Isolated Word Speech Recognition System Based On FPGA

Xiaohui Hu[1]
1 School of Physics and Telecommunication, South China Normal University, Guangzhou,P.R.China
xiaohui_huhu@sina.com

Haolan Zhang[2], Lvjun Zhan[1], Yun Xue[1], Weixing Zhou[1], Gansen Zhao [3]
2 NIT,ZheJiang University,Ningbo,P.R.China
haolan.zhang@gmail.com

3 School of Computer Science, South China Normal University,Guangzhou,P.R.China,510006
zhaogansen@gmail.com

***Abstract*—The paper introduces an isolated word speech recognition system in which the speech signal is acquired in real time. Half Raised-Sine function is applied to the MFCC parameters of the audio files, and improved DTW algorithm is implemented. Simulation results show that compared the conventional DTW with the improved DTW algorithm, the latter can obtain a better recognition rate and faster response time. Finally, the system was implemented on FPGA and yields satisfactory performance.**

***Index Terms*—the isolated word, MFCC, Half Raised-Sine function(HRSF), DTW, FPGA**

I. INTRODUCTION

Automatic Speech Recognition is technology that allows a computer to identify the words that a person speaks into a microphone or telephone and convert it to written text. A fundamental problem of speech recognition is a reasonable selection of phonetic features. Linear prediction (LP) analysis is widely used in speech feature parameter extraction technology. There are many successful application systems built based on LP technology. However, linear prediction model is a pure mathematics model without consideration on processing features of human auditory system. Other techniques used are: Linear Predictive Cepstral Coefficients (LPCC); Perceptual Linear Prediction (PLP); Mel-Frequency Cepstral Coefficients (MFCC); and Neural Predictive Coding (NPC) [1], [2], [3]. MFCC is a popular technique because it is based on the known variation of the human ear's critical frequency bandwidth. MFCC coefficients are obtained by de-correlating the output log energies of a filter bank which consists of triangular filters, linearly spaced on the Mel frequency scale. There have been many attempts to enhance the robustness of MFCC features. For example, cepstral mean and variance normalization (CMVN) [4], RASTA filtering [5], temporal structure normalization [6], feature warping [7], and MVA processing [8], HRSF[9]（Half Raised-Sine function） are commonly used for enhancing MFCC robustness against additive noises and channel distortions.

Dynamic time warping (DTW) algorithm has been widely used in speech recognition, especially the speaker-independent speech recognition template matching processing. DTW is a simple and effective method in speech recognition. The algorithm is based on dynamic programming, solving the problem of the random timeline in the various parts of each word. However, the DTW algorithm needs test voices to match all the models, and then to find the most similar model corresponding to the speaker as a recognition result. The more models are stored, more time is needed for recognition. So improving the matching speed is a crisis matter for the application of speech recognition system. This problem can be addressed by considering an implementation strategy which uses the parallel processing paradigm to provide the opportunity for reducing processing times required. Recently, parallel techniques[9]-[11] have been proposed for its implementation in order to achieve real time speech recognition, taking advantage of VLSI technology.

This paper is organized as follows: In Section2, the parameter feature method MFCC and HRSF are explained. Section3 deals with DTW algorithm and implements the improved DTW algorithm. The whole system completed on FPGA is introduced in section4 which including the simulation test, FPGA implementation and the final experimental results on speaker-independent and speaker-dependent speech recognition. Section5 outlines the conclusion and future work.

II. FEATURE EXTRACTION METHOD

A. Mel-Frequency Cepstral Coefficients (MFCC)

For each tone with an actual frequency f measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale.

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (1)$$

Where f_{mel} is the subjective pitch in Mels corresponding to a frequency in Hz. This leads to the definition of MFCC, a baseline acoustic feature set for speech and speaker recognition applications.

