International Journal of Control and Automation
Vol. 9, No. 12 (2016), pp.191-202
http//dx.doi.org/10.14257/ijca.2016.9.12.17

The Application of Extreme Learning Machine and Support Vector Machine in Speech Endpoint Detection

Zhigang Feng, Junlei Feng and Fangyuan Dai

School of Automation, Shenyang Aerospace University, Shenyang, Liaoning, China
fzg1023@yeah.net

Abstract

In this paper, a general voice activity detection (VAD) method based on pattern recognition is proposed, and a specific algorithm of endpoint detection is researched. In this method, the Extreme Learning Machine (ELM) and Genetic Algorithm (GA) optimization Support Vector Machine (SVM) is used as the training and recognition model. The simulation results indicates that ELM and GA-SVM have the same superior endpoint detection accuracy, and recognition time were similar, but the training time of ELM only up to a 1/2000 of the GA-SVM, the robustness of ELM and GA-SVM is greatly improved in noisy environment compare with the traditional VAD that depends on time-domain energy and zero crossing rate.

Keywords*: Endpoint Detection, Extreme Learning Machine, Support Vector Machine, Genetic Algorithm*

1. Introduction

In speech recognition systems, the accuracy of voice activity detection (VAD) could reduce the amount of calculation and improve the accuracy of speech recognition. The VAD is the most important part of the voice adaptive enhancement algorithm and speech coding system. Studies have shown that more than half of the recognition errors came from the voice activity detector even the speech recognition system in a quiet environment [1]. Currently, there are several algorithms of VAD. In summary, those algorithms can be divided into three categories: short term energy, double threshold and information entropy endpoint detection methods. The first two detection methods use the time domain feature of the speech voice signals, such as short term energy and zero crossing detection methods. Those methods can obtain satisfactory detection results under high signal-noise ratio conditions or the noise signal is a particular type [2], but cannot detect voice signal effectively or even cannot work under low signal-noise ratio. The algorithm based on information entropy has certain abilities to distinction between mute and voice, the performance of this algorithm is not very satisfactory when the ambient noise is strong.

In this paper, in order to overcome the shortcomings of existing technologies, an endpoint detection algorithm based on pattern recognition is proposed, and the support vector machine optimized by extreme learning machine (ELM) and genetic algorithms (GA) as the training and recognition models [3]. This method uses the combination of time domain energy, Linear Prediction Coefficient (LPC), Mel Frequency Cepstrum Coefficient (MFCC), Differential Linear Predictive coefficient (Δ-ΔLPC), Differential Mel Frequency Cepstrum coefficient (Δ-ΔMFCC) as the feature data to divided the voice signal and judged frame by frame, filtering the frame label according to certain rules, followed by frame conversion, the starting and ending points of the speech could be gotten finally. Experimental results show that the proposed algorithm has high detection accuracy of the VAD and high robustness to ambient noise.

ISSN: 2005-4297 IJCA

2. The Theory of Endpoint Detection

The VAD algorithm could identify the beginning and ending point of a voice signal, to filte out the unwanted signals and extract the useful signals. In this paper, the pattern recognition method was used to detect the activity of voice. The algorithm consists of two steps: Training and Recognition [5]. The theory of this VAD algorithm is shown in Figure 1. In training stage, firstly, the voice and noise sequence is identified, then characteristic parameters of the sequence are obtained through feature extraction, finally, the training library is established using the extracted features. In recognition stage, the characteristic parameters of the voice sequence which needed to be detected are extracted, then voice activity detection results are gotten using the training library, finally the endpoint could be gotten after filter and frame point conversion [6].

