ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

FPGA Acceleration of Chinese speech recognition algorithm

Xuanang Chen[1*], Qijia Tang[2, a], Boyu Wang[3, b], Xinyu Lin[4, c], Deyi, Wan[5, d], Liyan Bao[6, e]

[1]Beijing Jiaotong University, Beijing, China

[2]Southwest Jiaotong University, Chengdu, Sichuan, China

[3]Wuhan University of Technology, Wuhan, Hubei, China

[4]University of Electronic Science and Technology of China, Chengdu, Sichuan, China

[5]Shandong Province Qingdao Second Middle School branch, Qingdao, Shandong, China

[6]Adcote School Shanghai, Hangzhou, Zhejiang, China

[a]1322540573@qq. com, [b]32183128589@qq. com, [c]3028609987@qq. com, [d]1006056275@qq. com, [e]3263592398@qq. com

*Corresponding author: 1104030651@qq. com

These authors contributed equally to this work

Abstract

In recent years, the AI algorithm of speech recognition based on English isolated characters has developed quite mature, but the algorithm acceleration based on FPGA and chip design are rarely discussed. This article describes that our research team, based on the in-depth investigation of the existing speech recognition algorithms, puts forward the improvement of the original algorithm, so that it can be applied to Chinese isolated character recognition with high accuracy, then writes the code into Verilog language that can be compiled, and finally integrates the software and selected hardware peripherals through FPGA, A Chinese speech recognition chip which can be applied to the smart home for the elderly is designed.

Keywords

Smart home, Chinese speech recognition, algorithm, FPGA.

1. Introduction

FPGA has become a new trend of IC industry. Field Programmable Gate Arrays are semiconductor devices that are based around a matrix of configurable logic blocks connected via programmable interconnects. One of the most salient features is its flexibility. FPGA functionality can change upon every power-up of the device. So, when a design engineer wants to make a change, they can simply download a new configuration file into the device and try out the change. This characteristic is exactly what is needed in the emerging field of speech recognition. Nowadays,the requirements of speech recognition technology are gradually increasing. Especially important in the filed of disability assistance. For example,the deaf people can communicate with others through speech-to-text app which is based on the speech recognition. We believe as speech recognition technology evolves,the discrimination against disability will be reduced. To improve the accuracy of speech recognition,AI and machine learning will be advanced solutions. Both of them need a large number of samples. And during the researches,the engineer needs to frequently optimize the program structures. To test a new alternation,use conventional ASICS will cost a fortune and waste a lot of time. So FPGA is obviously

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

a better choice. We reckon that Speech recognition and FPGA will achieve each other and develop together.

2. Literature review

J. Whittington, K. Deo, T. Kleinschmidt and M. Mason(2008) utilized FPGA to realize an speech recognization system especially in in-car environment. They applied model compensation and modified the recognization algorithms in order to obtain accurate recognition result at the high levels of noise, with only about ten percent usage of the overall available FPGA resources. Xiaoxiao Bian(2012)used FPGA to build a smart home system,An advanced speech recognization system that can help the host easy to command the whole system. Due to the number of sensors ,it's easy for FPGA to optimize the structure of the system. Xiaoyu Ren(2015) did research on speech blind guidance system based on FPGA, the functional modules of his system are implemented on Quartus II, and the blind lane recognition, zebra crossing recognition, traffic light recognition and other modules of the system are simulated respectively. Another essential achievement made by JingXiang Zeng(2017) is a desktop help meal robot, which is a hardware system based on FPGA and Arm. Wen-Chung Tsai(2018) conducted a research of a toilet voice control system based on FPGA. During the test, the flusher, nozzle, and heater of the toilet could be successfully manipulated as user gave voice commands. Yuanjiang Wu and Sheng Li(2020) used speech recognization to help people to classify the garbage,the hardware device was implemented based on FPGA. Training the system to capture the item of garbage,citizen can tell the robot the name of the garbage,the robot will guide the right case,based on the speech recognization and the FPGA.