MFCC coefficients are a set of Discrete Cosine Transform(DCT) decorrelated parameters, which are computed through a transformation of the logarithmically

doi:10.4304/jcp.8.12.3216-3222

compressed filter-output energies, derived through a perceptually spaced triangular filter bank that processes the Discrete Fourier Transformed (DFT) speech signal.

$$H_m(n), m = 0,1,\cdots,M-1, n = 0,1,\cdots,\frac{N}{2}-1 \quad (2)$$

Where m is the number of filters, N is the points of the frame of the speech. The filter in the frequency domain is simple triangle, whose center frequency is f_m, which distributes uniformly in the Mel frequency axis. Band-pass filter parameters should be calculated beforehand, which can be used directly in the calculation of MFCC parameters. The filterbank designed in this paper contains 24 filters, the speech signal frame length is 256, signal sampling frequency is 8KHz, the filterbank waveform is stated as follows:

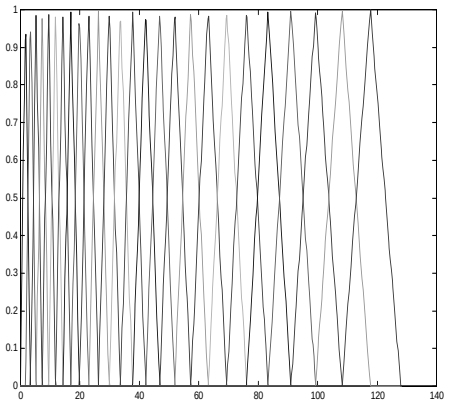

Figure1. The waveform of filterbank

The computing of MFCC parameter is shown as following:

(1) Identify the points of each frame of the speech sample sequence, in this paper N = 256. Obtain discrete power spectrum X (n) after discrete FFT transform.

(2) Calculate the power value of X (n) through M filters, get M parameters $P_m, m = 0,1,\cdots,M-1$.

(3) Calculate of the natural logarithm of P_m, then we have $L_0, L_1, \cdots, L_{m-1}$

(4) After the discrete cosine transform of $L_0, L_1, \cdots, L_{m-1}$ get $D_m, m = 0,1,\cdots,M-1$

(5) Discard D_0 take $D_1, D_2, \cdots, D_k$ as the MFCC parameters.

B. Half Raised-Sine Function

The standard MFCC parameters only reflects the static characteristics voice parameters, but a human ear is more sensitive to the dynamic characteristics of the voice. It is possible to obtain more detailed speech features by using a derivation on the MFCC acoustic vectors. This approach permits the computation of the delta MFCC (DMFCCs), as the first order derivatives of the MFCC. Then, the delta-delta MFCC (DDMFCCs) are derived from DMFCC, being the second order derivatives of MFCCs.

Differential parameter is calculated as the following:

$$d(n) = \frac{1}{\sqrt{\sum_{i=-k}^{k} i^2}} \sum_{i=-k}^{k} i \bullet c(n+i) \quad (3)$$

Where c and d are expressed in a speech frame parameter, K is a constant, usually 2, then the differential parameter is the linear combination of parameters of the two frame ahead and the following two frame. In practical, the MFCC parameters and the differential order component parameters are combined into a vector, as the speech parameters of one frame.

Researches prove that the contribution for recognition rate of each component in the feature vector is different. In speech recognition, higher order MFCC components are more susceptible to the influence of noise than the lower order MFCC components, so if we put half raised sine function in use can add weight to smaller high-order component, and reduce weight of the low order component which could be easily interfered by noise.

The formula is shown as following.

$$\overline{c_i} = r_i c_i \quad (4)$$

$$r_i = 0.5 + 0.5\sin(\pi i / L) \quad (5)$$

Where i=0，1… L-1; L is the order of the feature.

III. THE ALGORITHM OF DYNAMIC TIME WARPING

A. The Conventional DTW Algorithm

Supposes reference template have M frame vector , $A_1, A_2, \ldots A_M$, A_M is the M_{th} feature vector of speech, and the testing template have N feature vector, $B_1, B_2, \ldots B_N$, B_N is the N_{th} feature vector of speech. $D(B_{(in)}, A_{(im)})$ is used to indicate the distance between the M_{th} feature in A and N_{th} feature in B.