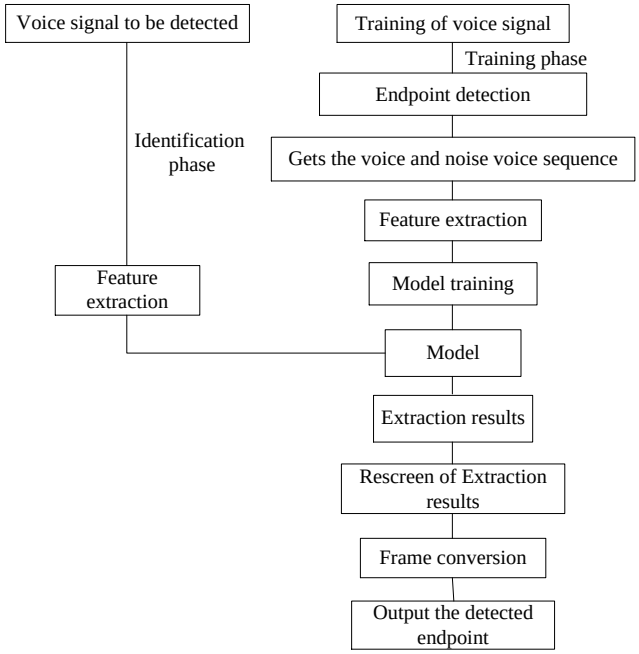

Figure 1. The Theory of This VAD Algorithm

3 Algorithm of VAD

3.1. Feature Selection

The merits of VAD algorithm depends on the choice of the characteristics. For VAD, the more characteristics the identification would be better. But add feature would increase the burden of machine and the detection time. So the merits of the characteristics and the complexity of the algorithm would be taken into account when select the characteristic data. In this paper, the combination of time domain energy coefficient, LPC, MFCC, Δ-ΔLPC and Δ-ΔMFCC was chosen as the characteristic data.

(1) Time domain energy

The characteristic of time domain energy has an excellent performance in the field of distinguishing between silent and sound. In silent stage, the E in formula (1) would have a significant difference between the sound stage.

$$E = \sum_{i=1}^{N} e(i)^2 \tag{1}$$

(2) LPC

LPC analysis technique is widely used speech feature parameter extraction technique, reflects the speaker's vocal tract characteristics to a certain extent [7], human's vocal tract characteristics have a difference with noise evaluation's vocal tract characteristics in a certain extent. Linear prediction starts from the mechanism of human voice. Considered the system's transfer function is fully consistent with all-pole digital filter. The modal can be expressed as formula (2), where P is the order of the linear prediction filter.

$$H(z) = \frac{1}{1 - \sum_{k=1}^{P} a_k z^{-k}} \tag{2}$$

Solving LPC is to solve the value of a_k in the formula (2). There are many solution of LPC characteristic, such as auto-correlation, covariance method, lattice method, inverse filter formula, Spectral estimation formula, maximum likelihood formula, inner product formula. In this paper, the inverse filter formula was chosen as the solution method and a_k could be obtained after finitely recursive operations.

(3) MFCC

MFCC parameter describes a frame's voice signal from the Cepstrum domain, which reflects the human auditory characteristics in part and has a certain ability to distinguish the voice and noise. The sensitivity of human ear is not linear relationship but similar to logarithmic relationship. The relationship between Mel frequency dial and linear frequency dial as shown in formula 3, where f stands for the frequency of linear dial and f_{mel} .stands for the dial of frequency of frequency dial.

$$f_{mel} = 2595 \lg(1 + \frac{f}{700}) = 1127 \ln(1 + \frac{f}{700}) \tag{3}$$

The general computational procession of MFCC is shown in Figure 2. The MFCC could be got after pre-emphasis, framing, adding-windows analysis, FFT transforms, Mel filter, logarithmic transformation, discrete cosine transform (DCT) and other operations.

