The Implementation of Speaker Recognition System with the Xilinx® FPGA Device

Chin-TA CHEN and Chih-Chung YANG
Electrical and Electronic Engineering College of Zhaoqing University, Zhaoqing City, Guandong province, China, 526061
*Corresponding author

Keywords: Speaker recognition, Euclidean distance, FPGA.

Abstract. This paper proposes a speaker identification system with a field-programmable gate array (FPGA) device of the Xilinx® Spartan III platform. The Fast Fourier Transformation (FFT) is employed to extract the features of the speeches from a group of Mandarin speakers. The Long Times Average Spectrum (LTA) is identified by calculating the Euclidean distance and matching up with the minimum distance found. Numerical experiments verify the efficiency and accuracy with the simulation in order to code the hardware digital circuits based on the proposed model. This speaker identification system applying the FPGA for a feature extraction and pattern recognition has been implemented. Test signals were generated using ModelSim® and Matlab® software. Comparison with results in Matlab® simulation, the proposed system can achieve a pretty good accuracy on identifying speakers with the FPGA device.

Introduction

Nowadays, devices based on embedded systems can be seen almost everywhere in our daily life. Most of them are multi-functional for people’s convenience. Speech coding and recognition technology has gradually become an integrated part of everyday life. The features of FPGA devices are most suitable for designing and developing complicated systems with multi-function. There are many advantages of FPGA devices, such as increasingly supported hardware, faster and faster working clock rate, shorter development time, avoiding design risks like ASIC, reducing prototype product costs, and providing powerful architecture for either DSP or microprocessor. Therefore, how to apply the pattern recognition technology to speaker identification application with hardware of the FPGA is the main focus of this paper.

With the help from the powerful simulation capability of the Matlab, a speaker identification simulation system is established firstly. Then, the analysis of the optimal identification rate on system accuracy, as well as the preliminary results of the simulation outcomes as a whole are discussed. Regarding the optimization, the speech recognition rate with analog samples is firstly compared in the similar setups with different lengths of window frames on digitized sampling points. Since the hardware of the FPGA does not support the operation of floating-point number, it introduces round-off numerical errors while performing arithmetic calculation that greatly affects the outcomes in the classification process. To alleviate the influence, the identification system can be improved on accuracy by the analysis of the amplification coefficients. Based on the analysis, the simulation and measurement results of the speaker identification system are analyzed. In order to make it useful in the practical application with wide range in the real world, the Xilinx ISE (Integrated Synthesis Environment) with settings for the Spartan III is developed to implement the system. The output of the system is verified with the MATLAB simulation results.

Background

Speaker Recognition System Architecture

Speaker identification is designed to make machines, like computer or microprocessor system, to recognize what we are talking, so a speaker identification system should be designed completely
from the very beginning. First, the analog signal of speech sound is recorded by microphone, and
converted to digital signal by analog/digital converter (A/D). Signals are per-processed by the
endpoint detection method to remove unused silences and retain valid signals to find the beginning
and end of the speaker's signal. After a windowing processing, the speech signal is cut into many
small frames, then the feature extraction process is applied to extract each person's voice features.
And the speech samples are grouped into either the training or the testing sets. If a speech signal is
assigned to the training set, it is stored in the database of features. If a speech signal is assigned to
the testing set, the speaker pattern which is the closest to the test signal is labeled as the output
result according to the identification system working flow by the feature comparison process.

**Long Time Average Spectrum**

In the study of human phonetics, it has been found that the more speech corpus there is the more
stable and convergent the amplitude of speech frequency will be. Its equation is as follows:

\[
M = \frac{1}{NF} \sum_{i=1}^{NF} X_i
\]

Where \( M \) is the Long Time Average Spectrum, \( NF \) is the total number of sound frames \( i \) is the
sound frame index, and \( X_i \) is the corresponding spectrum vector of the specified sound frame \( i \).

In this study, 256 points are taken from each speech sound frame and 128 spectrum vectors are
obtained by the FFT algorithm. The average spectrum of different frame numbers is also different,
but they tend to be stable in a certain range. This feature is used to analyze the individual’s voice
characteristics of each person, and the same number of frames is used for analysis.

**Bayesian Rule**

In the speaker recognition system, speaker model training is to find out those unique feature vectors’
distribution from speakers’ sentences. Then, the Bayesian theorem is applied to match the speakers’
model with the maximum likelihood to the tester. The decision-making method derived from
conditional probability theory is called prior probability for the subjective probability predicted by
the decision-maker for each environmental condition. After further experimental verification, the
subjective probability is transformed into the posterior probability by Bayesian rule. By means of
the probability required by Bayesian law, its accuracy is relatively objective and trustworthy.

