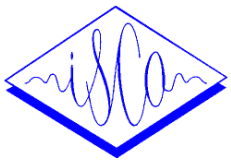

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

AN ACCELERATOR FOR HIGH-SPEED SPOKEN WORD-SPOTTING AND NOISE IMMUNITY LEARNING SYSTEM

Hiroyuki Tsuboi, Hiroshi Kanazawa and Yoichi Takebayashi

Toshiba Corporation, Research & Development Center
Saiwai-ku, Kawasaki, 210 Japan

ABSTRACT

An accelerator utilizing four digital signal processors (DSPs) has been developed to facilitate real-time speech recognition. The accelerator has been implemented in a real-time robust speaker-independent word recognition system. This system employs word-spotting based on Noise Immunity Learning to avoid word boundary detection errors and to increase recognition accuracy in noisy environments. The accelerator board, including four DSPs with shared memory, has 132 MFLOPS peak performance. Since more than 90% of the computational load is inner product calculation, the DSPs share a vocabulary for the purpose of load-balancing. The architecture of the accelerator consists of off-the-shelf components connected in such a way for improved performance in the application. The speed of a single accelerator was shown to be approximately 20 times faster than that of a current high speed workstation with 32 MFLOPS peak performance.

I. INTRODUCTION

In the past decade, there have been many efforts directed toward improving the performance of continuous or large vocabulary speech recognition for natural human-computer interaction. In order to achieve high recognition accuracy, large amounts of computation are required for various complex processes associated with analysis, recognition, and learning. Several speech recognition systems, such as SPHINX and OSPREY, have employed their own special-purpose accelerators [1][2]. Other systems, such as ASPEN and the system on MARS machine, have used general-purpose recognition accelerators [3][4].

These systems have produced good recognition results in noise-free environments. However, their performance is drastically reduced with the addition of background noise, which is present in real-world applications. Noise robustness is crucial for human-computer interactive systems that support speech input. In response to this need, the authors have implemented a speech recognition system employing high-performance word-spotting based on the Multiple Similarity method and Noise Immunity Learning [5].

This system requires greater computational power than conventional recognition systems. The newly implemented methods require iterative computations for spectral analysis and similarity calculations. Therefore, the digital signal processor (DSP) accelerator has been employed to increase processing speed and to enable real-time recognition.

In this paper, a noise-robust speech recognition method based on word-spotting with Noise Immunity Learning and its required computation are first described. The design approach to the accelerator and its architecture are presented. The system environment and process flow are discussed, and evaluation results are shown.

II. WORD-SPOTTING BASED ON NOISE IMMUNITY LEARNING

Noise robustness is a crucial factor in the realization of practical, real-world applications of speech input. Under noisy environments, recognition accuracy decreases due to speech pattern variations and word boundary detection errors. A word-spotting method has been proposed for the purpose of robust speaker-independent word recognition. In order to avoid word boundary detection errors at the recognition stage, the method employs word-spotting based on the Multiple Similarity, which was shown to be effective for noisy speech data [5].

Figure 1 shows a block diagram of the proposed word-spotting method based on Noise Immunity Learning. In the learning process, noisy speech data is synthesized by mixing pure speech data and noise data. The noisy speech data is used in the learning process to create reliable word reference vectors. The learning process requires much more computation than the recognition process.

Figure 2 shows word-spotting using continuous pattern matching between fixed dimensional word feature vectors and word reference vectors. The word feature vectors are extracted by linearly sampling speech parameters in terms of assumed start and end points. Multiple Similarity values are time-continuously computed using these vectors. The word is spotted and recognized based on the criterion of similarity threshold and maximum similarity.

In the Multiple Similarity method, the similarity value S for the lth class between an input vector X and reference

vectors $\phi_m^{(l)}$ is defined as follows:

$$S^{(l)}[X] = \sum_{m=1}^{M} \frac{\lambda_m^{(l)} (X, \phi_m^{(l)})^2}{\lambda_1^{(l)} \| X \|^2} \quad (1)$$

Each reference vector for class l is obtained by first creating a covariance matrix from many input vectors X and then performing KL-expansion on the matrix.

In order to enable real-time word-spotting, the similarity computations must be performed during the analysis frame period; for example, 16ms. Most of the computations are inner product calculation; therefore, the DSP is suitable for this purpose. In the learning process, synthesized noisy speech data is used to create reliable word reference vectors for word-spotting. Word feature vectors with maximum similarity values are then automatically extracted by word-spotting. The signal to noise ratio (SNR) of the synthesized noisy data is gradually decreased during the learning process to obtain noise immunity and to improve the performance of word-spotting on noisy data.

Processing for word-spotting and spectral analysis comprises approximately 90% of total computation in the recognition stage; therefore, the authors have developed high speed hardware to account for the computational demand of the new method. The hardware is described in the following sections.

