Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 2633 – 2637

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Embedded Speaker Recognition System Design and Implementation Based on FPGA

Jingjiao Li, Dong An*, Lili Lang, Dan Yang

School of Information Science & Engineering, Northeastern University, Shenyang 110004, China.

Abstract

For the hardcore processor such as DSP, the existence of embedded speech recognition system taking more time on train and recognition, this paper presents an FPGA-based platform with the principle of vector quantization speech recognition system implementations. In vector quantization using genetic algorithm for speaker recognition systems, the parallel hardware structure of the program can greatly reduce the calculation the time-consuming. After testing, the implementation program under the premise of ensuring the recognition rate, which can effectively reduce the time-consuming of the training and recognition.

Key words: Speaker Recognition; Endpoint detection; MFCC; Vector quantization; FPGA

1. Introduction

Speaker recognition is a biometric modality that uses an individual's voice for recognition purposes [1]. The speaker recognition process relies on features influenced by both the physical structure of an individual's vocal tract and the behavioral characteristics of the individual [2]. Speaker recognition technology as a non-contact identification technology, in the judicial, military, and information services, etc. Embedded speaker recognition system, at present, is usually based on DSP processors and other hardware platforms, training and recognition time-consuming [3], bad real-time. In this paper, on the basis of the vector quantization (VQ) for speaker recognition algorithm, the Cyclone II2C35 series FPGA to achieve embedded speech recognition system, using the characteristics of vector quantization and the genetic algorithm, considering the training and recognition time-consuming, resource consumption and

* Corresponding author. Tel.: +86-024-836-78543.
E-mail address: 249350656@qq.com.

doi:10.1016/j.proeng.2012.01.363

other factors. Finally, after verification, the speaker recognition system has a high recognition rate, better than the hardcore processor system in real-time aspects.

2. System Design

This design of the application of speaker recognition system is closed set text-dependent identification. Were measured by collecting and analyzing the voice signal, the system can identify the identity of the persons registered. Specific functions are as follows:

- Collection to be identified by three seconds the length of the voice data;
- Voice data feature extraction;
- Test vector quantization codebook;

Through the analysis of specific functions, the overall framework of the system can be used in the form of Figure 1: Collection to be identified in 3 seconds of voice data stored in the SRAM chip. Voice data were collected simultaneously in the endpoint detection. End of the completion of endpoint detection, effective voice data extraction Mel Cepstrum [4]. With codebook in this library identified the treatment of speech features to quantify, to determine the minimum distortion codebook. Finally, the threshold comparison, the recognition results with digital display.

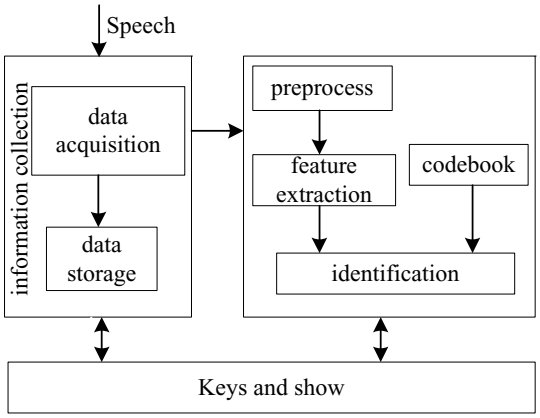

Figure. 1 System architecture

3. The hardware design and implementation

The system is fully functional on the Altera EP2C35F672C6 board, the WM8731 audio codec chip on the information collection.

3.1 Pre-emphasis unit

Under normal circumstances, the value of pre-emphasis coefficient closes to 1, between 0.9 and 1, typically 0.93. To facilitate hardware programming, a first-order FIR filter with a differential equation is expressed as:

$$S[i]=S[i]-\frac{15}{16}S[i-1]=S[i]-(S[i-1]-\frac{S[i-1]}{16}),0\le i\le N \quad (1)$$

S[i] of the original speech signal sequence, N is the length of the voice. 16 divided by 4 can be used to achieve the right, there would be simplified to shift operations division, and reduces the computational complexity. The hardware circuit is shown in Figure 2:

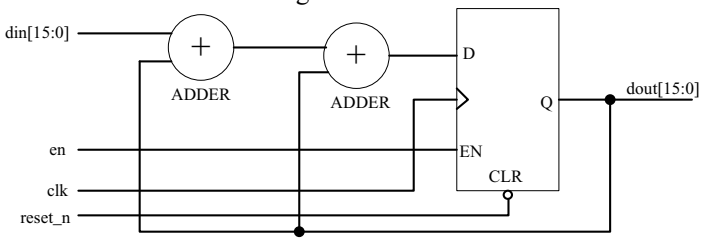

Figure. 2 RTL-level view of pre-emphasis

3.2 Sub-frame unit

The sub-frame of voice data uses counter. A data length 128 points, 64 points frame shift. The short-term energy and zero short-term average rates are calculated by a summation process. Half between frames of data overlap, so the actual operation only half of the points can be calculated. Specific optimization as follows: Using a binary counter 64 count of voice data, the accumulation module sum of short-term energy and short-term average rate of zero respectively. Clearing the counter after the counter overflowed, and the results of the accumulator stored in the depth of 2 shifts register. Shift register shifts once every 64 clocks, and calculate the sum of two values in the register; the results are data value of short-term energy and zero-rate short-term average value.

3.3 Detection Unit

Figure 3 for the endpoint detection module of the RTL-level view. Vad_part1_sp2 for data conversion unit; vad_d1 for valid bit delay unit; vad_part2_sp1 for the short-term energy and the short-term average zero crossing rate calculation units; vad_part3_sp1 for the endpoint detection unit.

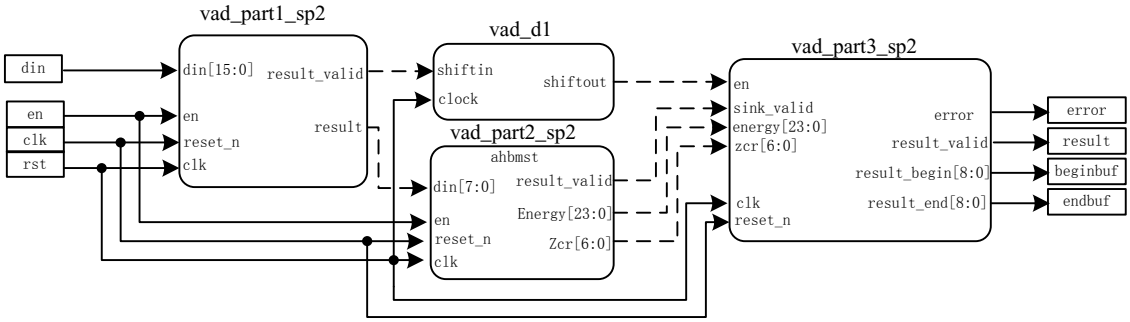

Figure 3 Endpoint detection RTL view

In the endpoint detection unit, en is the input enable flag; input energy and zcr are the average short-time energy and short-term zero-crossing rate; enGate threshold for the short-term energy register; zcrGate for the short-time average zero crossing rate threshold register; BeginLen frame length threshold for the starting point; EndLen frame length threshold for the end point; flag1 as a starting point detection complete flag, flag 2 is an end point detection complete flag.

3.4 VQ recognition module

Identification is to find the minimum distortion codebook in the process, the hardware structure shown in Figure 4. The unknown vector sequence and each codebook for the Euclidean distance test; the resulting cumulative error value is the value of the distortion codebook. Minimum distortion of the codebook is the sentence object. Among them, the counting statistics module for the unknown vector sequence, the address generator control code output. When, after all the codebook for the input vector quantization, the decision module detects the minimum distortion codebook, the output results.