The distance is usually expressed by Euclidean Distance. DTW algorithm is to find a warping function $i_m = W(i_n)$ which will make the timeline n of testing template mapped to the timeline m of reference template non-linear and enable the function to meet the equation as follow.

$$D = \min_{\Phi(i_n)} \sum_{i_n=1}^{N} d(B(i_n), A(W(i_n))) \quad (6)$$

When the A and B are identical, they are mapped to be a straight line whose slope is 1, when the A and B aren't identical, in order to make the M_{th} sample in A and N_{th} sample in B alignment, the corresponding point is not in a straight line, but forms a curve, this curve corresponds to a function that is the Warping Function. If each frame number of the testing template is marked on the horizontal axis in a two-dimensional coordinate system and the reference template marked on the ordinate axis, we can draw a vertical and horizontal line which can form a network through the integer coordinate. In this network, each spot (n, m) expresses the intersection of the frame in the testing template and the frame in the reference template. The DTW algorithm may sum up to seek for a way through a number of grid points, which is the frame

number which will calculate the frame distance between the two templates.

In practice, the function $W(n)$ is restrained:

$$W(1) = 1, W(n) = M \tag{7}$$

$$W(n) - W(n-1) = \begin{Bmatrix} 0,1,2 & W(n) \neq W(n-1) \\ 1,2 & W(n) = W(n-1) \end{Bmatrix} \tag{8}$$

Thus as fig2, the path must start from the bottom left corner and end at the upper right corner; secondly, in order to prevent the blind search, there is no toleration of the path in favor to the horizontal or vertical axis, usually the smallest is 1/2 and the biggest slope is 2. The partial constrained path is shown in Fig3. The previous grid of (i_n, i_m) can only be (i_{n-1}, i_m), (i_{n-1}, i_{m-1}),(i_{n-1}, i_{m-2}).

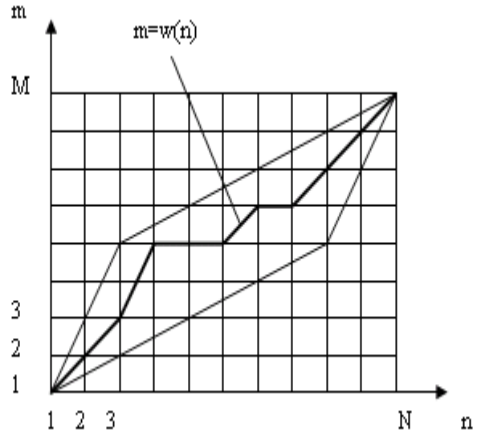

Figure2. The optimal path

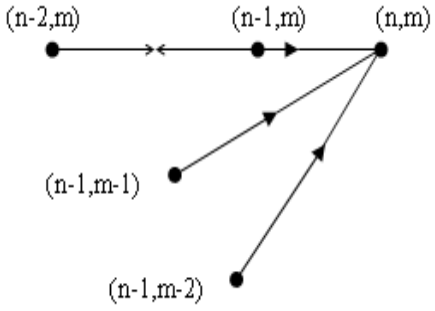

Figure3. Partial constrained path

Provide the d[n,m]=d[n,w(n)] is the distance between B_n and A_m, then the minimum sum distance is:

$$D(n,m) = \min_{w(j)} \sum_{j=1}^{n} d\left[j, w(j)\right] \tag{9}$$

And

$$D(n+1,m) = d[n+1,m] + \min[D(n,m)g(n,m), D(n,m-1), D(n,m-2)] \tag{10}$$

Where g(n, m) is the nonlinear weighting.

$$g(n,m) = \begin{cases} 1 & w(n) \neq w(n-1) \\ \infty & w(n) = w(n-1) \end{cases} \tag{11}$$

When carries on the speech recognition, the testing template will match to the reference template and receive the minimum matching distance D_{min}(N,M) which is the recognition result.

To guarantee that the optimum path to (n, m) does not stay flat for two consecutive frames. The final desired solution to (9) is:

$$D = D_A(N, M) \tag{12}$$

$$\bar{D} = D_A(N, M) / N \tag{13}$$

Thus the DTW algorithm requires on the order of NM distance calculations, and NM sets of combinatorics [(10) and (11)] to obtain the best path and the total distance for each reference pattern. We now consider alternative finding techniques which seek to reduce the number of local distance calculations.

B. The Improved DTW Algorithm

We have shown that the conventional DTW algorithm solves the problem of optimally time aligning a test and a reference pattern, at the same time providing a measure of the similarity (distance) between test and reference along the alignment path. By restructuring the entire time alignment problem as a problem in finding the best path through a finite grid of points, we can take advantage of a large class of ordered tree and graph searching algorithms, to find the best path with substantially reduced computation of local distances.