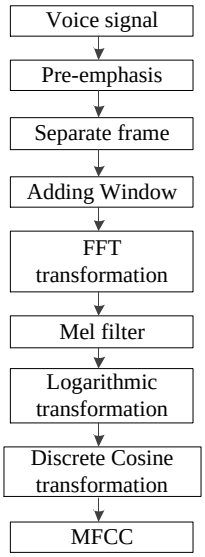

Figure 2. Flowchart of MFCC Calculate

(4) Δ-Δ features

MFCC and LPC considered the information in a frame, doesn't take into account the information between frames when extracted features. The Δ-Δ features were used in this paper to retain the information between frames. Δ feature refers to the feature which could be gotten after feature vector does Fourier transform of the voice frame under the time period of voice frame sequence. The feature would be called Δ-Δ feature if Δ feature does Fourier transform again. In actual operation, often choose the simplified Δ-Δ features. As shown in formula (4), d_t stands for the Δ feature in the t frame. In this paper, the $c_{t-\theta}$ and $c_{t+\theta}$ stands for the $t-\theta$ or $t+\theta$ frame's LPC or MFCC feature vector, N stand for the number of voice frame when the t frame's sequence changes.

$$d_t = \frac{\sum_{\theta=1}^{N} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{N} \theta^2} \tag{4}$$

3.2. Determination of Preprocessing Parameters and Extraction Feature Order

The preprocessing of signal contains the pre-emphasis, framing, adding-windows analysis of signals. The pre-emphasis of signals is in order to compensate the high frequency energy dissipation. The framing and adding-windows analysis is to eliminate the smoothing attenuation of the signal at a window boundary. The extracted characteristic parameters include the LPC and MFCC vector dimension, Δ-Δ related frame number. The preprocessing and the extracted characteristic table as shown in Table 1.

Table 1. The Preprocessing and the Extracted Characteristic Table

Parameter	Value
Pre-emphasis coefficient	0.95
Recording frequency	16Khz
Frame length(ms)	20
Frame shifting(ms)	10
window function	Hamming window
LPC order	12
MFCC order	16
Δ-Δ frame number	3

3.3. Endpoint Detection Algorithm

The schematic diagram of endpoint detection algorithm is shown in Figure 3. In training stage, the voice played in Cooledit software as the endpoint detection source. The gotten voice and noise signal should be pre-emphasis firstly and then framing and add-windows analysis. Secondly, extraction the time domain energy features, LPC, Δ-ΔLPC, MFCC, Δ-ΔMFCC features, then normalization the feature date except the time domain feature. Because the time domain feature after normalization didn't have the separability for the voice and noise signals. According to the supervised way to training, the SVM training model would be gotten. In testing phase, the feature extraction would be done after the pre-emphasis processing frame by frame of the voice signal sequence to be detected. Then put the feature into the model gotten from the training stage, the testing results would be gotten after processing frame by frame. The endpoint would be output after filtered the testing results.

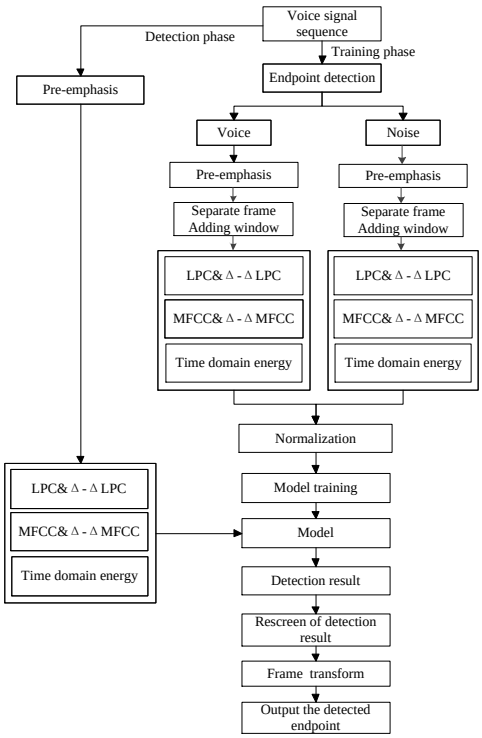

Figure 3. The Endpoint Detection Algorithm

3.4. Introduction of ELM Algorithm

ELM is a new algorithm for the single hidden-layer feedforward neural networks (SLFN). The algorithm generated the connection weights between the input layer and hidden layer and the thresholds of the neurons in the hidden layer randomly. No adjustments in the training process would be needed. Only the number of neurons in the hidden layer needed to be set, the unique solution could be gotten. Compared with conventional learning algorithms (such as BP algorithm), ELM have the advantages as high learning speed, good generalization performance. Typical SLFN can be described as shown in Figure 4, the network formed by the input layer, hidden layer and output layer. The hidden layer and output layer is fully-connected [8].

