Advances in Computer Science Research, (ACSR), volume 52
2016 International Conference on Computer Engineering and Information Systems (CEIS-16)

Speech Recognition for Finite Passwords in Smart Home

Jun Yang, Cui-Yun Gao, Ming Zhang
School of electronic and information engineering, Anhui Jianzhu University, Hefei, China
E-mail: yang_jun_ahjzu@163.com, gaocuiyun@ahjzu.edu.cn, zm1128@ahjzu.edu.cn

***Abstract*-Dynamic Time Warping (DTW) is the most widely used algorithm for finite passwords or isolated words recognition system. However, the recognition rate using DTW is strongly influenced by the precision of voice activity detection(VAD), and the traditional DTW algorithm has more path to search when compared with the improved DTW algorithm will cost more time. To solve the problem, a novel recognition method for limited passwords using a new VAD algorithm of improved short-time Teager energy with double dynamic thresholds combined with the improved DTW algorithm for limited passwords or isolated words recognition is proposed in this paper. Experiment results show that the proposed method has better noise robustness and recognition rate than the short-time energy and zero-crossing rate with DTW or the improved DTW algorithm. Because of using the improved DTW, it is also real-time.**

Keywords-finite passwords; VAD; improved DTW; teager energy

I. INTRODUCTION

The influence of noises in house often result in low recognition rate for finite passwords. Therefore, it's necessary to research a kind of speech recognition algorithm with strong robustness to noise, high speed and high recognition rate in the real use for smart home.

To realize the limited password or the closed set of isolated word recognition, DTW algorithm shows almost the same recognition rate as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) and Artificial Neural Network (ANN) algorithms [1,2,3]. However, the search path of DTW contains a whole rectangular area made of the template of reference and test. That lead to too much calculation and low recognition speed. The improved DTW algorithm will limit the search path between two parallel lines[4], which will shorten the recognition time while the recognition rate only drops a little.

As the pre-processing part of DTW algorithm, the detection accuracy rate of voice activity detection (VAD) will effect on recognition rate directly. The most common VAD algorithm is short time energy and Zero Crossing Rate (ZCR)[5], but the algorithm has poor robustness to noise. In recent years, some kinds of VAD algorithms based on the spectral entropy[6], cepstral distance[7], Teager Energy Operator (TEO)[8], Wavelet and Wavelet packet transform[9], Hilbert huang transform (HHT)[10] and Artificial neural network (ANN)[11] were proposed by researchers. Among these methods, the spectral entropy and cepstral distance algorithm still have high false alarm rate or miss rate under low Signal to noise ratio (SNR) environment. Though the detection accuracy rate is higher when the VAD algorithm adopts Wavelet and Wavelet packet transform, Hilbert huang transform (HHT) and Artificial neural network (ANN), but it also cost longer time. The VAD based on TEO is simple, which only processes the signal in time-domain. And this algorithm has the capabilities of eliminate zero-mean noise and enhance voice. With these advantages, the VAD algorithm base on the short-time TEO with double threshold is proposed by LI Jie[8]. The method can balanced the detection rate and calculate time, but some voiced parts which have lower energy may be missed.

Thus, this paper proposes an improved short-time Teager energy and improved DTW algorithm for limited passwords recognition.

II. PRINCIPLE OF VOICE RECOGNITION

The basic structure of limited passwords recognition system[4] is showed as Figure1,including the pre-processing module, feature extraction module, speech templates module, pattern matching module, decision rules and so on.

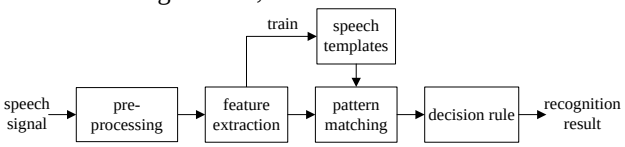

Figure 1. The basic structure of limited passwords recognition system.

On the basis of reference [8] method, the improved short-term Teager energy with double threshold method for VAD is proposed in this paper.