According to the definition of Bayesian law, a theorem of conditional probability and edge
probability of random events A and B is given. Suppose the eigenvectors \( \{A_1,A_2,…,A_k\} \) is one of
the partitions of sample space \( S, P(A_i) > 0, i=1,2,……k \), For any other eigenvector B:

\[
P(A_j | B) = \frac{P(A_j)P(B | A_j)}{P(B)} = \frac{P(A_j)P(B | A_j)}{\sum_{i=1}^{k} P(A_j)P(B | A_j)}
\]

Where \( P(A | B) \) is the conditional possibility of A when B occurs.

The similarity between test samples and training samples is obtained by Bayesian rule. The
similarity between the test samples and the training samples is obtained by Bayesian rule, and the
minimum difference between the test samples and the training samples is found by comparing the
test samples with the training samples, and the best path for the identification results is obtained.

**Speaker Identification System Based on FPGA Design and Simulation Results**

In this implementation, the main speaker identification system is designed with the FPGA hardware
device. The Xilinx ISE design suite greatly simplifies the development process on the Spartan III.
In the meantime, both the feature extraction and speaker identification comparison processes are
integrated in our system architecture. The functional verification is verified by applying ModelSim
simulation system waveform.
**Introduction to Design Environment.** As Fig.1, the feature extraction and identification comparison function of the speaker recognition system is implemented using the FPGA, which is supported with the special software Xilinx ISE. The ModelSim, which is a HDL (Hardware Description Language) simulation environment by Mentor Graphic, is adopted for the Spartan III by Xilinx. Firstly, Xilinx ISE is used to design the system functions, and modelSim is used to simulate the system. Finally, it is transferred to the FPGA for verification.

This paper designs the speaker identification system from feature extraction to identification comparison. The software tools for test and verification are Xilinx-ISE and ModelSim. As Fig. 2

![Figure 2. Xilinx ISE: Integrated Synthesis Environment.](image)

**Preliminary Simulation Result.** As Fig. 3, taking the speaker signal numbered 1 as the input sample for test, the system responds with identification results of the imported speaker to be number 1, which is correct as expected. The measured results are identified correctly. The system spent 0.598 seconds for this measuring process. The difference between the measured results and MATLAB is 99.97%. Table 1 showed the hardware resource utilization.

![Figure 3. Speaker Identification Results of Test Number 1.](image)

**Table 1. Hardware resource utilization.**

<table>
<thead>
<tr>
<th>Device Utilization Summary</th>
<th>Logic</th>
<th>Used</th>
<th>Available</th>
<th>Utilization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slices</td>
<td>7653</td>
<td>7680</td>
<td></td>
<td>99%</td>
</tr>
<tr>
<td>Slice Flip Flops</td>
<td>802</td>
<td>15360</td>
<td></td>
<td>5%</td>
</tr>
<tr>
<td>4 input LUTS</td>
<td>13843</td>
<td>15360</td>
<td></td>
<td>90%</td>
</tr>
<tr>
<td>Bonded IOBs</td>
<td>140</td>
<td>173</td>
<td></td>
<td>80%</td>
</tr>
<tr>
<td>MULT18X18s</td>
<td>22</td>
<td>24</td>
<td></td>
<td>91%</td>
</tr>
<tr>
<td>GCLKs</td>
<td>2</td>
<td>8</td>
<td></td>
<td>25%</td>
</tr>
</tbody>
</table>
Conclusion

The main purpose of this paper is to implement a speaker identification system with the FPGA hardware, and to analyze the optimal recognition rate by comparing different frame numbers and sampling points. In the meanwhile, the effect of errors introducing by the fixed-point operation is explored on extracting analog feature vectors with the FPGA, as well as the accuracy of feature vectors for different amplification coefficients. This speaker identification system designed by Xilinx ISE software can take the measuring time from 0.591 seconds to 0.848 seconds with the implementation of FPGA. Therefore, the higher the resource utilization rate, the easier the impact on the hardware will be. As to the identification rate, the Euclidean distance of the feature vector is used to calculate the identification rate. At present, the recognition rate is only estimated at about 60%. If the Mahalanobis distance is used, the rate can be improved to 80%, and quadratic distance will increase another 10% of the recognition rate. However, there are some difficulties in the realization of some principle formulas and it takes a lot of time to study. Therefore, in addition to understanding the speaker identification deeply and improving the recognition rate, it can also make research and design towards word recognition that is widely used in daily life.

References