Voice Control System Based on Zynq FPGA

Weili Liu
Integrated Circuit Design and Integration System, Microelectronics Academy, Dalian University of Technology, Dalian, Liaoning Province, China
Email: 1786828936wl@sina.com

Abstract. In this paper, a speech recognition system is designed and the relevant hard-ware environment is built through Zynq FPGA AX7020 platform. In this system, the feature extraction of speech signals is based on MFCC (mel-frequency cepstrum coefficients). The algorithm in this system includes Pre-emphasis, Framing and Windowing, FFT, MFCC Parameter Calculation, VED Endpoint Detection and DTW Operations processes commonly used in speech system design. In this paper, the speech recognition algorithm is implemented on Zynq FPGA platform and also simulated in Matlab. At the same time, audio module AN831 is adopted in the hardware part to realize the sound acquisition function. Through the comparison of experimental data, it is found that the average recognition rate in Matlab simulation is 86.67%, while in Zynq FPGA platform test, the recognition rate is 80.67%. Using Zynq FPGA AX7020 as a platform to realize speech recognition, the recognition rate is higher than the traditional speech recognition technology.

Keywords. Zynq FPGA; speech recognition; feature extraction; MFCC.

1. Introduction
In recent years, due to the development of artificial intelligence, the demand for speech recognition technology has increased dramatically. Speech recognition system is widely used. Voice query services on the Internet, dictation input on smartphones, and these things are large vocabulary continuous speech recognition applications. In the elderly mobile phone voice dialing, smart home, car navigation voice control system, is the miniaturized, portable voice products application. The best realization of speech recognition system is to allow human and machine to communicate by voice naturally and in real time [1].

In the traditional speech recognition technology, special speech recognition software needs to be added, which will make the circuit complex, low recognition rate and long delay [2]. With the continuous development of FPGA technology, more and more applications in high-tech fields, speech recognition technology has made great breakthroughs. This is because FPGA combines the advantages of DSP and ASIC, has the characteristics of large scale, high integration, fast speed, low power consumption, strong configurability, and easy to realize flow and parallel structure [3]. Meanwhile, the Method of MFCC is also widely used in feature extraction in ASR. It can better meet the requirements of improving accuracy and recognition rate of speech recognition.

2. Analysis of Speech Recognition System

2.1. Extraction of MFCC Parameters
MFCC is the abbreviation of Mel Frequency Cepstrum Coefficient. Mel frequency is a concept proposed according to the auditory characteristics of the human ear, and it corresponds to frequency in a
nonlinear way. Mel frequency cepstral coefficient (MFCC) is the calculated Hz spectrum characteristics based on the relationship between them. MFCC has been widely used in speech recognition. The physical meaning of MFCC is a set of eigenvectors that encode the physical information (spectrum envelope and details) of speech. The extraction of MFCC parameters requires a series of speech signal preprocessing to make speech matching and recognition more convenient [4]. The specific MFCC parameter extraction flow chart is as follows in figure 1.

![Figure 1. MFCC parameter extraction flow chart.](image)

2.1.1. Pre-emphasis of Speech Signals. Due to the influence of lip and vocal cords, the high-frequency component resolution of speech signals is very low, which means that the human pronunciation system inhibits the high frequency. In order to increase the high frequency component of speech and flatten the signal spectrum, the speech signal is pre-emphasis [4]. The pre-emphasis transfer function is as follows:

\[ H(z) = 1 - \mu z^{-1} \]  

(1)

2.1.2. Framing and Windowing of Speech Signals. In the process of vocalization, the speech is continuous in a short time, so the speech signal can be divided into several short signals for processing. Framing can make the speech signal not fluctuate greatly, and it can be divided into one frame according to a certain sampling point, so that it can become relatively stable during analysis at a slightly varying angle. In the experiment, Voicebox’s built-in Enframe function is used to frame.

When framing, a finite length window is used for weighting calculation. In the Hanning Window, Rectangular Window and Hamming Window, we use the hamming window which is most suitable for increasing the resolution ability by comparison [5, 6]. The window function moves along the direction of time, turning the speech signal into a periodic signal, making it more continuous and convenient for Fourier transform, so as to analyze the speech of any frame. The Hamming window function is as follows:

\[ W = (n - \alpha) - \alpha \cos(2\pi n/(N - 1)), 0 \leq n \leq N - 1 \]  

(2)

The value of \( \alpha \) is generally 0.46.