A. Improved Short-term Teager Energy with Double Threshold

For discrete-time signal $x(n)$, TEO is defined as:

$$\Psi(x(n)) = x(n)^2 - x(n-1)x(n+1) \tag{1}$$

There are two thresholds when doing VAD according to LI Jie's paper. One threshold is the relatively low and it is sensitive to the changes of signal. Another threshold is relatively high. The threshold can be exceeded only when the signal reaching a certain intensity. The two thresholds are calculated as:

$$T_1 = \max(k_1 * tmx + (1-k_1) * Tmn, k_2 * Tave + (1-k_2) * tmn) \tag{2}$$

$$T_2 = \max(T_1 + k_3 * (tmx - tmn), T_1 + k_4 * (Tave - Tmn)) \tag{3}$$

Tave and *Tmn* are the average and minimum energy of TEO for the entire signals. The *tmx* and *tmn* are maximum and minimum energy of TEO in the first continuous 0.125s section and the last continuous 0.125s section of the signal. In the two formulas, k_1 = 0.035, k_2 = 0.010, k_3 = 0.400, k_4 = 0.025.

The VAD algorithm in [8] gives the result of detecting voiced segments by double threshold detection. When a password contains several syllables, it maybe miss some voiced parts which have lower energy. The improved algorithm can judge the voiced segment after double threshold detection. If the detected two adjacent voice segments are less than 0.15s, the two voice segments are merged into one voice segment. The maximum length of silence and noise are also need to filter the burst noise and prevent from missing the voiced parts of low energy[12]. Through a large number of experiments, we takes 12 frames as the maximum length for silence and 10 frames for noise in this paper.

B. *Feature Extraction and Pattern Matching*

The coefficients including 12 order Mel Frequency Cepstrum Coefficients (MFCC) and its delta coefficients are

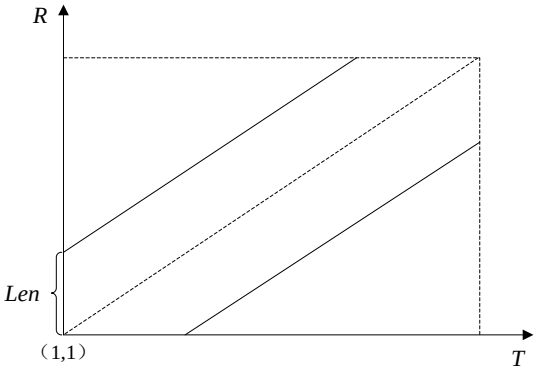

Figure 2. Improved DTW algorithm path.

adopted as the features in this paper. Each password template is calculated by the method of reference [13]. Using the improved DTW algorithm shown in Figure 2 do pattern matching. In this paper, *Len* equals 26.

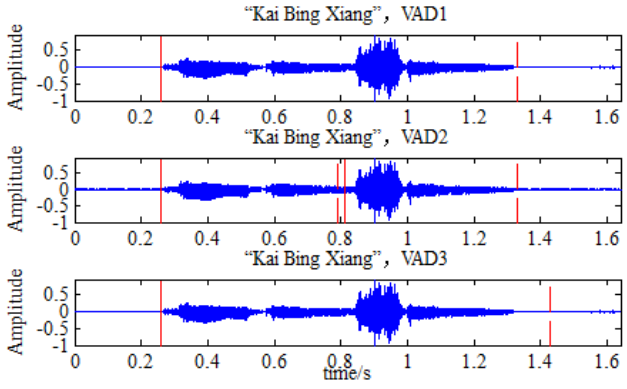

Figure 3. "Kai Bing Xiang", three methods' VAD effect.

III. EXPERIMENTAL DATA

Experimental platform is a laptop with 2.3GHz CPU Clock Speed and 4.00GB RAM. The operating system is Windows 7, 32 bits. The software simulations are performed using Matlab R2011b. The experimental recording tool is laptop or mobile phone. Sampling rate is no less than 16KHz, and quantization digit is 16 bits.

There are 8 kinds of Chinese passwords which are: Kai Men (open the door), Guan Men (shut the door), Kai Deng (turn on the light), Guan Deng (turn off the light), Kai Bing Xiang (open the refrigerator), Guan Bing Xiang (close the refrigerator), Kai Kong Tiao (turn on the air conditioning) and Guan Kong Tiao (turn off the air conditioner).We recorded 35 persons(including men and women)'s speech data, each kind password were recorded for four times. And the recording time of each password is not less than 1.5s. Every password has 140 recordings.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The experiment adopts 4 methods, comparing the effect of these four kinds of method. Method 1 uses improved Teager energy with double threshold method combined with DTW algorithm. Method 2 is the short-time energy and short-time zero crossing ratio combined with traditional DTW algorithm. Method 3 is improved Teager energy double threshold method combined with improved DTW algorithm. Method 4 is short-time energy and short-time zero crossing ratio combined with improved DTW algorithm.