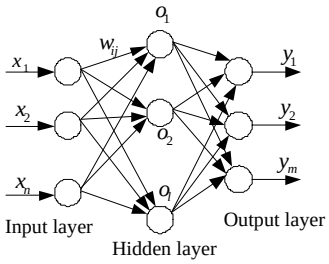

Figure 4. The Typical Structure of SLFN

Known training samples $\{(x_i, t_i)\}$, $x_i \in R^p$, $t_i \in R^q$, $i = 1, 2, ..., N$ contains L hidden layer nodes and excitation functions for standard single hidden-layer feedforward neural networks in mathematical model could be expressed in the form of formula (5).

$$\sum_{i=1}^{L} \beta_i f(w_i x_j + b_i) = O_j, \; j = 1, ..., N \tag{5}$$

In the formula, β_i stand for the output weights of i-th hidden layer node field connecting output neurons, w_i stands for the offset of i-th hidden layer node, O_j stands for the output of the j-th enter sample. There are L hidden layer nodes, the excitation function $f(x)$ of standard single hidden layer feedforward neural network could be zero-error approach to N samples as $\sum_{j=1}^{L} \| o_j - t_j \| = 0$, β_i, w_i, b_i makes the formula (6) established.

$$\sum_{i=1}^{L} \beta_i f(w_i x_i + b_i) = t_j, j = 1,..., N \tag{6}$$

The formula (6) can be expressed as $H\beta = T$. Among them, H is called the hidden layer output matrix of neural network. The i-th column of H represents the output vector when i-th hidden layer node's input is $X_1, X_2, ..., X_N$.

$$\begin{aligned} &H(w_1, ..., w_L, b_1, ..., b_L, x_1, ..., x_N) = \\ &[f(w_1 x_1 + b_1) ... f(w_L x_1 + b_L); \\ &...; \\ &f(w_1 x_N + b_1) ... f(w_L x_N + b_L)]_{N \times L} \\ &\beta = [\beta_1^T ... \beta_L^T]_{L \times M}^T \\ &T = [t_1^T ... t_N^T]_{N \times M}^T \end{aligned} \tag{7}$$

ELM algorithm described as follow: known training samples $\{(x_i, t_i)\}$, $i = 1, 2, ..., N$ hidden layer nodes' number is N. The learning algorithm of SLFN's ELM have three-step process with excitation functions of $f(x)$:

Set the input weights w_i and offset b_i;

Calculate the output matrix *H*;

Calculate output weights β: $\beta = H^2 T$.

In training stage of algorithm, the input x_i is the eigenvector of voice frame, the output y_i is the label of voice frame (voice tags is 1, noise tags is 0); In identify phase of algorithm, the algorithm would output the label y_i of voice frame use the mode gotten from training phase when the eigenvector x_i of voice was inputted.

3.5. Introduction of Support Vector Machines

Support Vector Machine (SVM) was a machine learning algorithm which introduced by Vapnik based on statistical theory and take structural risk minimization principle. This algorithm shows many unique advantages in solving the small sample, nonlinear and multidimensional pattern recognition problems. The fundamental thought of the algorithm is translate the input space into high-dimensional feature space and finding an optimal hyperplane in higher dimensional space make the category interval maximum. However, SVM have many limitations, such as the option of kernel, the setting of penalty factor c and kernel parameter γ, the influence of the SVM's training speed by the size of the training set. Genetic Algorithms is often used to select the penalty factor and kernel parameter. Meanwhile the particle swarm optimization or other modern intelligent algorithms are used for parameter optimization. The search algorithm of SVM optimized by genetic algorithm (GA-SVM) as shown in Figure 5. At beginning, setting the maximum evolution algebra, the maximum population number, the variation range of parameter c and γ, the times of cross validation. Then initialize the algorithm, the colony began fitness degrees calculation, selection, crossover and mutation. The calculation stopped when the maximum evolution algebra is reached. The SVM is impemented base on libsvm, the SVM forecast accuracy is used as the fitness function. The algorithm output the optimal parameter, punishment factor c and kernel parameter γ.