The distance between vector U_i and V_i is considered in the existing algorithms, where minimum distance is the requirement for the matching points, and the minimum sum weighted distance is the dynamic similarity measurement of the U and V sequences. While this paper uses similarity between vector U_i and V_i for calculation. The max similarity is needed to search the matching points, and the maximum sum similarity is the measurement of the U and V sequences. Similarity L (i, j) is shown as follow:

$$L(i,j) = \frac{U_i V_j}{|U_i| \bullet |V_j|} \tag{14}$$

Usually L（i，j）<= 1，while L（i，j）=1, vector U is the same as vector V.

There must exist N*M matrix to calculate maximum similarity between a reference pattern of N frames and a test pattern of M frames, which is time consuming. So we find several points $U(m_1), V(n_1), U(m_2), V(n_2), .\cdots U(m_k), V(n_k)$ in the matrix, as $L(U(m_1), V(n_1)), L(U(m_2), V(n_2)), .\cdots L(U(m_k), V(n_k))$ where is the maximum, and $(n_1 < n_2 < \cdots < n_k), (m_1 < m_2 < \cdots < m_k)$ Then we only need to consider :

$$\begin{aligned} &\{L[((U(1), V(1)), (U(m_1), V(n_1))], \\ &L[(U(m_1), V(n_1)), (U(m_2), V(n_2))]... \\ &L[(U(m_{k-1}), V(m_{k-1})), (U(m_k), V(n_k))\} \end{aligned} \tag{15}$$

So we obtain:

$$L[U,V]=\sum_{i=1}^{k-1}L[(U(m_i),V(n_i)),(U(m_{i+1}),V(n_{i+1})]] \quad (16)$$

Where K is the number of matching points.

$$L[((U(m_i),V(n_i),(U(m_{i+1}),V(m_{i+1})]] = \sum_{n=n_i,m=m_i}^{n_{i+1},m_{i+1}} L[U(m),V(n)] \quad (17)$$

Thus the N*M matrix is divided into several sub matrix and the computing time is reduced relatively.

IV. System Implemntation

A. Simulation

The simulation experiments have been done on the platform Matlab2010a. The recognizer was trained by using a Chinese isolated word database. The database consists of 10 words which were uttered by 60 different speakers, including 30 males and 30 females. Each speaker pronounced every word 5 times. 2000 utterances from this database were used as training data and the 1000 remaining utterances were used as test data. The sampling rate for the speech signal is 8KHz, each frame consists of 256 sample points. 12 MFCC and the log energy component were used together with their delta and delta-delta coefficients, Then HRSF is put in use for contrast. Table 1 shows the recognition results for different approaches on the given dataset.

TABLE I.

THE RESULTS OF IMPROVED ALGORITHM AND TRADITIONAL ALGORITHM

Algorithm	Recognition rate (%)	Recognition time (ms)
MFCC + The Conventional DTW	86.3	413.22
HRSF + The Conventional DTW	92.7	436.17
MFCC+The Improved DTW	90.5	286.37
HRSF+The improved DTW	96.1	309.46

In this research, the performance of speaker verification based on different methods was evaluated. The percentage of Identification for the MFCC is (86.3% and 90.5%) is lower than HRSF(92.7% and 96.1%). A improvement in the performance is seen by using HRSF. From table 1 we can conclude that the average recognition time decreases rapidly by about 29.5%.

B. FPGA Implementation

An isolated word recognition system was developed for evaluating the proposed approach. The flowchart of the system is shown as figure4.

The hardware of the system is implemented on Altera's DE2 board which mainly includes CycloneII 2C35 FPGA chip, serial Flash EPCS16, 4MB Flash memory, kB SRAM, 8MSDARM 512, MIC 8, bitADC 0809, LCD, serial communication interface, PS2, etc .

The audio signal is acquired by 16 bit Audio CODEC audio module on the board, and the timer interrupt.is used to control the system. The system enables audio module read audio data from the buffer of the module then sent it to DDR SDRAM . When the audio module buffer is empty, the following data process will continue as audio data pre-processing, endpoint detection, MFCC coefficients extraction and HRSF upgrading. If the process is in the training phase, the upgraded MFCC coefficients will be stored as reference templates into FLASH. While in the recognition stage, the coefficients will be transferred to the DDR memory model as the test template to match the reference template. The improved dynamic time warping (DTW) algorithm.is applied in pattern matching procedure and the recognition results are given as the output .

