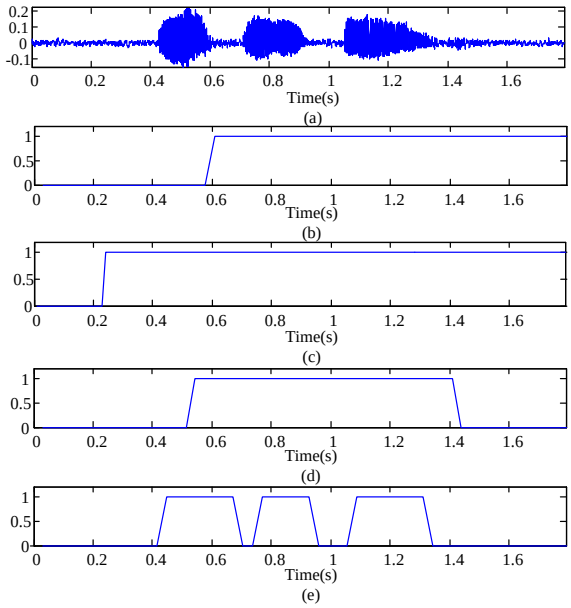

Fig. 8. (a) Noisy speech waveform (Babble noise, SNR=10dB), (b) GSM AMR VAD results, (c) G.729 VAD results, (d) BLZC VAD results, (e) MLZC VAD results

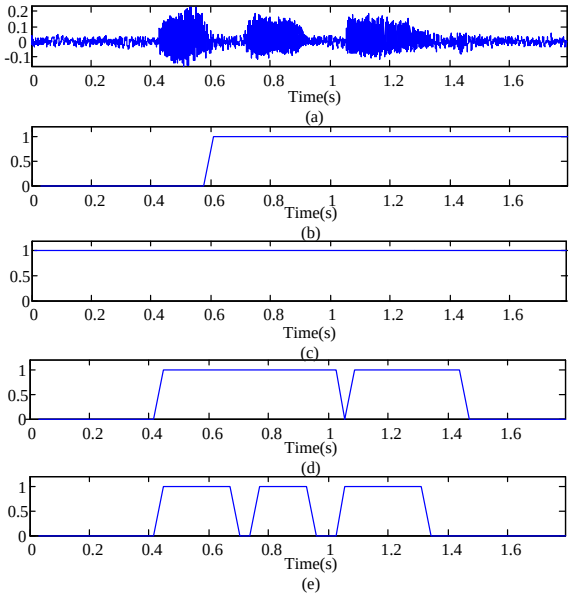

Fig. 9. (a) Noisy speech waveform (Babble noise, SNR=5dB), (b) GSM AMR VAD results, (c) G.729 VAD results, (d) BLZC VAD results, (e) MLZC VAD results

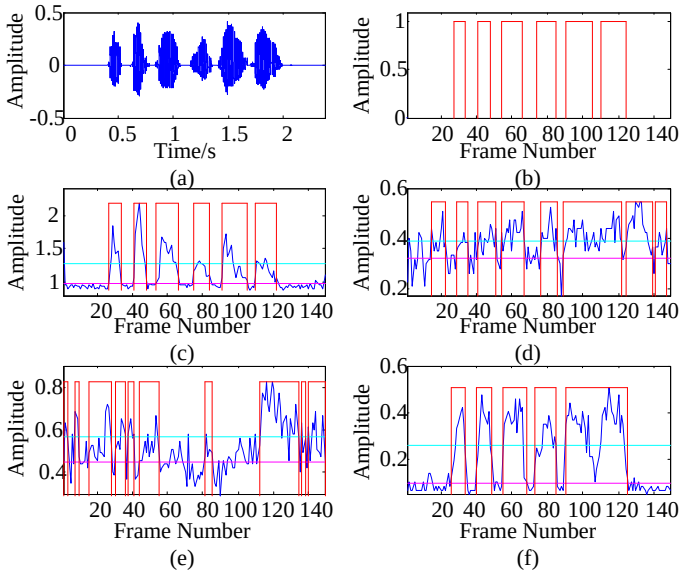

Fig. 10. VAD results by BLZC in different noise environments (SNR=5dB), (a) clean speech waveform, (b) VAD results by hands, (c)~(f) VAD results by BLZC in White, Babble, Factory, and Volvo noise, respectively