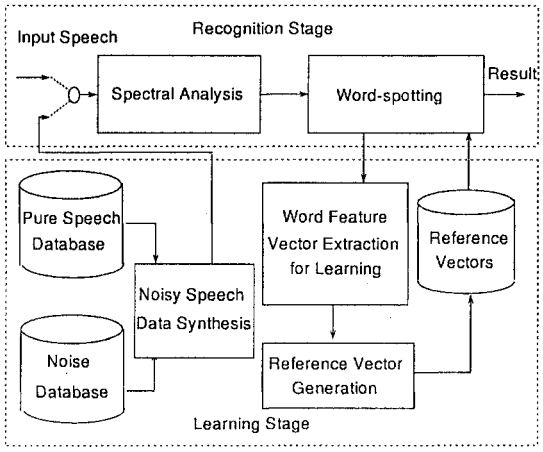

Figure 1 Block diagram for word-spotting based on Noise Immunity Learning

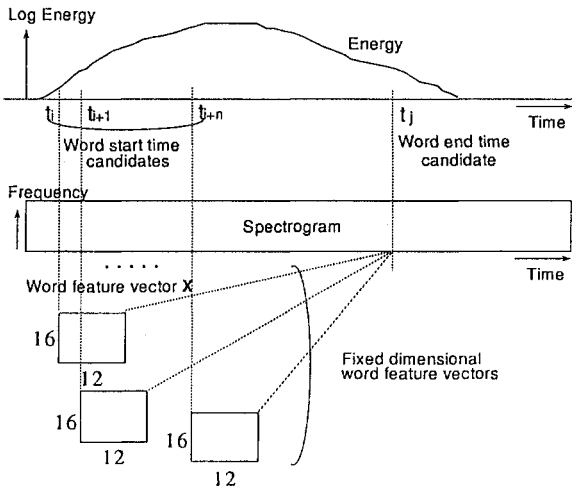

Figure 2 Recognition process of word-spotting

III. ARCHITECTURE

3.1 Approach

In order to support this speech recognition task, high-speed hardware is needed on the workstation. Other computer environments, such as graphics and window systems, could be effectively used for both speech and human interface research. Such an integrated system is possible if an accelerator is employed to facilitate real-time robust speech recognition. The workstation with the accelerator enhances speech research efficiency and enables real-time interaction with speech input and other media such as graphics.

3.2 Design

The following points were considered in the design of the accelerator:

(1) Easy of development

Off-the-shelf DSPs were used to achieve the desired computational power which is not available on current workstations. Most of the computations for spectral analysis, word-spotting and Noise Immunity Learning are inner product calculation; therefore, floating point DSP was used. Multiple DSPs were implemented on the accelerator board.

(2) Expandibility of computational power

The proposed word-spotting method requires a great deal of computation. The amount is proportional to number of words in the recognition task. For increased vocabulary size, a VME-bus interface is used to enable parallel processing of the multiple accelerators for increased computational power.

(3) Applicability to various speech processing systems

The accelerator has been designed not only for real-time noise-robust word-spotting, but also for other speech processing systems. These include off-line simulation of Noise Immunity Learning and various signal processing algorithms. Flexible data access among DSPs is necessary to efficiently implement various speech and signal processing algorithms. Shared memory is accessed by the multiple DSPs to achieve such a function.

3.3 Architecture

The specifications for the accelerator are listed in Table 1, and the architecture is shown in Figure 3.

The DSPs (TMS320C30) were used for high performance floating point calculation. They share a 512-Kbyte data memory and have two private memories: a 32-Kbyte program memory and a 256-Kbyte local data memory. Direct memory access (DMA), controlled by the workstation, was employed for data transfer between the workstation and shared memories.

To achieve exclusive control and minimum overhead for shared memory access, general bus arbitration logic was implemented. It uses the following order of priority: workstation, DMA, DSPs(#0-#3). Each DSP can access shared memory by a request and release interlock operation. The workstation also has this ability through the VME bus interface.

For easy access from the workstation, shared memory and each DSP local memory were mapped to the workstation memory. DSP software was then down-loaded by the workstation.

Table 1 Accelerator specification

Processor	4 DSPs (TMS320C30)
Cycle Time	60ns
Peak Performance	132 MFLOPS
Memory	1.54 Mbytes
Bus	VME Bus