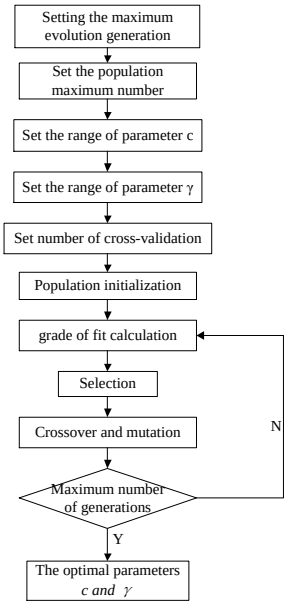

Figure 5. The Flowchart of GA-SVM Algorithm

Using genetic algorithm for SVM optimization is to get the optimum training model during the training phase and make the output of SVM more accurate and reliable in prediction phase. In training phase, the inputs of SVM are the eigenvectors and the labels of speech frames. In identification phase, the input is the eigenvector of speech frame; the output is the label of speech frame [9].

3.6. The Rescreen of Detection Results

There would be some errors in the judgment result because the algorithm is judged by frame. The designed filtering rules for erroneous judgments of the noise to speech labels is shown in Figure 6. The frame judging results needs to be gotten first. In this article the voice label was assigned to 1, the noise label was assigned to 0. The test results in Figure 3 did dislocation subtract, the low to high transition (1-0>1) of dislocation subtract indicate the beginning of speech, the high to low transition (0-1<-1) indicate the end of the speech. Through the loop searching, the low to high transit frame and the high to low transit frame were set to the starting frame and the end frame when the frames between the positive and negative jump were greater than 4. Until the cycle number reached the maximum one, the starting and end point of voice could be gotten after the frame converting of filtered frame labels.

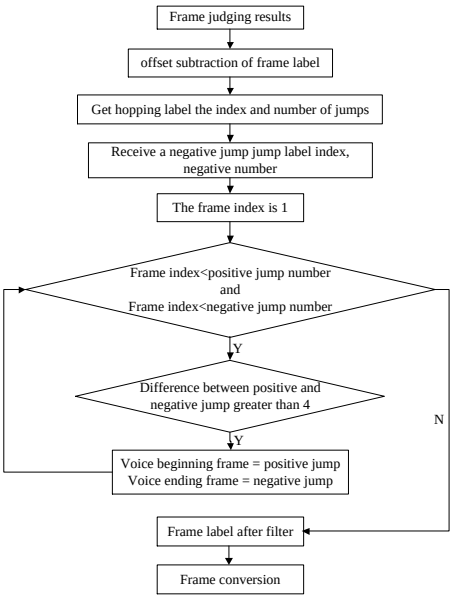

Figure 6. Rescreen Rules of Survey Results

4. Endpoint Detection Algorithm Performance Evaluations

The software and hardware platforms of SVM and ELM experiment are as follows: CPU E7500 (Duo 2.93GHz), 4GB of RAM, the operating system is Windows7 Ultimate Edition 32bit, the simulation environment is Matlab 2009a.

4.1. The Training of SVM

Token the voice of ONE to NINETEEN from 4 different people's pure English speech database as training samples. The voice sequences and noise sequence is gotten from the actual playback of voices pecimen. Take the parameters in Table 1 and Table 2 doing the SVM training which optimized by genetic algorithm, the parameters of SVM can be gotten as shown in Table 2.

