A Gaussian Probability Accelerator for SPHINX 3

Binu K. Mathew, Al Davis, Zhen Fang

UUCS-03-02

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

November 18, 2002
Revised on July 22, 2003

Abstract

Accurate real-time speech recognition is not currently possible in the mobile embedded space where the need for natural voice interfaces is clearly important. The continuous nature of speech recognition coupled with an inherently large working set creates significant cache interference with other processes. Hence real-time recognition is problematic even on high-performance general-purpose platforms. This paper provides a detailed analysis of CMU's latest speech recognizer (Sphinx 3.2), identifies three distinct processing phases, and quantifies the architectural requirements for each phase. Several optimizations are then described which expose parallelism and drastically reduce the bandwidth and power requirements for real-time recognition. A special-purpose accelerator for the dominant Gaussian probability phase is developed for a 0.25μ CMOS process which is then analyzed and compared with Sphinx's measured energy and performance on a 0.13μ 2.4 GHz Pentium4 system. The results show an improvement in power consumption by a factor of 29 at equivalent processing throughput. However after normalizing for process, the special-purpose approach has twice the throughput, and consumes 104 times less energy than the general-purpose accelerator. The energy-delay product is a better comparison metric due to the inherent design trade-offs between energy consumption and performance. The energy-delay product of the special-purpose approach is 196 times better than the Pentium4. These results provide strong evidence that real-time large vocabulary speech recognition can be done within a power budget commensurate with embedded processing using today's technology.