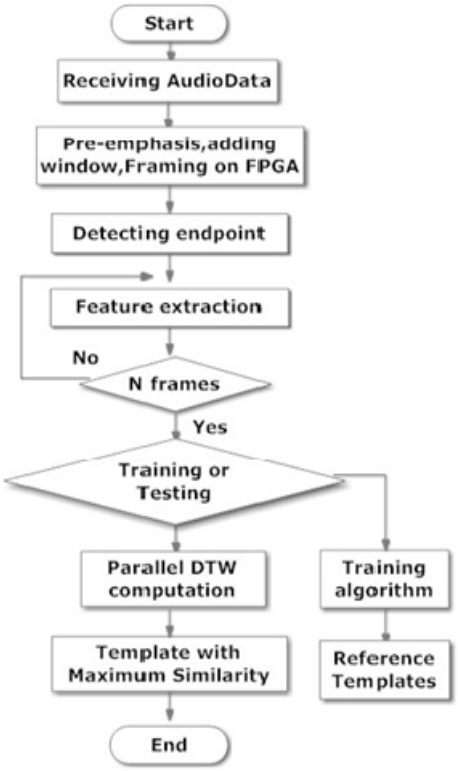

Figure 4. The flowchart of the speech recognition system

1) Framing & Pre-emphasis

For spectrum or track analysis, Pre-emphasis process wides and smoothes the spectrum of the signal by raising the high frequency part. With short-time stationary characteristics, the speech signal can be divided into frames to reduce the negative effect causing by time variant of the speech signal.

The equation is shown as (18):

$$Sign(n) = S(n) - \alpha * S(n-1) \quad (18)$$

Where α is 0.9375; s(n) is digital speech signal while sign(n) is pre-emphasized speech signal. In FPGA, it can be computed as (19), The $S_{n-1}/16$ can be obtained through the data shifting, the whole formula effectively improves the speed of data processing without decreasing the data accuracy by avoiding floating-point, multiplication and division operations.

Framing:

$$S_w(n) = \sum_{-\infty}^{\infty} s(m)w(n-m) \quad (19)$$

Where s(n) is the origin signal, $s_w(n)$ is the framed signal, w(n) is window function.

$$w(n) = \begin{cases} 0.54 - 0.64\cos[2\pi n/(N-1)] N = 0,1...N-1 \\ 0, N \text{ is others} \end{cases} \quad (20)$$

2) Endpoint detection and feature extraction

The speech signal is serial input into FPGA after sampling and A/D converting (the sampling frequency is 8 kHz and 16 bit sampling depth) . The FPGA chip has a user-set hardware module which can receive the signal in real time, compute the difference from the prior data and the square value of the data simultaneously, then store the results in SRAM. While the memory is full, the module informs CPU to read data by the interrupt mode. According to the formula (21), (22) by calculating the short-time energy of each frame and zero-crossing rate, the end point is detected and data is valid through the double threshold. In equation (21), (22), N is the sampling points per frame.

$$ZCR(i) = \sum_{n=1}^{N-1} \left| S_i(n+1) - S_i(n) \right| \quad (21)$$

$$E(i) = \sum_{n=1}^{N-1} S_i^2(n) \quad (22)$$

Because the sampling frequency is much lower than the frequency of the system, MFCC coefficients extraction can be completed in idle time during the process of speech acquisition and detection.

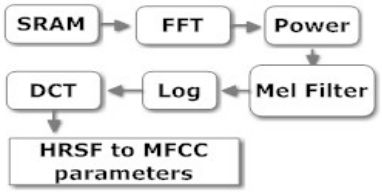

Figure5. The flow chart of the feature extraction

In the training phase, the reference templates are established after MFCC coefficients extraction, and stored in the Flash memory. Each reference template contains the flag and the character array. While in the recognition stage, the MFCC parameters extracted in real time and the reference template are sent into the DTW function module and the number of template with the maximum similarity is output as the candidate results.

