The Implementation of Chinese and English Bilingual Speech Recognition System-on-Chip

Shunli Ding, Hong Cao, and Jia Liu

Abstract—This paper presents a high performance embedded non-specific, medium vocabulary Chinese-English bilingual speech recognition system using the continuous density hidden Markov model and a two-pass search strategy based on a 16-bit fixed-point digital signal processing (DSP). This system selecting MFCC parameters as recognition feature. Improve the system real-time through a dedicated hardware circuit design. We extract specialized hardware co-processing circuit characterized by structural features through abstracting algorithm critical path that speedy computation concerns much in speech recognition, so as to greatly enhance the overall performance by little chip costs. The experimental result suggested that the identification rate is 97.6% when entries sum is 600. The characteristic storage space was reduced 31%, and the real-time rate of two stage identification is 0.7.

Index Terms—Speech recognition, system on chip, dsp, model selection.

I. INTRODUCTION

With the development of semi-conductor industry and speech recognition and coding technology, the demand for embedded speech recognition product in the field of intelligent electronics, toys and automotive electronics has been increasing. However, within the limit in hardware resources and algorithm complexity, traditional embedded speech recognition product is only applied to the low end market, being far away from meeting the demand of speech recognition product in the field of electronic market. This paper implemented speaker-independent chinese-english bilingual speech recognition system with high performance through improving algorithm and the deep hardware and software collaborative design.

Because of the limit in arithmetic speed and storage space, the paper employed a two stage identification strategy based on CHMM [1]-[3] and chose proper eigenvector and model to reduce the computing complexity and storage space in order to improve the efficiency of the recognition and real-time capabilities. Consequently, the high performance, speaker-independent, medium vocabulary chinese-english bilingual speech recognition system was developed in the embedded system [4]-[7] with the 16 bits fixed-point DSP as the core. This system designed dedicated co-processor circuit on chip based on the characteristics of speech recognition algorithm, greatly improving the system performance.

II. HARDWARE PLATFORM

The hardware platform of the speech recognition system-on-chip is designed by the laboratory and chip design companies, and its structure is shown in Fig. 1.

![Speech recognition system-on-chip structure.](image)

This chip with Harvard bus, four pipelined architecture, it is composed of a 16 bit DSP, 64KB RAM, 256 KB ROM. Required for speech recognition in the audio amplification circuit and the ADC, DAC, the automatic gain control (AGC) circuit and other modules are integrated on the chip. The chips specifically designed for speech recognition applications, designed to effectively reduce hardware overhead, lower application costs through hardware and software co-design.

As the speech recognition algorithm is in high demand for embedded system chip resources, in the dissertation, some logic abstraction has been applied to the critical path which requires more computing performance, and the specific hardware co-processor circuit with the structural characteristics has been extracted. Software algorithm has been treated as hardware, the overall system operation efficiency has been largely improved at a smaller chip cost increase. As a result, the comprehensive price quality will be leading in the domestic industry. The introduction for specific co-processor circuit function module was demonstrated in Fig. 2.

Array unit: to complete the function of sequence accumulation, was able to improve FFT and MFCC features effectively as well as increasing the speed of GMM calculating.
Rapid mover of data: to transfer the data from OTP/SRAM to SRAM rapidly, so as to improve the efficiency of the model and netlist.

Seeker for extreme data: to seek and identify the location of maximum and minimum value, thus sorting capabilities be realized by coordinating with the software circulation and data exchange.

32-bit divider: to complete 32-bit signed/unsigned number division. It would be time-consuming if the chip division was operated by software, and the calculating efficiency would be increased effectively by using hardware divider.

Viterbi arithmetic unit: to do viterbi search arithmetic. In the whole speech recognition algorithm, the time consumed by Viterbi arithmetic accounts for 30% to 70% and the ratio becomes more significant with the increase of the terms number. Consequently, the comprehensive systematic performance was significantly improved by substituting hardware Viterbi arithmetic for software.

III. IMPLEMENTATION OF SOC

The basic framework of speech recognition system is demonstrated in Fig. 3. Recognition process can be roughly divided into feature extraction, model training and identify network decoding. For isolate word speech recognition[8],[9], the selection of acoustic model has nothing to do with the hardware platform, so the model training can be completed on PC computer. But the feature extraction and network decoding must be implemented on hardware platform.

Speech recognition system mainly has three kinds of characteristic parameters: LPCC, MFCC and PLPC [10]-[13]. We chose MFCC in this paper according to the demand of recognition performance and limit of hardware resource. Storage space and recognition time are increased along with the increase of the characteristic dimension, therefore needing to choose the proper characteristic parameters for the sake of good real-time performance in the embedded recognition system, rather than using 39 dimensions feather composed by MFCC [14],[15], first-order difference, second-order difference, normalized transient energy and differential power. We identified the contribution of feature components to recognition performance, and chose the 27 dimensions feather components with biggest contribution, effectively reducing the SRAM area resources consumption by storage characteristic parameters.