3. Research process

This part is going to introduce the principle of speech recognition algorithm, FPGA development process, FPGA core algorithm module architecture and FPGA peripherals in detail.

3.1 Matlab Algorithm

Dynamic Time Warping(DTW) Algorithm : In speech recognition and speaker Recognition, the most commonly used phonetic feature is the Mel-scale Frequency Cepstral Coefficients(MFCC). According to the research of the human ear hearing mechanism, the human ear has different hearing sensitivity to sound waves of different frequencies. The speech signal from 200Hz to 5000Hz has a great influence on the clarity of speech. Since the distance of the upward wave of the lower frequency sound in the inner cochlear basilar membrane is greater than that of the higher frequency sound, in general, the bass is easy to mask the treble, and the treble is more difficult to mask the bass. The critical bandwidth of sound masking at low frequencies is smaller than higher frequencies. Therefore, people arrange a set of bandpass filters according to the critical bandwidth from dense to sparse in the frequency band from low to high frequencies to filter the input signal. The signal energy output by each band-pass filter is taken as the basic feature of the signal, and this feature can be used as the input feature of the voice after further processing. This feature does not depend on the nature of the signal, does not make any assumptions and restrictions on the input signal, and uses the research results of the auditory model, and this parameter is more in line with the auditory characteristics of the human ear. It still has good recognition performance when the signal-to-noise ratio is reduced.

3.2 MFCC extraction

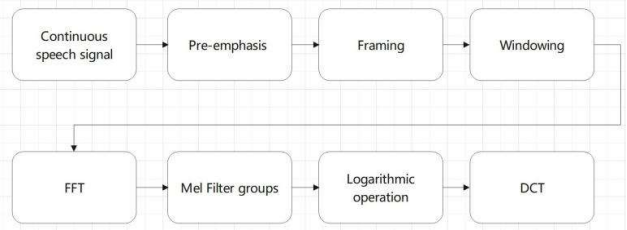

Fig. 1. MFCC extraction flowchart

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

Fig. 1 represents the fundamental algorithm modules to acquire MFCC parameter, each step is introduced in detail as follows:

Pretreatment: it comprises four main signal processing steps, which are, endpoint detection, pre-emphasis , framing and widowing.

Fast Fourier Transform(FFT): the transformation of the signal in the time domain is usually difficult to see the characteristics of the signal, so it is usually converted to the energy distribution in the frequency domain for observation. Different energy distributions can represent the characteristics of different voices. Fast Fourier transform is performed on each frame signal after frame division and windowing to obtain the frequency spectrum of each frame. And take the modulus square of the frequency spectrum of the speech signal to obtain the power spectrum of the speech signal.

Mel Transform: in the selection of the MEL filter, the filter bank usually selects a triangular filter, which smoothes the frequency spectrum and eliminates harmonics, highlights the formants of the original voice, and can also reduce the amount of calculation. But other shapes can also be selected, such as a sinusoidal filter. Pass the energy spectrum through a set of Mel-scale triangular filter banks to define a filter bank with M filters (the number of filters is similar to the number of critical bands), the filter used is a triangular filter, the center The frequency is f(m), m=1, 2. M. M usually takes 22-26. The interval between each f(m) decreases as the value of m decreases, and widens as the value of m increases.

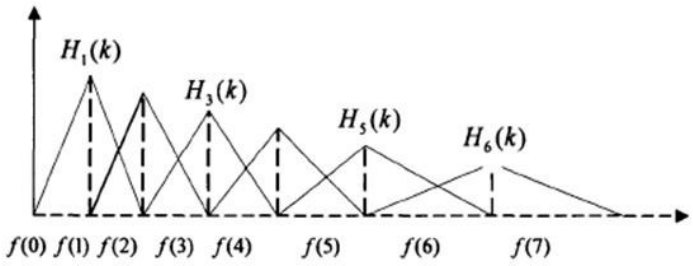

Fig. 2. Mel frequency scale filter bank

Discrete Cosine Transform: bring the logarithmic energy obtained in the previous step into the discrete cosine transform to obtain the L-order Mel parameter.