Framing is to multiply \( s(n) \) by a certain window function \( \omega(n) \), so as to form windowed speech signal:

\[ S'(n) = S(n) \times W(n) \]  

(3)

2.1.3. FFT Processing of Speech Signals. FFT is one of the most commonly used methods in speech signal analysis. In the time domain, the signal has the discrete aperiodic characteristic, only in the
3

frequency domain the signal is converted into the continuous periodic signal. FFT can transform the
signal from time domain to frequency domain for energy analysis and get the power spectrum of the
signal.

2.1.4. Mel Filter Bank. When the speech recognition system performs nonlinear processing, the
nonlinear frequency resolution can be obtained by applying the Mel filter bank [7]. The experiment
uses the triangle filter bank, a Mel filter bank whose number of filters are generally 22 to 26, and this
system uses 24 filters. The interval between each F (m) shrinks with the decrease of the m value and
widens with the increase of the m value, which can approximately ignore the influence of the change
of human tone.

2.1.5. Logarithmic Operation and MFCC Parameter Calculation. The logarithmic energy output from
the filter bank is calculated. After logarithmic operation and DCT transformation, the first-order
dynamic difference is added to form the MFCC coefficient. The first and last frames are also removed,
because the first difference parameter of these two frames is 0. The purpose of the above approach is
to combine the individual frames. Logarithmic operation formula:

\[ s(m) = \ln \left( \sum_{k=0}^{N-1} |X_a(k)|^2 H_m(k) \right), 0 \leq m \leq M \]  

(4)

2.2. VED Endpoint Detection and DTW Operations

2.2.1. VED Endpoint Detection. VED is all called Voice Endpoint Detection. In order to compress
the frame length of speech signal and make it more compact, it is necessary to extract the non-silent part
of speech signal [8]. The whole endpoint detection can be divided into four segments: the mute
segment, the transition segment, the speech segment and the end segment. The time period is
determined by whether the two parameters x1 and x2 extracted from the speech signal exceed the
threshold. That is, the waveform in the voice signal frame can judge the voiceless and voiced sound by
the number of zeros.

2.2.2. DTW Operations. DTW (Dynamic Time Warping) is the simplest algorithm in the speech
recognition of isolated words [9]. Based on the idea of dynamic programming, this algorithm is used
to solve the problem of people’s different pronunciation length to match the corresponding template.
Through three constraints (monotonicity, continuity, and certainty of boundary conditions), the
shortest path can be obtained as long as it can be matched to the end point on the two-dimensional
path graph of the reference and test speech frames. Finally, after MFCC eigenvalue is extracted, the
test value is matched according to DTW algorithm and reference template.

3. Introduction of the Hardware System

In the design of this system, Zynq is the main device and WM8731 is the slave device.

3.1. Zynq FPGA AX7020 Module

The highlight of the Zynq FPGA is that the FPGA contains a complete ARM processing subsystem,
and the processor subsystem is integrated with a memory controller and a large number of peripherals.
ZYNQ is the first product to combine high-performance ARM processor and high-performance FPGA
in a single chip, reducing the difficulty of software development, high level of integration, making
communication easier, and thus having great ad-:vantages in the design of speech recognition system.

3.2. Audio Module AN831

This system uses the audio module AN831, which uses the audio chip WM8731 [10]. In this system,
the chip mainly performs the Function of A/D and D/A conversion during the acquisition and
playback of sound signals. In this system, the sampling frequency of ADC and DAC of this chip is set
at 48KHz. The inside of WM8731 has 11 registers. ADCDAT, DACDAT, ADCLRC and DACLRC are synchronized with the bit clock CLK, and data transmission is conducted once at the falling edge of each BCLK. BCLK, DACDAT, DACLRC and ADCLRC are input signals of WM8731. ADCDAT is the output signal of WM8731. The Structure diagram of audio module AN831 is as follows in figure 2:

![Structure diagram of audio module AN831](image1)

**Figure 2.** Structure diagram of audio module AN831.

4. The Speech Recognition Algorithm is Implemented on Zynq FPGA

MFCC process is implemented by Verilog hardware description language (HDL) [11]. It includes Hamming window, FFT, power spectrum, Mel-Filter, logarithm, and DCT processes. The overall block diagram is as follows. I2S-IP core can be used to connect microphone and audio output module. And a DMA is used to process the stream data. The hardware part of the overall framework is as follows in figure 3:

![The hardware part of the overall framework](image2)

**Figure 3.** The hardware part of the overall framework.

Before implementing the algorithm, you need to set up a hardware environment in Vivado. Select EMIO to configure WM8731, and turn on GPIO EMIO to connect the buttons and LED lights. And set the width of EMIO to 3, in which two lights are used for recording and playing indicator light, and one is used for pressing the button to control recording and playing. The FCLK_CLK1 frequency is set to 12.288MHz. Turn on an interrupt to connect a middle break in DMA. Open read and write ports, select SG mode, set stream data Width to 32 since audio data is 32 bits wide. Add THE ADI’s I2S-IP.

