A 40-nm 144-mW VLSI Processor for Real-time 60-kWord Continuous Speech Recognition

Guangji He,Takanobu Sugahara,Tsuyoshi Fujinaga,Yuki Miyamoto,
Hiroki Noguchi, Shintaro Izumi, Hiroshi Kawaguchi, and Masahiko Yoshimoto

Kobe University, Kobe, 657-8501 Japan
achilles@cs28.cs.kobe-u.ac.jp

Abstract - **We have developed a low-power VLSI chip for 60-kWord real-time continuous speech recognition based on a context-dependent Hidden Markov Model (HMM). Our implementation includes a cache architecture using locality of speech recognition, beam pruning using a dynamic threshold, two-stage language model searching, highly parallel Gaussian Mixture Model (GMM) computation based on the mixture level, a variable-frame look-ahead scheme, and elastic pipeline operation between the Viterbi transition and GMM processing. Results show that our implementation achieves 95% bandwidth reduction (70.86 MB/s) and 78% required frequency reduction (126.5 MHz). The test chip, fabricated using 40 nm CMOS technology, contains 1.9 M transistors for logic and 7.8 Mbit on-chip memory. It dissipates 144 mW at 126.5 MHz and 1.1 V for 60 kWord real-time continuous speech recognition.**

I Introduction

A hardware approach for large vocabulary continutous speech recognition (LVCSR), with implementation by VLSI or an FPGA is requested recently, especially for mobile equipment and the intelligent robot, because of its advantageous processing speed and power consumption. Lin et al. investigated FPGA implementations for 5k-word speech recognition [1], but it consumes too much power and is not cost-effective as it needs two FPGAs .Yoshizawa et al. proposed a scalable architecture for speech recognition[2], but their chip costs 136mW even with a limited vocabulary of 800-word. Choi et al. investigated FPGA implementations for 5k-word and 20k-word[3, 4], they implemented special memory interface for several part of the recognition engine to apply optimized DRAM access, which reduces the delay for loading data, but they did not reduce the large amount of external DRAM access which will cause a lot of power consumption for IO. Comparison of power consumption among the recent published hardware-based speech recognizers is shown in Fig.1. To date, the hardware approach has never achieved real-time operation with a 60-k word language model because of the numerous computation work-load and external memory bandwidth. For low power and real-time 60-k processing, we have to reduce both the memory bandwidth and operating clock frequency.

II. Algorithm

Figure. 2 presents the speech recognition flow with the HMM algorithm. The following items describe concrete stages. **Step 1:** Feature vector extraction: a feature vector is extracted on a frame-by-frame basis. **Step 2:** GMM calculation: a phonemic-model GMM is read and GMM probability, log $[b_j(x_t)]$, is calculated for all active state nodes. **Step 3:** Viterbi transition: $\delta_t(j)$ is calculated for all active state nodes using GMM probabilities. **Step 4:** Beam pruning: according to the beam width, active state nodes having a higher score (accumulated probability) are selected; the others are dumped. **Step 5:** Output sentence: The word-end state and having the maximum score is output as a speech recognition result after final-frame calculation and determination of the transition sequence.

III. Architecture

The overall chip architecture is depicted in Fig.3. The proposed architecture is comprised of a global Sequencer, GMM-core, Viterbi-core, double GMM result Buffer to support pipeline operation. The GMM core have 16 mixture processing blocks each of which has 1.6Kb register to preserve the mixture parameter. All the blocks are processed simultaneously for the look-ahead frames which is saved in MFCC buffer. The parameters will be reused until all the look-ahead frames are processed. The mixture results will soon be caculated by the add-log processor based on a look-up table. The mixture computation, Add-log calculation, and parameter reading are processed in pipeline. Figure.4 shows the viterbi architecture, we find that the Active Node, Active Node Map and the Bi-gram accounts for 96% of the memory bandwidth for viterbi transition as shown in Fig.5. So we introduce all the active node data and part of the bi-gram and active node map data into the cache memory by using the locality of speech recognition that some of the data which has been used for this frame may have a high probability to be reused in the following frames. We maximize the cache memory size of bi-gram and active node map to 0.73Mbit, which can give a hit rate of 75%.

V. Implementation

The chip has been fabricated in 40nm CMOS technology as shown in Fig.6. It occupies 2.2×2.5 mm^2 containing 1.9M transistors for Logic and 7.75Mbit on-chip SRAM (TableI). The clock gating is implemented in GMM result RAM and GMM Core. Figure.7 shows a measured data of power consumption versus operating frequencies versus beam_width. The bigger beam_width can give higher accuracy, but it will also cause larger computation, This chip can work at 126.5MHz operation for a beam_width of 3000 while the power consumption is 144mw with a accuracy of

978-1-4673-3030-5/13/$31.00 ©2013 IEEE

91.39% and 140MHz for a beam_width of 3800 while the power consumption is 204.8mW with a accuracy of 91.92%.

Acknowledgements

The VLSI chip in this study was fabricated in the chip fabrication program of VLSI Design and Education Center (VDEC), the University of Tokyo. This development was performed by the author for STARC as part of the Japanese Ministry of Economy, Trade and Industry sponsored “Silicon Implementation Support Program for Next Generation Semiconductor Circuit Architectures”.