In the short-time energy and short-time zero crossing rate of VAD, three threshold must be set, the energy of high and low threshold and the rate of zero crossing threshold[5]. The three thresholds are calculated as:

$$E_1 = \alpha_1 * Det + eth \tag{4}$$

$$E_2 = \alpha_2 * Det + eth \tag{5}$$

$$Zcr = \alpha_3 * Zcrth \tag{6}$$

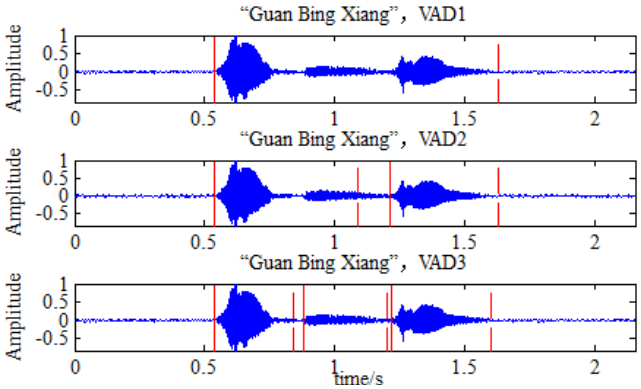

Figure 4. "Guan Bing Xiang", three methods' VAD effect.

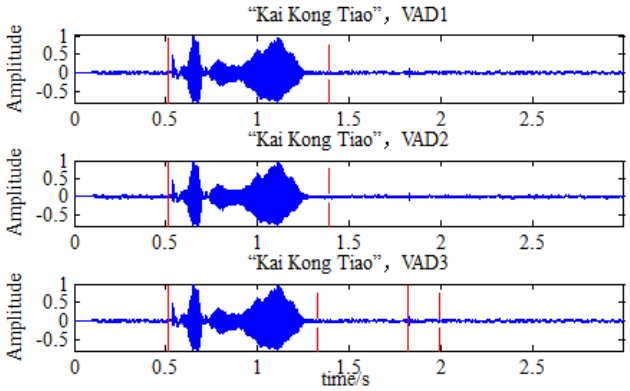

Figure 5. "Kai Kong Tiao", three methods' VAD effect.

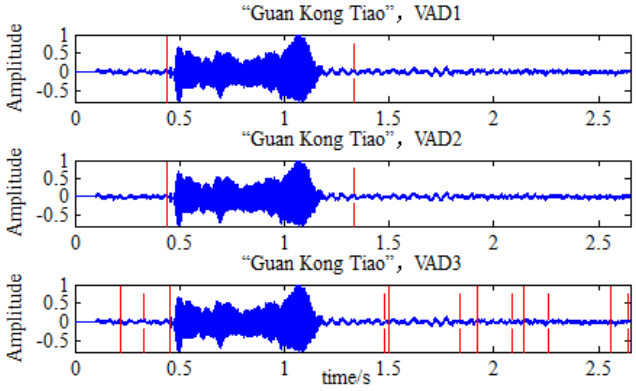

Figure 6. "Guan Kong Tiao", three methods' VAD effect.

TABLE I. THERECOGNITION EFFECT OF METHOD 1 AND 2

speech passwords		Kai Men	Kai Deng	Kai Bing Xiang	Kai Kong Tiao	Guan Men	Guan Deng	Guan Bing Xiang	Guan Kong Tiao	Mean Value
recognition rate	method 1	57.62%	85.71%	80.00%	88.10%	63.33%	70.00%	93.34%	81.90%	77.50%
	method 2	57.62%	80.48%	72.38%	77.62%	60.00%	70.00%	75.71%	86.19%	72.50%
recognition time	method 1	257ms	214ms	296ms	267ms	221ms	208ms	292ms	266ms	253ms
	method 2	223ms	191ms	267ms	227ms	205ms	190ms	260ms	233ms	224ms

TABLE II. THE RECOGNITION EFFECT OF METHOD 3 AND 4

speech passwords		Kai Men	Kai Deng	Kai Bing Xiang	Kai Kong Tiao	Guan Men	Guan Deng	Guan Bing Xiang	Guan Kong Tiao	Mean Value
recognition rate	method 3	60.95%	80.71%	75.72%	83.57%	66.91%	68.81%	92.15%	82.15%	76.13%
	method 4	59.05%	81.43%	72.38%	76.19%	57.62%	69.04%	75.71%	84.76%	72.02%
recognition time	method 3	195ms	168ms	232ms	210ms	179ms	170ms	232ms	213ms	200ms
	method 4	168ms	145ms	200ms	175ms	155ms	148ms	195ms	178ms	171ms

eth is the average energy in the first continuous 0.125s section and the last continuous 0.125s section of the signal. *Det* is the difference value of the maximum energy in entire signal with the *eth* . *Zcrth* is the average zero crossing rate in entire signal. $\alpha_1 = 0.002$, $\alpha_2 = 0.003$, $\alpha_3 = 0.400$.