Extraction of Dynamic Difference Parameters: The standard cepstrum parameter MFCC only reflects the static characteristics of the speech parameters, and the dynamic characteristics of the speech can be described by the difference spectrum of these static characteristics. Experiments show that combining dynamic and static features can effectively improve the recognition performance of the system.

3.3 DTW Algorithm Pattern Matching

Forward seeking the best match cumulative distance:

In order to perform the alignment operation, we need to construct a Y * X matrix first. The matrix element (i, j) represents the distance d(Qi, Cj) between the two points Qi and Cj. This distance calculation uses the Euclidean distance. That is, $D(Qi,Cj) = (Qi-Cj)^2$. Finally, the warping distance obtained is solved using dynamic programming.

Backtracking to find the best path: Starting from the point (ix,jx)=(XY), according to the path limitation, find the grid point with the smallest cumulative distance among the three possible previous grid points in the previous column as the previous grid point, and so on, and then find the grid Continuing the previous grid point of the point until the grid point (i1,j1)=(1,1) is found. Then these grid points are the corresponding grid points for the best matching. These grid points are the most matching grid point, connect them can get the Best path.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

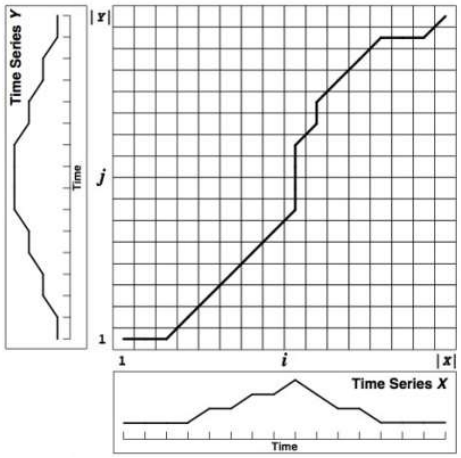

Fig. 3. Forward seeking

DTW Template Training:In recognition based on template matching, a feature template library must be established first, that is, each word in the command library is read aloud multiple times, and the generated speech is preprocessed, endpoint detection, feature extraction to obtain feature parameters, and training to generate feature templates. Multi-template average training method is used here. The specific word is read aloud several times, and each feature vector sequence is extracted respectively, and then these feature vector sequences are averaged on the DTW path to obtain the final feature vector template.

3.4 MATLAB experiments and results

In the following experiments, we use several groups of corpora that may be used in smart home, in which the semantics is indicated in parentheses after Chinese corpora.

Table. 1. test content

Test voice file	Corresponding test voice content	Identification results	Corresponding reference file
1. wav	stop	stop	r_2. wav
2. wav	hello	hello	r_1. wav
3. wav	seven	seven	r_4. wav
4. wav	monkey	monkey	r_3. wav
5. wav	left	left	r_5. wav
6. wav	down	down	r_7. wav
7. wav	up	up	r_6. wav
8. wav	turn on the living room light	turn on the living room light	r_11. wav
9. wav	turn off the living room light	turn off the living room light	r_12. wav
10. wav	stop	stop	r_10. wav
11. wav	start	turn on the living room light	r_8. wav
12. wav	turn on the kitchen light	seven	r_4. wav
13. wav	turn off the kitchen light	turn off the kitchen light	r_13. wav
14. wav	turn right	turn right	r_14. wav
15. wav	turn left	turn left	r_14. wav
16. wav	yes	yes	r_16. wav
17. wav	no	no	r_17. wav

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

When the amount of data is small, the recognition rate of the algorithm can reach 81. 25% (the clarity of pronunciation will also affect the recognition rate).