Table 2. The Parameter Table of SVM Optimized By Genetic Algorithm

Parameters	Value
maximum evolution generation	100
The maximum number of population	20
c ranges	[0,100]
γ ranges	[0,100]
Number of cross-validation	5
Kernel type	RBF
Optimal value c	4.1154
Optimal value γ	0.1004
Training time(s)	249.3364
Accuracy of training (%)	99.9023

4.2 The Training of ELM

The relationship between the number of hidden layer and the training time and accuracy is shown in Table 3. Different hidden layer number would get different accuracy as shown in Table 3 [11]. But the maximum training time is only 1/2000 of the SVM training time. To compare the training effect of ELM between different hidden layer

numbers, the accuracy/training time is measured as performance metric. The metric's value is bigger the training effect is better and vice versa. Apparently, the performance is optimum when the hidden layer number is 55,

Table 3. The Relationship Between Training Time and Accuracy of Different ELM's Hidden Layers

Hidden layers	Training time(s)	Accuracy (%)	Accuracy/Training time
35	0.0780	99.3164	12.7329
40	0.0624	99.4141	15.9317
45	0.1248	99.1211	7.9424
50	0.0936	99.2188	10.6003
55	0.0468	99.707	21.3049

4.3 The Comparison of Endpoint Detection Effectiveness

Figure 7 is the endpoint detection effectiveness comparison figure. Figure 7a is the time domain waveform of three English words "TWO", "THREE", and "FOUR"; Figure 7b is the corresponding zero crossing rates figure; Figure 7c is endpoint detection waveform of the traditional time domain energy and zero crossing rates; Figure 7d is the frame label detection results of GA-SVM training model; Figure 7e is the frame figure after filtered by "section 2.6 The rescreen of detection results"; Figure 7f is the endpoint detection waveform from GA-SVM training mode; Figure 7g is the frame label detection results of ELM training model; Figure 7h is the frame figure after filtered by "section 2.6 The rescreen of detection results"; Figure 7i is the endpoint detection waveforms from ELM training model.

From Figure 7a, the voice signal sustained 2.5s. It is needed to notice that there have miscalculation in the 3-th, 14-th, 15-th, and 21-th to 26 frame in Figure 7d and 7-th to 8-th, 20-th to 25-th frame n Figure 7g. Compared with the play efforts of the COOLEDIT software could found that the part which have been cut off in Figure 7f and Figure 7h compared with Figure 7a are noise signals (airflow sound, non-voice). Use the same method could confirm that the parts which have been cut off in Figure 7f and 7i are noises too. In comparison, the endpoint detection in Figure 7f and 7i is more strictly than figure 7c. It is said that ELM and GA-SVM endpoint detection method's effort is better than traditional time domain energy and zero crossing rate method.

To test algorithm's robustness, added some different signal noise rate Gaussian white noise into signal before pre aggravated. Used the correct rate, virtual seized rate and missed rate as the algorithm's performance evaluation standard. The detection result of three algorithms as shown in Table 4. From the data in the table, the difference between the detection times of three different algorithms is slightly. But with the SNR reduced, the accuracy of short time energy and zero crossing rate detection method is reduced sharply; The accuracy of the ELM and SVM detection algorithm is consistent, and didn't change significantly when SNR changing. The algorithm shows a certainly robustness.

Table 4. The Recognition Rate of the Three Voice Endpoint Detection Algorithm

SNR(db)	ELM			SVM			Short Energy and Zero Crossing Rate		
	Accuracy	Virtual seized rate	False detection rate	Accuracy	Virtual seized rate	False detection rate	Accuracy	Virtual seized rate	False detection rate
-5	98.12	1.45	0.43	98.67	0.138	1.192	89.84	4.26	5.9

0	97.79	1.6	0.61	99.16	0.07	0.77	91.78	3.13	5.09
10	98.26	1.23	0.51	97.86	0.19	1.95	90.46	3.61	5.93
15	98.41	1.27	0.32	98.43	0.178	1.392	93.88	5.76	0.36
Pure voice	99.67	0.33	0	99.56	0.21	0.23	95.78	2.06	2.16