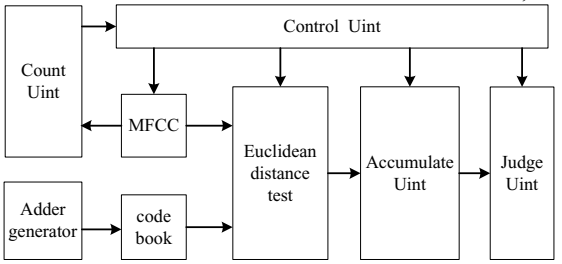

Figure 4 VQ identify the hardware structure

This system is for specific populations to identify, code the library is relatively fixed, without the need for real-time updates. Therefore, the code of this training can be done on a PC. Codebook storage using the single-precision floating point number, a code book size of 128 * 12 * 32bits, the total of 6KB.

Decision unit is for detecting the minimum distortion code. Taking into account the environment and the pronunciation of the speaker recognition accuracy, in the process to determine a threshold value set: When the minimum distortion codebook error is greater than the threshold, indicating serious environmental disturbance or voice disorder, refused to recognize; Otherwise, the test valid, the output recognition results.

4. Performance Test

In the FPGA implementation of speaker recognition system selected voice data length of 128, Mel Cepstrum of 12 bands. System clock is 50MHz, the sampling rate of voice signal 8KHz, data bits to 12bits, acquisition time of 3 seconds. Due to hardware resource constraints, the system can't store too many codebooks the same time. In the test, using the cross-packet manner.

4.1 Testing recognition rate

The 301 tests were grouped by gender. Texting of each group, respectively, were text relevant and text-independent text. In the laboratory environment, the test text is the Northeastern University. The results are shown in Table 1.

Table 1 Text-related test

	Testing times	Correct recognition times	Error recognition times	Refusal times	Correct recognition rate
male	231	210	11	10	90.9%
female	70	58	6	6	82.8%
total	301	268	17	16	89%

The test contents of text-independent are random, the test results shown in Table 2.

Table 2 Text-independent test

	Testing times	Correct recognition times	Error recognition times	Refusal times	Correct recognition rate
male	231	120	71	30	51.9%
female	70	39	11	20	55.7%
total	301	159	82	50	52.8%

The results: Text-related recognition rate of 89%, text-independent recognition rate of 52.8%, hardware and software simulation results are basically the same.

4.2 Real-time testing

The real-time tested on speaker recognition system. The test environment is: Intel (R) Core (TM) 2 Quad Q8400 2.66G quad-core processors, DDR2 800 (PC6400) 2048Mbyte memory, the software Matlab R2010b. Hardware environment: Altera provided EP2C35F672C6 development board, the system clock is 50MHz. Test results shown in Table 3.

Table 3 codebook system run-time

Matlab (ms)	FPGA (ms)
264.583	14.976

The results: In the system clock of 50MHz conditions, FPGA processing speed is 17.6 times the PC has better real-time performance.

5. Conclusion

For the real-time problem, this paper presents a solution that makes the FPGA as the hardware platform. This program uses the speaker recognition algorithm based on MFCC and VQ. The system consists of five parts: Signal Acquisition, Endpoint Detection, Feature Extraction, Training and Identification. The experiment results show that the time-consuming is 15.932ms on the 4 codebooks and 50MHz-clock system, the identification rate is 93.3% on the 12 codebooks system. This kind of design improves the system's recognition speed, which is an effective program to solve the real-time problem.

References

[1] Choi, Y.-K., K. You. A Real-Time FPGA-Based 20 000-Word Speech Recognizer With Optimized DRAM Access. Ieee Transactions on Circuits and Systems I-Regular Papers 2010; **57**(8): 2119-2131.

[2] Manikandan, J., B. Venkataramani. FPGA Implementation of Support Vector Machine based Isolated Digit Recognition System. 22nd International Conference on Vlsi Design Held Jointly with 8th International Conference on Embedded Systems, 2009; Proceedings: 347-352.

[3] Sarkar, G. and G. Saha. Real Time Implementation of Speaker Identification System with Frame Picking Algorithm. Proceedings of the International Conference and Exhibition on Biometrics Technology. G. A. D. C. L. K. A. Atkinson. 2010; 2: 173-180.

[4] Tamulevicius, G., V. Arminas, et al. (2010). Hardware Accelerated FPGA Implementation of Lithuanian Isolated Word Recognition System. Elektronika Ir Elektrotechnika.2010; **23**(3): 57-62.