References

[1]S. Yoshizawa et al., “Scalable architecture for word HMM-based speech recognition and implementation in complete system,” *IEEE Trans. Circuits Syst. I*, Vol. 53, pp. 70-77, 2006.
[2]E. C. Lin et al., “A multi-FPGA 10x-real-time high-speed search engine for a 5000-word vocabulary speech recognizer,” *Proc. 17th ACM/SIGDA Int. Symp. FPGA*, 2009.
[3]Y. Choi et al., “A real-time FPGA-based 20,000-word speech recognizer with optimized DRAM access,” *IEEE Trans. Circuits Syst. I*, Vol. 57, pp. 95-105, 2006.
[4]Y. Choi et al., “FPGA-based implementation of a real-time 5000-word continuous speech recognizer,” *Proc. 16th Eur. Signal Process. Conf.*, 2008.
[5]T.Ma et al., “Novel ci-backoff scheme for real-time embedded speech recognition,” Proc. IEEE ICASSP, 2010.
[6]G. He et al., “A 40 nm 144 mW VLSI processor for realtime 60 kword continuous speech recognition,” Proc. IEEE CICC, 2011.

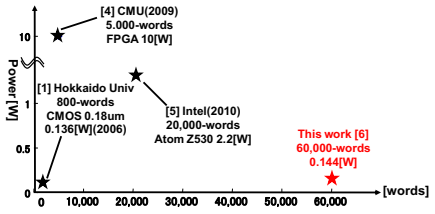

Fig. 1. Comparison of power consumption.

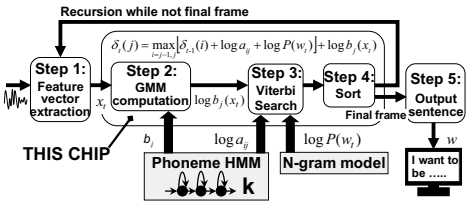

Fig. 2 Speech recognition flow with HMM algorithm computation

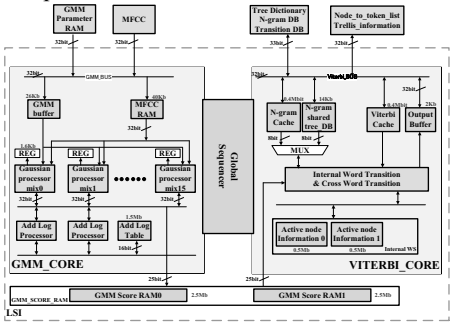

Fig.3 Proposed speech recognition adrchitecture

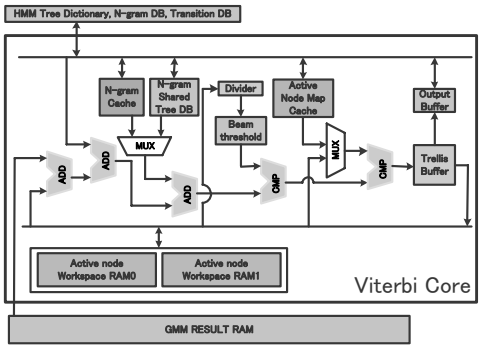

Fig.4 Viterbi Cache Architecture

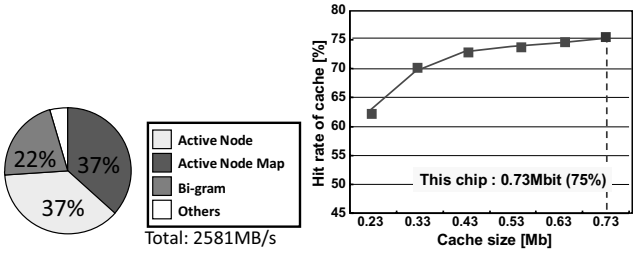

Fig.5 External memory bandwidth of Viterbi and Cache hit rate

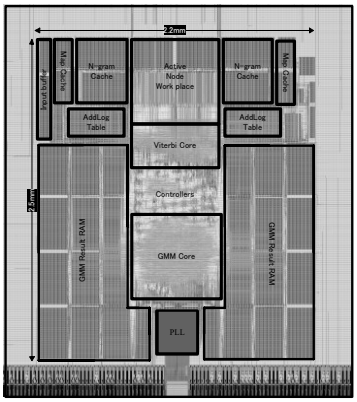

Fig. 6. Chip microphotograph.

TABLE I
Summary of chip implementation

Process Technology		40-nm CMOS
Core area		2.2 mm × 2.5 mm
Chip area		5 mm × 5 mm
Transister Count (Logic)	GMM	1.1 M
	Viterbi	0.3 M
	Other	0.5 M
	Total	1.9 M
On-Chip Memory (SRAM)		7.8 Mbit
Supply voltage		1.1V
I/O voltage		3.3V
Evaluation environment		ADVAN SOC Tester System
Operating Frequency		126.5 MHz for 60 kWord real-time processing
Measured Power (core)		144 mW
Measured Power (IO)		58.2 mW
Leakage current		2.28 mA

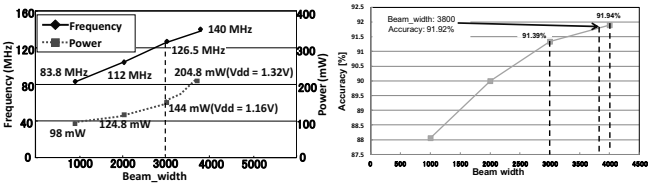

Fig. 7. Measurement results of frequency vs. power consumption vs. beam-width vs recognition accuracy.