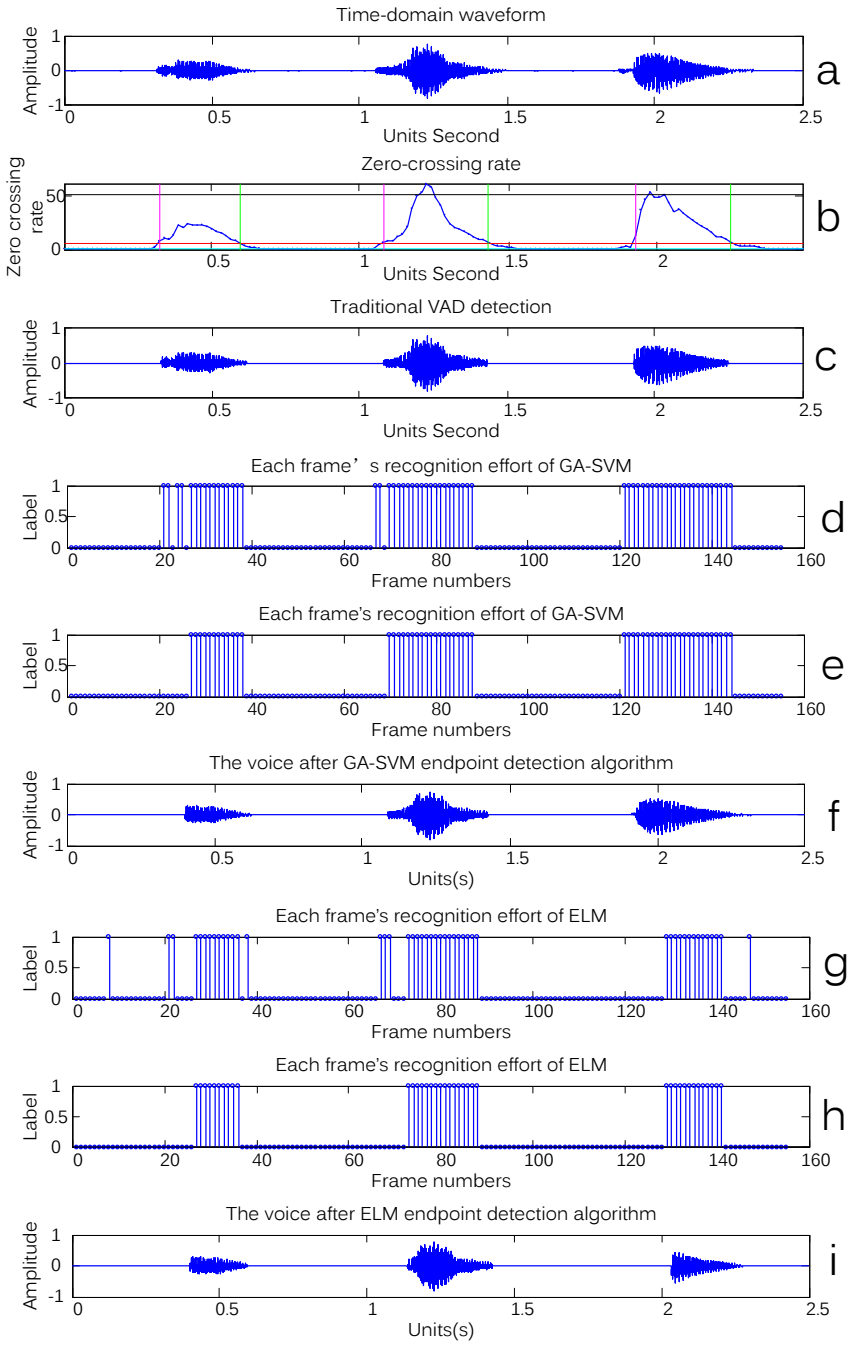

Figure 7. The Endpoint Detection Efforts Comparison Chart

5. Conclusions

In this paper, a new endpoint detection algorithm is proposed which takes the speech endpoint detection as a binary classification in pattern recognition. Experimental simulation showed that the algorithm are greatly improved on the endpoint detection accuracy and robustness to noise compared with traditional method of short time energy and zero crossing detection. The recognition time is almost same between ELM and GA-SVM, but the recognition accuracy, reliability and the training time of ELM is better than GA-SVM's. In this algorithm, considered the difference between the voice and noise when extract the features, set the frame filter rules. This algorithm performs a better robustness to different noise signal.