There are 4 reference words for speech recognition of household appliance control: r_1. wav (turn on the living room light), r_2. wav (turn off the living room light), r_3. wav (kitchen light on), r_4. wav (turn off the kitchen light). Next, select the four speech as the test speech respectively, and repeat the recognition results of the test ten times.

First, turn on the living room light and repeat it ten times to generate voice from 1. wav to 10. wav. See the recognition results.

Calculating matching results...
The recognition result of test template 1 is: 1
The identification result of test template 2 is: 1
The identification result of test template 3 is: 3
The identification result of test template 4 is: 1
The identification result of test template 5 is: 1
The identification result of test template 6 is: 1
The identification result of test template 7 is: 1
The identification result of test template 8 is: 1
The identification result of test template 9 is: 1
The identification result of test template 10 is: 3

Fig. 4. result for opening living room light test

It can be seen from the above that for the test voice, turning on the living room light is correctly recognized 7 times, and the wrong recognition is turning on the kitchen light 3 times, with a recognition accuracy of 70%.

Repeat turning off the living room light for ten times, and the statistical identification results are as follows:

Calculating matching results...
The recognition result of test template 1 is: 2
The identification result of test template 2 is: 2
The identification result of test template 3 is: 2
The identification result of test template 4 is: 4
The identification result of test template 5 is: 2
The identification result of test template 6 is: 2
The identification result of test template 7 is: 2
The identification result of test template 8 is: 2
The identification result of test template 9 is: 2
The identification result of test template 10 is: 2

Fig. 5. result for closing living room light test

It can be seen from the above that for the test voice, turning off the living room light is correctly recognized 9 times, and the error recognition is turning off the kitchen light once, with a recognition accuracy of 90%.

Turn on the kitchen light for ten times, and the statistical identification results are as follows:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

Calculating matching results...
The recognition result of test template 1 is: 3
The identification result of test template 2 is: 3
The identification result of test template 3 is: 3
The identification result of test template 4 is: 3
The identification result of test template 5 is: 1
The identification result of test template 6 is: 4
The identification result of test template 7 is: 3
The identification result of test template 8 is: 3
The identification result of test template 9 is: 3
The identification result of test template 10 is: 3

Fig. 6. result for opening kitchen room light test

It can be seen from the above that for the test voice, turning on the kitchen light is correctly recognized 8 times, the error recognition is turning on the living room light once, and the error recognition is turning off the kitchen light once, with a recognition accuracy of 80%.

Calculating matching results...
The recognition result of test template 1 is: 2
The identification result of test template 2 is: 4
The identification result of test template 3 is: 4
The identification result of test template 4 is: 4
The identification result of test template 5 is: 4
The identification result of test template 6 is: 4
The identification result of test template 7 is: 4
The identification result of test template 8 is: 4
The identification result of test template 9 is: 4
The identification result of test template 10 is: 2

Fig. 7. result for opening kitchen room light test

3.5 Fpga: Baisc Design Process

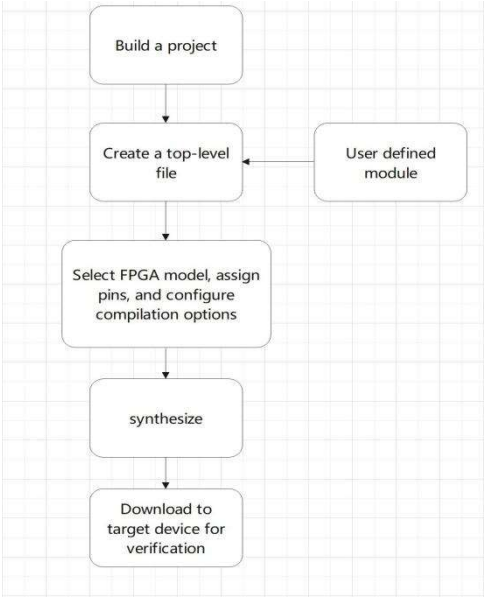

Fig. 8. shows a typical FPGA design process.