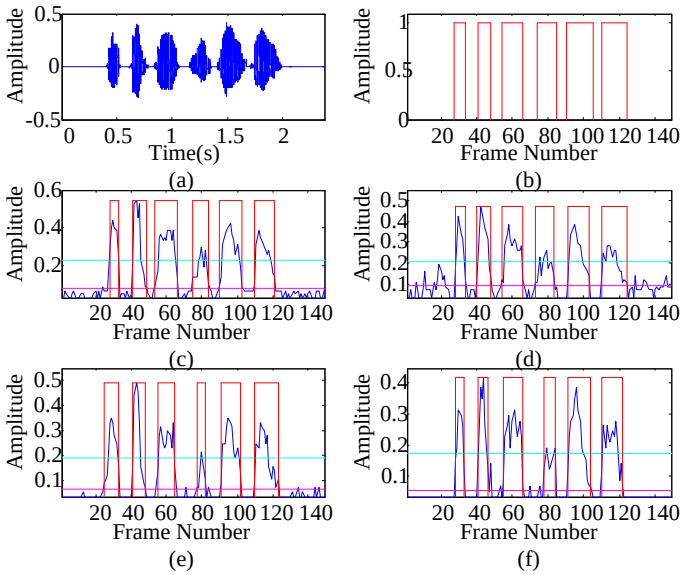

Fig. 11. VAD results by MLZC in different noise environments (SNR=5dB), (a) clean speech waveform, (b) VAD results by hands, (c)~(f) VAD results by MLZC in White, Babble, Factory, and Volvo noise, respectively

as a small amount of voice truncated error in the Factory and Volvo noise environments and can accurately detect the speech endpoints in other cases.

Due to truncation error and extended error which exist in the VAD, the experimental analysis is based on the Weighted Average Error measurement (WA) and its definition is [27, 32]:

$$\mathrm{WA}=\frac{K_C \cdot \mathrm{CLP} + K_W \cdot \mathrm{WDN}}{fNum}. \tag{20}$$

where, CLP stands for truncated error frame, WDN stands for extended error frame, and *fNum* means the total frame number of the sampled signal. K_C and K_W are weighting coefficients. According to people's subjective feelings, extended signals are more acceptable than truncated [32], so K_C=1.4, K_W=0.6. The WA results in different noise and SNR environments are shown in Table 1. It can be found that the WA values of MLZC in different situations are lower than BLZC, which demonstrates that MLZC has a better VAD performance than BLZC, and it also can be found MLZC in the Volvo noise has a lower WA, which indicates MLZC would have a good application prospect in the vehicle environment.

Table 1. WA results by two methods in different noise and SNR environments (%)

Noise	VAD	0dB	5dB	10dB	15dB	20dB
White	BLZC	26.71	13.67	9.26	8.19	7.65
	MLZC	12.21	11.28	7.52	6.57	6.04
Babble	BLZC	29.13	27.78	25.63	24.69	22.55
	MLZC	22.14	17.32	15.83	14.09	8.85
Factory	BLZC	57.85	57.71	50.58	48.38	46.93
	MLZC	25.90	21.61	19.73	17.72	9.79
Volvo	BLZC	8.18	7.78	6.97	5.91	5.50
	MLZC	6.97	6.58	5.63	4.81	4.25
Average	BLZC	30.47	26.73	23.11	21.79	20.65
	MLZC	16.81	14.19	12.17	10.79	7.23

5. Conclusions

In this paper, we propose a new VAD method that is multi-valued coarse-graining Lempel-Ziv Complexity (MLZC), which use fuzzy *c*-Means clustering algorithm and Bayesian information criterion algorithm to estimate the thresholds of the MLZC characteristic, and dual-thresholds method for VAD. Experimental results show that at low SNR environments, MLZC method is superior to the binary coarse-graining Lempel-Ziv Complexity (BLZC) method, especially in the vehicle interior noise environments, where MLZC method shows better detection performance. Therefore, we can say that MLZC method has a good application prospect and can provide accurate VAD techniques for car navigation.

In summary, there are several advantages that can be seen in the proposed VAD method: 1) Compared with the binary coarse-graining method, the multi-valued coarse-graining method can better perform the characteristics of speech signals. 2) We propose the novel non-linear feature of MLZC for VAD which could capture underlying model differences of speech and noise. 3) We use fuzzy *c*-Means clustering algorithm and Bayesian information criterion algorithm to estimate the thresholds, which more robust and heuristic-rules-free than previous thresholds estimation algorithms. In future work, we will apply the proposed VAD method to the speech recognition and speech applications in the car. Of course, it needs further verification.

Acknowledgements. This work is supported by Hunan Provincial Natural Science Foundation of P. R. China (Grant No.10JJ2046), and the Planned Science and Technology Key Project of Hunan Province, P. R. China (Grant No.2010GK2002).