We must select the right acoustic model because of limit in operating capacity and store space of unspecific speech recognition system on chip. If the simpler monophon model [16]-[19] is selected, although it will take up less space, it will affect the operating accuracy. On the contrary, if the more accurate triphone model is selected, although it can satisfy the high recognition rate, it needs the large store space and operating capacity which embedded system can’t bear.

So this paper introduced a two phase recognition algorithm in order to assure the high recognition rate and low consume of hardware. As shown in Fig. 4, the monophon model was selected in the first stage to get the multiple candidate entries by identifying quickly, and the triphone model was selected in the second phase to get the final result based on the multiple candidate entries. The two phases were relatively independent. In the second phase, algorithm can reuse the memory; in the first phase, the multiple candidate entries which is got by rapid identification were few, and the calculation of viterbi decoding is greatly reduced in the second phase. This design of two phase identification frame took full advantage of store space and improved real time rate of system.

IV. SOFTWARE DESIGN

The software architecture has adopted a hierarchical design so as to optimize the portability of the embedded system, which is demonstrated in Fig. 5.

1) The driver layer, which exists in the form of interrupt service in more cases, such as A / D, D / A etc, contacts directly with the hardware more. The design of driver layer has improved the safety and reliability of the system.

2) The service layer: basic math functions required repeatedly in the algorithm are encapsulated in the Universal module layer, so that the speed of the whole system operation can be improved effectively through the optimization of the basic functions.

3) The function layer, including the feature extraction, post processing characteristics, GMM score calculation, the Viterbi decoding etc, is the core algorithm layer of the speech recognition, determines the recognition rate and timeliness of the system.
4) The application layer is the highest level of system-on-chip, controls the system overall framework and calls different modules. Application layer does not care about the specific method of the functional modules, only to call the appropriate function to achieve specific functions due to the particular need. The layered design software has improved the maintainability and scalability of the system-on-a-chip effectively, large-scale changes to the system subject are not needed when the function or module adjusts, thus reduces the difficulty of the secondary development.

![Software division](image)

V. PERFORMANCE

As mentioned before, the study found the amount of the GMM score calculation and Viterbi decoding operations in a high proportion of the speech recognition algorithm, in this article, dedicated hardware co-processing circuit was designed to enhance the system performance. Examination has been made to test the improvement of computing array and Viterbi hardware circuit after the completion of System-on-chip system. The experimental results are as follows:

Under the minimum SRAM application conditions, with 30 terms, the real-time rate has reached to 0.87 when using computing array, if not, the rate would have increased to 2.23. It means the use of computing array improved the GMM score calculation speed significantly.

In order to test the effect of the Viterbi hardware circuit module on the overall system performance, the article has done the test based on 100 words, 600 words and 2000 words respectively, as the results shown in Table I.

<table>
<thead>
<tr>
<th>TABLE I: PERFORMANCE COMPARISON OF HARDWARE VITERBI AND SOFTWARE VITERBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardware Viterbi</td>
</tr>
<tr>
<td>------------------</td>
</tr>
<tr>
<td>100 Words 600 Words 2000 Words</td>
</tr>
</tbody>
</table>

From the above table, the real-time rate will have five times increase or so by using Viterbi hardware circuit module compared to software operations only. The improvement of embedded system overall performance is very impressive.

VI. CONCLUSION

This paper introduced a Embedded chinese-english bilingual speech recognition system based on 16bit fixed-point DSP. This system used a two phase recognition algorithm. The 208 state monophon model was selected in the first phase to identify quickly and get eight candidate entries; the 358 state triphone model was selected in the second phase to get the final result. Identification capacity has been improved, memory space and computational complexity have been reduced through hardware specialized module design, select of feature vector and model state share. The experimental result suggested that the identification rate is 97.6% when entries sum is 600. The characteristic storage space was reduced 31%, and the real time rate of two phase identification.

ACKNOWLEDGMENT

Thanks to Zhang Weijiang and MAX who gave help and guidance during the process of system implementation.

REFERENCES


Shunli Ding was born in Jining Shandong province in 1984.9. Now he is a postgraduate student in Institute of Microelectronics of Tsinghua University, Beijing, China; his research aspect is speech signal processing and speech recognition.

Hong Cao was born in Gansu in 1984. He got Master's degree in Institute of Microelectronics of Tsinghua University, Beijing, China. Now he is working in HELIOS-ADSP Technology Co.Ltd his research aspect is speech signal processing and speech recognition.

Jia Liu was born in 1954, in Beijing. He is a professor and Doctor guider in EE Department of Tsinghua University, Research Aspects: Speech Recognition/ Speech coding, speech recognition chip design, Multimedia signal communication system.