First, establish the top-level file. If it is established in the form of graphics, it is equivalent to a schematic diagram, and all modules are integrated together with bus connection. If it is created with a text file, be sure to name the module name with the project name. For users' own system, users

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

generally need to establish their own functional modules. You can use Verilog HDL or schematic input. This process is like you want to make a circuit. Now there is no function you want in the market, so you can make such a chip and add it to the circuit. In the top-level file, after instantiating and calling each module, it is made into the user's circuit system. The next step is to select the chip model and assign pins to the top-level diagram. So far, the top-level files of the system can be compiled. It is best to conduct syntax analysis first, and then conduct overall synthesis. Optimization logic combination, synthesis logic and routing constraints can be set during synthesis. After synthesis, get *. Sof file or *. POF file, which is the target download file. The former is downloaded to ram in FPGA through JTAG. If the power is off, the information will be lost; The latter is downloaded to the E2PROM in the configuration chip. After restarting, the system automatically guides the loading of the configuration file into the FPGA to achieve the purpose of power failure without losing information. But the way of online debugging programming is very fast, and the latter will have to wait for a long time. In addition, the self-contained simulation tool software has been relatively perfect, and more professional third-party software verification is generally used by the company for finished products.

3.6 Peripheral Devices Overview

The chips used in this system are based on new starting point FPGA development boards of ALIENTEK company, China;use Puzhong WM8978 MP3 module as audio decoding chip.

The following figure is the physical drawing of the development board:

Fig. 9. FPGA

The FPGA chip adopts EP4CE10F17C8 of Cyclone IV E series, which has 10320 logic units, 414kbits embedded storage resources and 23 18 × 18 embedded multiplier, 2 general phase locked loops, 10 global clock networks, 8 user IO banks and up to 179 user I / O, and the crystal oscillator is 50MHz. The system uses the development board to integrate SDRAM, SD card socket, LED lights and keys. The development board has rich resources and appropriate price, which can perfectly meet the development of this system.

The on-board SDRAM of the development board is W9825G6DH-6, with 4 l-banks, 13 bit row address, 9 bit column address and 16 bit data bus bit width. The total storage space is 256Mbit, or 32MB.

SD card uses SDHC card and FAT32 format, which has good compatibility between computer and development board.

Since the development board does not have an on-board ADC, it adopts the external universal WM8978 MP3 module. The resources of this module used by the system include: WM8978 and the microphone adopt electret condenser microphone.

Circuit diagram of Puzhong WM8978 MP3 module is shown below:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

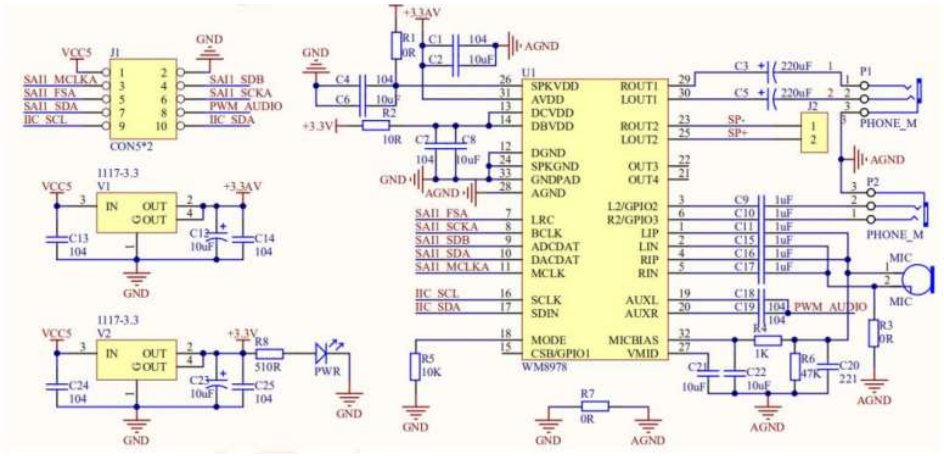

Fig. 10. WM8978 MP3 module

This is an integrated module. The system only uses the audio acquisition part. Therefore, the following pins are connected to the FPGA development board pins:

Table. 2. PINs distributions

MODULE PIN NAME	NAME	DESCRIPTION	FPGA PIN TYPE
VCC5	-	5V Power Supply	-
GND	-	Ground	-
SAI1_MCLKA	MCLK	Master Clock	output
SAI1_SDB	ADCDAT	ADC Output	output
SAI1_FSA	LRC	ADC and DAC Sample Rate Clock	input
SAI1_SCKA	BCLK	Digital Audio Port Clock	input
IIC_SCL	SCLK	IIC Clock	output
IIC_SDA	SDIN	IIC Data	inout

The communication between WM8978 and FPGA adopts IIC protocol to set the working mode of WM8978, and the audio interface communication adopts inter IC sound bus.

MCLK is 12Mhz, which is provided by FPGA. Wm8978 works in main mode and provides LRC and BCLK.

ADC is set to 48kHZ sampling rate, dual channel and 32bit sampling depth.

Since the left and right channels input by the microphone are the same, the data of the left channel is taken, and the low 16 bit data sampled each time is stored in a storage unit of SDRAM for subsequent processing. Each recording time is 3 seconds, and a total of 144000 SDRAM storage units are required.

A total of 1 key is used. When the key is pressed, the recording starts; Four LEDs are used. After identification, the corresponding LED lights up to represent the identification result.

3.7 Fpga System Frame Design And Analysis

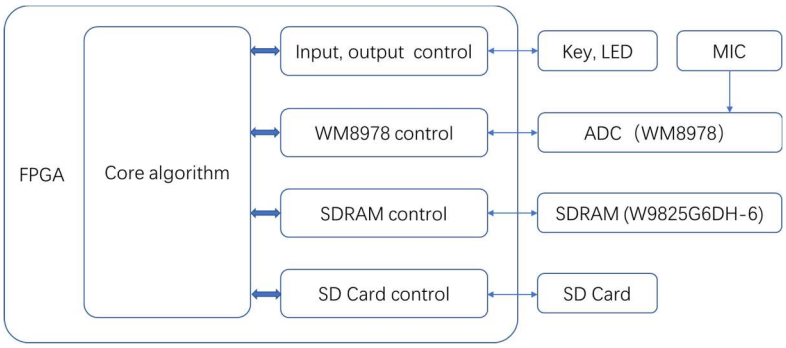

Fig. 11. FPGA system framework

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

The system design makes full use of cyclone IV E chip and new starting point development board hardware resources. An isolated word speech recognition system is realized. Firstly, the samples are input for training through the computer MATLAB software, and the first-order MFCC characteristic parameters are extracted through the software algorithm program designed on the MATLAB platform. The generated reference template is taken as the average of multiple input training samples. Through the above algorithm, 100 samples are used for training in the experiment, The reference templates of 10 isolated words 0 to 9 are obtained and stored in 1g SD card. The whole process of identification is realized on the development board. First, the tester reads the number through the microphone according to the system prompt, and the system recognizes all steps. It includes endpoint detection of multi segment speech, effective speech access, FIR filtering of speech data, compressing speech data into first-order MFCC parameters, testing people's real-time extraction of feature parameters and pre stored reference template, and obtaining recognition results through the recognition module. If the LED light is on successfully, it means that recognition is successful. The main hardware modules of the system are: audio acquisition module, each storage unit interface driver module, SD card read-write modules, LED lights and keys, etc. The main software modules of the system include: high-speed voice signal preprocessing module based on Verilog HDL, high-speed voice signal feature parameter extraction module based on Verilog HDL, voice signal matching and recognition module, LED display control module, Nios II SD card file system control module, etc. The main block diagram of the system is shown in Fig. 6.