3) The parallelism of DTW algorithm

The DTW algorithm is realized on the IP core hardware(Figure6). The IP core is finally added to the FPGA PLB bus system for system integration by using EDK tool after the design, compiling, simulation, synthesis, and verification. The IP core contains certain processing element (PE) (Figure7) queue to realize parallel computing[9][10], reference templates management unit, the test templates management, similarity data management unit, PE, and control unit.

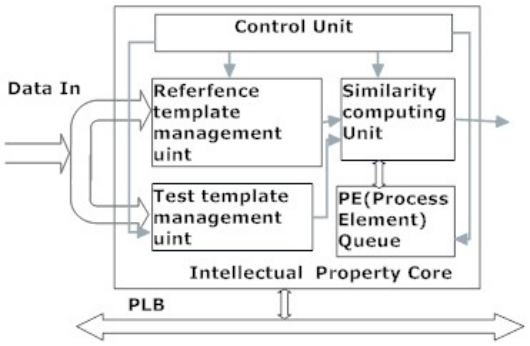

Figure 6. The internal structure of the Intellectual Property core

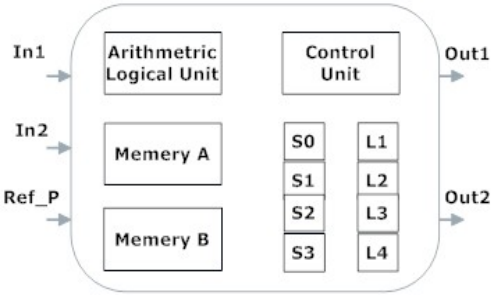

Figure 7. The internal structure of the Processing Element

The basic building blocks of each PE is shown as Figure 7 which includes:

i) A control unit(CU): Instruction decoding is performed by the CU which next generates those signals in order to control the various arithmetic elements of the PE. The instruction memory contains the program (depending on the improved DTW algorithm as mentioned in Section III).

ii) An arithmetic unit performs a number of operations required for the computation of the partial sum and the local similarities.

iii) Memory unit consisting of two memory blocks Memery A and Memery B whose sizes are equal to p where p is the dimension of the feature vectors U_i ,V_j.

iv) Registers S0,S1,S2,S3,L1,L2,L3,L4 retaining the partial sums $S_t(k-1)$,$S_t(k)$,$S_t(k+l)$,$S_{t-1}(k+1)$ and the local distances $L_t(k)$,$L_t(k-1)$,$L_t(k+1)$ and $L_{t-1}(k+l)$.

v) Two input ports In1 and In2 and two output ports Out1and Out2 for I/O transfers. In1(P_k)and In2(P_k) are the input ports of P_k connected to the output ports Out1(P_{k-1}) and Out2(P_{k+1}) respectively. Out1(P_k)and Out2(P_k) are the output ports of P_k connected to the input ports In1(P_{k+1})and In2(P_{k+1}),respectively, as shown in Figure8.

vi) Each P_k is connected via another port to the RPB bus.

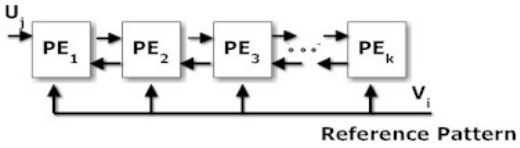

Figure8. The structure of PE queue

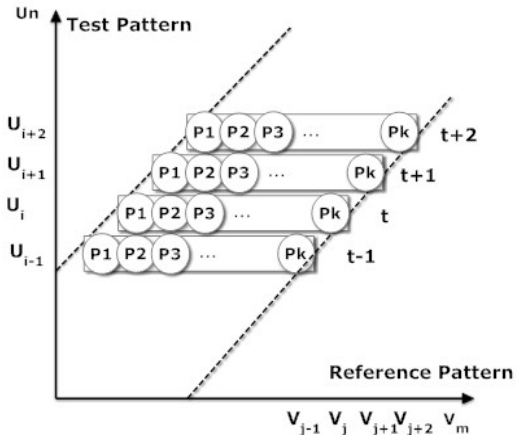

Figure9. Scheme of parallel DTW computation performed by PE's

The system can trace the best path with high-speed through the parallel processing and pipeline processing [5]. The reference template is connected to one input port of every PE_k, at the same time, the In2 port of PE_1 is connected with the test template. In order to solve the problems caused by too much PEs, which consumed too much hardware resources, K could be chosen as M/2 (a number M for the reference templates) or another appropriate value according to the space of SRAM. The scheme of parallel DTW computation is shown as Figure9.