With the development of speech feature extraction and pattern recognition technologies, the feature extraction module and pattern training and recognition module could be replaced with the latest model in this algorithm. The voice detection algorithm is a general endpoint detection algorithm, can be used in other signal's endpoint detection method.

Acknowledgements

The authors would like to thank the financial support of the financial support of Natural Science Foundation of Liaoning 2013024010.

References

[1] X.-J., J.-F. Yuan, J.-F. Luan and L.-J. Huang, "An end point detection method based on short term energy and high order difference", Journal of Beijing normal university (natural science), vol.48, no. 2, **(2012)**, pp. 324-334.
[2] C.-L. Zhang, X.-Y. Zeng and S.-G. Wang, "A voice activity detection algorithm based on the variance of critical band power spectrum", Technical Acoustics, vol.31, no. 2, **(2012)**, pp. 204-207.
[3] X.-L. Zhang, J. Wu and P. Lv, "Support vector machine based vad using multiple observation compound feature", J Tsinghua Univ(Sci & Tech), vol.51, no. 9, **(2011)**, pp. 1209-1214.
[4] H.-Z. Wang, Y.-C. Xu and M.-J. Li, "Voice activity detection algorithm based on mel frequency sepstrum coefficient (MFCC) similarity", Journal of Jilin University (Engineering and Technology Edition), vol. 42, no.5, **(2012)**, pp.1332-1334.
[5] J.- F. Wang, "GUO Ming. Research on VAD feature of exponent function warping group delay function", Journal of Jilin University (Engineering and Technology Edition), vol. 43, no., **(2013)**, pp. 435-438.
[6] J. W. Shin, J.-H. Chang and N. S. Kim, "Voice activity detection based on statistical models and machine learning approaches", Computer Speech and Language, vol. 24, no.3, **(2010)**, pp.515-530.
[7] P.-S., S.-M. Lee, "Speech enhancement through voice activity detection using speech absence probability based on Teager energy", Journal of Central South University. vol. 20, no. 2, **(2013)**, pp. 424-432.
[8] A.M. Aibinun, M.J.E. Salami and A.A. Shafie, "Artificial neural network based autoregressive modeling technique with application in voice activity detection", Engineering Applications of Artificial Intelligence. vol. 25, no. 6, **(2012)**, pp. 1265-1276.
[9] S.-H. Chen, Rodrigo, C. Guido, T.-K. Truong and Y. Chang, "Improved voice activity detection algorithm using wavelet and support vector machine", Computer Speech and Language. vol. 24, no. 3, **(2010)**, pp. 531-543.
[10] Y. Wan, H.-Q. Tong and Y.-Y. Zhu, "Parameters optimization of multi kernel support vector machine based on genetic algorithm", Wuhan Univ.(Nat.Sci.Ed.), vol. 58, no. 3, **(2012)**, pp. 255-259.
[11] H.-J. Lu, J.-W. Zhang, X.-P. Ma and W.-B. Zheng, "Tumou classification using extreme learning machine ensemble", Mathmetics in practice and theory, vol. 42, no. 17, **(2012)**, pp.148-151.

Authors

Zhigang Feng, he is an associate professor of Shenyang Aerospace University. He received his Doctor's Degree from Harbin Institute of Technology, P.R. China at 2009. His main research direction includes system fault diagnosis, self-validating sensor and self-validating actuator.

Junlei Feng, he received the bachelor degree of Automatization from Shenyang University, China, in 2013. He is currently study for a master degree in the School of Automation, Shenyang Aerospace University.