The functional modules are described below:

Voice signal acquisition module:

The microphone line is input, and the voice is converted into 32 bit PCM code buffer at 48kHz A / D sampling rate through audio codec chip WM8978.

Voice signal preprocessing module:

As part of the algorithm processing, the main function is to normalize the data in the buffer, first-order high pass filtering, divide frames, calculate the short-time energy and short-time zero crossing according to frames, remove the silence and noise at the beginning and end, obtain effective speech frames and store them in SRAM. Prepare for speech feature extraction.

Speech signal feature extraction module:

The voice data in SRAM is read by frame, and the power spectrum of one frame of data is obtained through FFT module. The power spectrum is weighted by 24 order Mel filter bank and taken as logarithm. Then, through DCT discrete cosine transform module, the 12-dimensional cepstrum domain feature vector is obtained, and then transformed into 24-dimensional feature vector through first-order difference as the test template. In this way, a frame of 256 dimensional time domain data is compressed to 24 dimensional cepstrum domain data, which provides efficient processing for template matching and reduces the difficulty and calculation of recognition.

Speech recognition module:

Several voice samples are recorded in advance, and the characteristic parameters of these voice samples are extracted and trained by using MATLAB platform on PC, and finally stored in SD card in the form of file. Nios II controls the SD card driver to obtain all reference templates. The user-defined DTW dynamic time warping module compares the European distance between the test template and the reference template in the SD card to obtain the recognition result.

LED display module:

Interactive module of user information input and feedback.

4. Conclusion

In the research, a Chinese isolated character recognition system based on FPGA was developed. The key to achieving voice recognition function is the extraction of Mel-scale Frequency Cepstral Coefficients(MFCC) parameter and Dynamic Time Warping(DTW) Algorithm. In the test of the

ISSN: 2414-1895 DOI: 10.6919/ICJE.202112_7(12).0057

voice command, generally the algorithm was capable of Chinese isolated character recognition. The voice recognization was then realized on a FPGA platform with 5 modules which complete the operations from voice requisition to result display. In short, the speech recognition system was proved to be suitable for the application on domestic appliances, for instance, the lighting system. It could be beneficial for the large-scale applications of smart-home devices in the future.

References

[1] Wu Yuanjiang, Li Sheng. Application of speech recognition in waste classification device [J]. Electromechanical engineering technology, 2020,49 (12): 82-85.

[2] Zeng Jinxiang, Gai Kerong, Yang Jinzhong. Design of desktop meal aid robot based on FPGA and arm [J]. Journal of Beijing Institute of technology, 2017,16 (01): 22-25.

[3] Wang Mingjuan. Design and implementation of speech recognition system based on FPGA [D]. Guangxi Normal University, 2009.

[4] Ren Xiaoyu. Research on speech blind guidance system based on FPGA technology [D]. Harbin University of technology, 2015.

[5] Zhang Yumu. Discussion on the development prospect of FPGA in automotive electronics [J]. Enterprise Herald, 2012 (20): 296.

[6] Bian Xiaoxiao, Chen Yuchao. Research and design of intelligent home integrated monitoring system [J]. Computer programming skills and maintenance, 2020 (11): 117-119.

[7] W. Tsai, Y. Lian, S. Hsu, Q. Zheng, Y. Su and J. Chen. An Implementation of Voice Recognition and Control System for Electric Equipment [J] 2018 International Symposium on Computer, Consumer and Control (IS3C), 2018, pp. 356-359.

[8] J. Whittington, K. Deo, T. Kleinschmidt and M. Mason, FPGA implementation of spectral subtraction for in-car speech enhancement and recognition, [C]. 2008 2nd International Conference on Signal Processing and Communication Systems.