References

1. Asano, F., Yamamoto, K., Hara, I., Ogata, J., Yoshimura, T., Motomura, Y.: Detection and separation of speech event using audio and video information fusion and its application to robust speech interface. EURASIP Journal on Applied Signal Processing 2004(11), 1727-1738 (2004)
2. Gilg, V., Beaugeant, C., Schonle, M., Andrassy, B.: Methodology for the design of a robust voice activity detector for speech enhancement. In: International Workshop on Acoustic Echo and Noise Control (IWAENC'2003). pp. 131-134. Kyoto, Japan (2003)
3. Tong, R., Ma, B., Lee, K.A., You, C. H., Kinnunen T., Sun, H.W.: Fusion of acoustic and tokenization features for speaker recognition. In: 5th International Symposium on Chinese Spoken Language Processing (ISCSLP'2006). pp. 566-577. Singapore (2006)
4. Zhang, L., Gao, Y. C., Bian, Z.Z., Lu, C.: Voice activity detection algorithm improvement in adaptive multi-rate speech coding of 3GPP. In: 1st International conference on Wireless Communications, Networking and Mobile Computing. vol. 2, pp. 1257-1260. Wuhan, China (2005)
5. Marko, T., Rosa, G.H., Pasi, F.: Automatic voice activity detection in different speech applications. In: Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop. p. 12. Adelaide, Australia (2008)
6. Rabiner, L.R., Sambur, M. R.: Voice-unvoiced-silence detection using the Itakura LPC distance measure. In: Proceeding international Commence Acoustic. Speech, Signal Processing (ICASSP'1977). pp, 323-326. Hartford, Connecticut, USA (1977)
7. Haigh, J.A. Mason, J. S.: Robust voice activity detection using cpectral features. In: IEEE Region 10 Conference on Proceedings, Computer, Communication, Control and Power Engineering (TENCON'1993). pp. 321-324. Beijing, China (1993)
8. Junqua, J. C., Mark, B., Reaves, B.: A robust algorithm for word boundary detection in the presence of noise. IEEE Trans on speech and Audio Processing 2(3), 406-412 (1994)
9. Beritelli, F., Casale, S., Ruggeri, G., Serrano, S.: Performances evaluation and comparison of G.729/AMR/fuzzy voice activity detectors. IEEE Signal Processing Letters 9(3), 85-88 (2002)
10. Shen, J. L., Hung, J.W., Lee, L.S.: Robust entropy-based endpoint detection for speech recognition in noisy environments. In: International Conference on Spoken Language Processing (ICSLP'1998). pp. 232-235. Sydney, Australia (1998)
11. Huang, L.S., Yang, C.H.: A novel approach to robust speech endpoint detection in car environments. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'2000). vol. 3, pp. 1751-1754. Istanbul, Turkey (2000)

12. Zhao, H., Zhao, L.X., Zhao, K., Wang, G.J.: Voice activity detection based on distance entropy in noisy environment. In: Fifth International Joint Conference on INC, IMS and IDC. pp. 1364-1367. Seoul, Korea (2009)
13. Teager, H., Teager, S.: Evidence for nonlinear sound production mechanisms in the vocal tract. In: Proceeding of NATO ASI on Speech Production and Speech Modeling. pp. 241-261. Boston, USA (1990)
14. Christoph, B., Bemd, P.: Permutation entropy-a natural complexity measure for time series. PhysRev Lett, vol. 88, p. 174102 (2002)
15. Lei, X.G. Zeng, Y.C., Li, L.: Noisy speech endpoint detection based on approximate entropy. Technical Acoustics 26(2), 121-125 (2007)
16. Fan, Y.L., Wu, C.Y., Li, Y.: Application of C$_{0}$ complexity measure in detecting speech. Chinese Journal of Sensors and Actuators 19(3), 750-753 (2006)
17. Huang, H.Y., Lin F.H.: A speech feature extraction method using complexity measure for voice activity detection in WGN. Speech Communication 51(9), 714-723 (2009)
18. Nikhil, R.P., James, C.B.: On cluster validity for the fuzzy \emph{c}-means model. IEEE Transactions on Fuzzy Systems 3(3), 370-379 (1995)
19. Tian, Y., Wu, J., Wang, Z.Y.: Fuzzy clustering and Bayesian information criterion based threshold estimation for robust voice activity detection. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'2003). vol. 1, pp. 444-447 (2003)
20. Chen, S. S., Gopalakrishnan, P.S.: Clustering via the Bayesian information criterion with applications in voice recognition. In: International Conference on Acoustics, Speech and Signal Processing (ICASSP'1998). vol. 1, pp. 645-648 (1998)
21. Lempel, A., Ziv, J.: On the complexity of finite sequences. IEEE Transactions on information theory 22(1), 75-81 (1976)
22. Zhang, D.Z.: Research on the correlation between the mutual information and Lempel-Ziv complexity of nonlinear time series. Acta Physica Sinica 56(6), 3152-3157 (2007)
23. Zhang, X.S., Roy, R.J.: EEG complexity as a measure of depth of anesthesia for patients. IEEE Transactions on Biomedical Engineering 48(12), 1424-1433 (2001)
24. Liu, F.T., He, G.G.: Complexity measure for macroscopical transportation system using Lempel-Ziv algorithm. Journal of Harbin Institute of Technology 40(12), 2058-2061 (2008)
25. Schuster, H.G., Leiserson, C.E., Rivest, R.L.: Deterministic chaos: An Introduction. pp. 7-8. Wiley-VCH Verlag GmbH \& Co. KGaA, Weinheim, Germany (2005)
26. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. Prentice Hall, New Jersey, USA (1999)
27. Zhang, S.Y., Guo, Y., Zhang, Q.: Robust voice activity detection feature design based on spectral kurtosis. In: 1st International Workshop on Education Technology and Computer Science (ETCS'2009). pp. 269-272. Wuhan, China (2009)
28. John, S.G., Lori, F.L., William, M.F., Jonathan, G.F., David, S.P., Nancy, L.Da., Victor, Z.: DARPA TIMIT Acoustic-Phonetic Speech Corpus. Linguistic Data Consortium, Philadelphia (1993). [Online]. Available: http://www.ldc.upenn.edu /Catalog/CatalogEntry.jsp?catalogId=LDC93S1 (current Nov. 2010)