In the hardware environment of Zynq FPGA, the speech recognition program is input, and the software performs pre-emphasis, framing, windowing, FFT, VED endpoint detection, DTW calculation and other steps successively. A total of 100 voice samples were record-ed from 0 to 9
Chinese pronunciations by 10 different voices. Five groups of voices were set as training groups for learning and the remaining five for recognition. The results are as follows in table 1.

| Voice type | Test times | Correct number | Recognition rate | Average recognition rate |
|------------|------------|----------------|------------------|--------------------------|
| Voice 1    | 30         | 25             | 83.33%           |                          |
| Voice 2    | 30         | 23             | 76.67%           |                          |
| Voice 3    | 30         | 23             | 76.67%           | 80.67%                   |
| Voice 4    | 30         | 24             | 80%              |                          |
| Voice 5    | 30         | 26             | 86.67%           |                          |

In the whole project of Matlab, including voicebox_test.m, Voice-box_train.m, MFCC.m, Hamming.m and other files. The original 100 samples are still used. Similarly, the original training group is still the training group and the learning group is still the learning group. The results of running the test are as follows in table 2.

| Voice type | Test times | Correct number | Recognition rate | Average recognition rate |
|------------|------------|----------------|------------------|--------------------------|
| Voice 1    | 30         | 27             | 90%              |                          |
| Voice 2    | 30         | 25             | 83.33%           |                          |
| Voice 3    | 30         | 25             | 83.33%           | 80.67%                   |
| Voice 4    | 30         | 26             | 86.67%           |                          |
| Voice 5    | 30         | 27             | 90%              |                          |

From the test results, the overall recognition rate of Matlab simulation is higher than that of Zynq FPGA platform simulation.

5. Conclusion
At the beginning of this paper, the recognition process of the speech recognition system is introduced as a whole. Meanwhile, the Zynq FPGA platform and the speech acquisition module AN831 are also introduced in detail. Then the speech recognition algorithm is simulated in Matlab, and finally implemented on the Zynq FPGA platform and tested. By comparing the recognition rate of simulation and platform test, it is found that the recognition rate of simulation is higher on the whole, which may be caused by the limited platform speed. However, Zynq FPGA platform, which combines high-performance ARM processor and high-performance FPGA in a single chip, enables speech recognition through this platform, and the recognition rate is much higher than that of traditional speech recognition technology without major changes in the algorithm.

References
[1] Zhang Z X, Han H L and Xue H W 2008 Analysis of speech recognition technology Computer Development and Application 21 (172) 33-35.
[2] Han D B, Zeng B, Ge L Q, et al. 2007 Voice control intelligent car design based on SPCE061A Computer Measurement and Control (9) 1183-1185.
[3] I Skog and P Handel 2009 In-car positioning and navigation technologies: A survey IEEE Transactions on Intelligent Transportation Systems (01) 4-21.
[4] Cheng L J, Jing X X and Yang H Y 2016 Design of on-board voice recognition system based on SOC FPGA Journal of Guilin University of Electronic Science and Technology (6).
[5] Li J C 2009 Research on Speech Recognition Algorithm Based on Voice-Controlled Vehicle chapter 3 pp 16-17.

[6] Wang Y P 2007 Study on Improving HMM Speech Recognition Algorithm by Genetic Algorithm chapter 2 p 10.

[7] Choo C, Chang Y U, Member, et al. 2015 FPGA-based hardware accelerator for feature extraction in automatic speech recognition Journal of Information and Communication Convergence Engineering 13 (3).

[8] Kepuska V Z and Hussien A 2015 Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions Journal of Computer and Communications 3 (6) 1-9.

[9] Jing X X and Shi X 2012 Speech recognition based on efficient DTW algorithm and its DSP implementation Procedia Engineering 9 (8) 832-836.

[10] Wang H R 2014 Design of audio acquisition and playback system based on SOPC embedded digital storage Journal of Shandong Agricultural University (Natural Science Edition) (2) 223-228.

[11] MFCC Project: C-Based Algorithm of MFCC https://code.google.com/p/mfcc-umbc/wiki/MFCCIntro.