C. Experimental Results

The experiment was divided into two parts: speaker-dependent speech recognition and speech-independent speech recognition.

The data for the former was the same as for the simulation test, which was sent into FPGA via the serial port in Wav format. The recognition results were displayed on the LCD finally. For the latter, Chinese speech sample database is applied, which has a high signal-to-noise ratio that can minimize the impact of noise on the recognition performance.

It can be seen in the experimental that the system average recognition rate is higher in the speaker-dependent recognition(reached 96.47%), while due to the different pronunciation, stress and the speed of each speaker, the average recognition rate of speaker-independent has declined slightly to 87.93% as shown in table 2.

TABLE II.

THE RESULTS OF SPEAKER-DEPENDENT AND SPEAKER-INDEPENDENT RECOGNITION ON FPGA

	Speaker-dependent	**Speaker-Independent**
Average recognition rate	96.21%	87.93%

V. CONCLUSIONS

In this paper, MFCC and HRSF methods for feature extraction, the conventional DTW and the improved DTW for recognition of isolated words is presented. The system including the parallel algorithm of DTW is implemented on FPGA. Results show that the performance of HRSF is better than MFCC since higher order MFCC components are more susceptible to the influence of noise than the lower order MFCC components. Compared the traditional DTW with improved DTW algorithm, the latter can obtain a better recognition rate and faster response time. The system obtained a better recognition rate on speaker-dependent speech recognition than speaker-independent speech recognition. The overall recognition rate can to be improved by future work on the feature extraction and endpoint detection.

ACKNOWLEDGMENT

The authors thank gratefully for the colleagues who have been concerned with the work and have provided much more powerfully technical supports. The work is supported by Guangdong Science and Technology Department under Grant No.[2010]170, No.2009B010800051, No.2009B010900056, No.2009B090300336, Guangdong Economy & Trade Committee under Grant No. GDEID2010IS034，Science and Technology Bureau of Guangzhou under Grant No. 2011J4300046. Guangdong-HK Research Fund Under Grant No. 20100101-5,Guangdong-Science Academy Cooperation Research Fund Under Grant No.2011A090100003,Zhujiang Science Star Under Grant No. 2011J2200089，China State Key Lab of Software Engineering Under Grant No. SKLSE2010-08

REFERENCES

[1] Ahmed Salman, Ejaz Muhammad and Khawar Khurshid, "Speaker verification using boosted cepstral features with gaussian distributions," IEEE International Multitopic Conference, 2007. INMIC 2007, 2007 pp.1 – 5.

[2] Anup Kumar Paul, Dipankar Das and Md. Mustafa Kamal, "Bangla speech recognition system using lpc and ann," Seventh International Conference on Advances in Pattern Recognition, 2009, pp. 171 – 174.

[3] C. Charbuillet, B. Gas, M. Chetouani and J. L. Zarader,"Complementary features for speaker verification based on genetic algorithms," IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 4 2007 pp. IV-285 - IV-288

[4] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Upper Saddle River, New Jersey: Prentice-Hall, 2001.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," IEEE Trans. Speech Audio Process., vol. 2, no. 4, pp. 578–589, Oct.1994.

[6] X. Xiao, E.-S. Chng, and H. Li, "Temporal structure normalization of speech feature for robust speech recognition," IEEE Signal Process. Lett., vol. 14, no. 7, pp. 500–503, Jul. 2007.

[7] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in Proc. Speaker Odyssey: Speaker Recognition Workshop (Odyssey 2001), Crete, Greece, Jun. 2001, pp. 213–218.

[8] C.-P. Chen and J. A. Bilmes, "MVA processing of speech features," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 1, pp. 257–270,Jan. 2007.

[9] GREGORY N. STAINHAOUER, GEORGE CARAYANNIS, New Parallel Implementations for DTW Algorithms, IEEE Transactions on Acoustics,Speech.and Signal Processing.VOL.38,NO.4.APRIL 1990,705-711

[10] Bae J Y,Fairhurst C M.Parallelism in Dynamic Time Warping of Automatic Signature Verification[C].Document Analysis and Recognition.Proceedings of the Third International Conference on.Canada,1995,1:426-429.