29. NOISEX-92 Database. [Online]. Available: http://spib.rice.edu/spib/select_noise.html (current Nov. 2010)
30. Voice Activity Detector (VAD) for adaptive multi-rate (AMR) speech traffic channels, ETSI, ETS1 EN 301 708 Recommendation (1999). [Online]. Available: http://www.3gpp.org/ftp/specs/html-info/0694.htm (current Nov. 2010)
31. A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70, ITU, ITU-T Rec. G.729-Annex B (1996). [Online]. Available: http://eu.sabotage.org/www/ITU/G/G0729nbe1.pdf (current Nov. 2010)
32. Xu, W., Ding, Q., Wang, B. X.: A speech endpoint detector based on eigenspace energy entropy. Journal of China institute of communications 24(11), 125-132, (2003)

Huan Zhao is a professor at the School of Information Science and Engineering, Hunan University. She obtained her B.Sc. degree and M.S. degree in Computer Application Technology at Hunan University in 1989 and 2004, respectively, and completed her Ph.D. in Computer Science and Technology at the same school in 2010. Her current research interests include speech information processing, embedded system design and embedded speech recognition. She served as visiting scholar at the University of Califomia-San Diego (UCSD), USA during the period of March 2008 to September 2008. The visiting scholarship was appointed and sponsored by the China Scholarship Council (CSC). Prof. Zhao is a Senior Member of China Computer Federation, Governing of Hunan Computer Society, China and China Education Ministry Steering Committee Member of Computer Education on Arts. She has published more than 40 papers and 6 books.

Gangjin Wang received his B.Sc. degree in Mathematics and Applied Mathematics at the School of Mathematics and Information Science, Henan Polytechnic University, P. R. China in 2008. Currently, he is a M.S. candidate of Hunan University, P. R. China. His current research interests include speech information processing and voice activity detection.

Cheng Xu is a professor at the School of Information Science and Engineering, Hunan University. He received B.Sc. degree and M.S degree from Hunan University in P. R. China in 1983 and 1986 respectively. He received Ph.D. degree in Mechanical Manufacture and Automation from Wuhan University of Technology in P. R. China in 2006. His current research interests include embedded system, network multimedia applications and control theory.

Fei Yu was born in Ningxiang, P. R. China, on February 06, 1973. Before Studying in Peoples' Friendship University of Russia, Russia, he joined and worked in Hunan University, Zhejiang University, Hunan Agricultural University, P. R. China. He has wide research interests, mainly information technology. In these areas he has published above 50 papers in journals or

conference proceedings and a book has published by Science Press, China (Fei Yu, Miaoliang Zhu, Cheng Xu, et al. Computer Network Security, 2003). Above 30 papers are indexed by SCI, EI. He has won various awards in the past. He served as many workshop chair, advisory committee or program committee member of various international ACM/IEEE conferences, and chaired a number of international conferences such as IITA'07, ISIP'08, ISECS'08, ISIP'09, ISECS'09 and ISISE'08. He have taken as a guest researcher in State Key Laboratory of Information Security, Graduate School of Chinese Academy of Sciences, Guangdong Province Key Lab of Electronic Commerce Market Application Technology, Jiangsu Provincial Key Lab of Image Processing and Jiangsu Provincial Key Laboratory of Computer Information Processing Technology.

Received: September 06, 2010; Accepted: January 19, 2011.