Figure 4 to 7 shows the effect of three VAD methods. VAD1 stands for the proposed VAD method. VAD2 stands for the short-time TEO with double threshold. VAD3 stands for the short-time energy and zero-crossing rate. The solid line indicates the start position of the voiced segment and the broken line indicates the end of voiced segment position. We can see that the improved Teager energy with double threshold method has better detection effect than other two VAD methods. Short-time energy and Zero Crossing rate has the worst detection effect.

70 samples for were chosen randomly from 140 samples of each kind of password as training set, the remaining data for testing, the experiments were done three times for different training set and testing set. The average recognition rate and time with different method of three times are recorded in Table 1 and 2.

From Table 1 to Table 2, we can find that method 1 has higher average recognition rate than method 2, and method 3 has higher average recognition rate than method 4. The reason is that the use of improved Teager energy with double threshold has higher VAD accuracy rate than short-term energy and short-term zero-crossing rate. Comparing method 3 with method 1, method 2 and method 4, the passwords recognition time is reduced more than 20% by using improved DTW instead of DTW. In summary, the method 3 achieved relatively high recognition rate and shorter recognition time.

V. CONCLUSION

An improved short-term Teager energy of VAD combined with improved DTW algorithm for finite passwords recognition is proposed in this paper. We compared the recognition rate and recognition time of improved short-time Teager energy with double thresholds, short-time energy and zero-crossing rate with DTW and improved DTW respectively. We chose 8 kinds of passwords of 35 people and the total number of samples is 1120. Experiments showed that the recognition rate is relatively

high and the recognition time is shorter when using the proposed method of this paper.

REFERENCES

[1] Cui X, Afify M, Gao Y, et al. Stereo hidden Markov modeling for noise robust speech recognition[J]. Computer Speech & Language, 2013, 27(2): 407-419.

[2] Ting H N, Yong B F, Mirhassani S M. Self-adjustable neural network for speech recognition[J]. Engineering Applications of Artificial Intelligence, 2013, 26(9): 2022-2027.

[3] Kim C, Seo K. Robust DTW-based recognition algorithm for hand-held consumer devices[J]. Consumer Electronics, IEEE Transactions on,2005, 51(2): 699-709.

[4] ZHANG Jun, Li Xue-bin. A Method of Isolated Word Recognition Based on DTW [J]. Computer Simulation, 2009 (10): 348-351.

[5] Zhi-Yong Song. The Application of MATLAB in Speech Signal Analysis and Synthesis[M]. BEIHANG UNIVERSITY PRESS, 2013.

[6] Wu B F, Wang K C. Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments[J]. Speech and Audio Processing, IEEE Transactions on, 2005, 13(5): 762-775.

[7] Kinnunen T, Chernenko E, Tuononen M, et al. Voice activity detection using MFCC features and support vector machine[C]//Int. Conf. on Speech and Computer (SPECOM07), Moscow, Russia. 2007, 2: 556-561.

[8] Li J, Zhou P, Jing X, et al. Speech endpoint detection method based on TEO in noisy environment[J]. Procedia Engineering, 2012, 29: 2655-2660.

[9] Eshaghi M, Mollaei M R K. Voice activity detection based on using wavelet packet[J]. Digital Signal Processing, 2010, 20(4): 1102-1115.

[10] Zhi-mao L, Hui J, Chun-xiang Z. Voice activity detection in complex environment based on Hilbert-Huang transform and order statistics filter[J]. Journal of Electronics & Information Technology, 2012, 34(1): 213-217.

[11] Aibinu A M, Salami M J E, Shafie A A. Artificial neural network based autoregressive modeling technique with application in voice activity detection[J]. Engineering Applications of Artificial Intelligence, 2012, 25(6): 1265-1276.

[12] Jinping H, Ruozhu C, Zhanming L. Discussion of improved DTW algorithm in speech recognition[J]. Microcomputer & Its Applications, 2011, 3: 012.

[13] Zhang J, Qin B. DTW speech recognition algorithm of optimization template matching[C]//World Automation Congress (WAC), 2012. IEEE, 2012: 1-4.