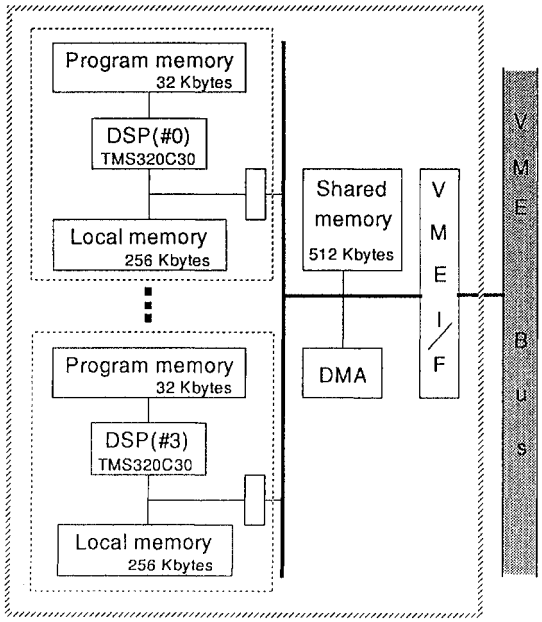

Figure 3 Accelerator architecture

IV. REAL-TIME RECOGNITION SYSTEM

The process flow for the real-time word-spotting system is shown in Figure 4. The system was implemented on a AS4260 workstation with an A/D converter and an accelerator. The recognition process is as follows:

(1) DSP software and the reference vectors for each word are down-loaded through the VME bus to the program memory and local memory, respectively. The A/D converter and accelerator are started.

(2) Digitized input data is transferred to the shared memory of the accelerator by DMA. The input data is transferred to local memory and analyzed to obtain spectral data. DSP(#0) performs this task using Fast Fourier Transform (FFT). The spectral data is then passed to the shared memory.

(3) Each DSP receives the spectral data in its local memory from the shared memory. Similarity values between the input data and reference vectors are calculated.

(4) Recognition results for the maximum similarity and its class are sent by DMA to the workstation through shared memory.

Pipelining is used in the above process and is illustrated in Figure 5. A length of pipe segment was determined according to the amount of data transfer and the response time after the end of utterance. The length of pipe segment is

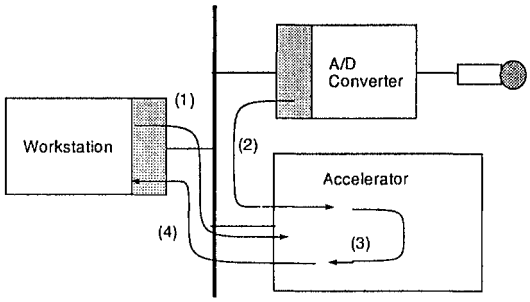

Figure 4 System Configuration

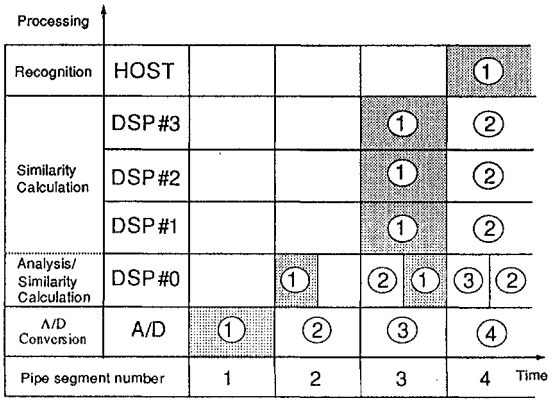

Figure 5 Pipeline Structure

80ms, and the recognition result is obtained 240ms after the end of the input utterance.

Driver routines were developed to facilitate accelerator access from the user program, thus read and write functions are able to send input data to the accelerator and to obtain recognition results. Also, the FIFO buffer of the driver routine and A/D converter allow easy programming for data processing on the workstation by maintaining data synchronization.

V. EVALUATION

Recognition experiments were carried out on thirteen city names uttered by 118 males and 31 females under the conditions listed in Table 2. Noisy speech data was synthesized by mixing the speech data and non-stational concourse noise. Data from 39 of the male speakers was used to evaluate recognition performance, the remaining data was used for Noise Immunity Learning. Recognition scores obtained by word-spotting alone and with the learning, were 88.5% and 99.2%, respectively, for SNR 10dB. Under more noisy conditions, SNR 5dB and 3dB, recognition scores were 96.5% and 95.0%, respectively, with the learning.

Experiments were also done in order to evaluate the effectiveness of the accelerator. The performance of the workstation with and without the accelerator were compared under the condition listed in Table 2. The results are shown in Tables 3. Without the accelerator, the total processing time for each analysis frame period is 250ms; where the spectral analysis using FFT takes 35ms and recognition takes 215ms. In contrast, with the accelerator, the processing time is 12ms; the spectral analysis of DSP(#0) takes 2.5ms and recognition of each DSP(#1-#3) takes 12ms. Therefore, the speed of a single accelerator was approximately 20 times faster than that of a current high speed workstation with 32 MFLOPS peak performance. Since a large majority of the Multiple Similarity calculation is simple multiply and accumulate operation, the performance was improved by four DSPs.

Table 2 Evaluation condition

Sampling frequency	12kHz
Frame period	16ms
Length of pipe segment	80ms
FFT point number	512 points
Feature vectors	192 dimensions (16channel x 12frame)
Vocabulary	13 city names
Accelerator	1 Board
Host machine	AS4260

Table 3 Comparison of Processing time

System configuration	Processing time per frame
AS4260	250ms
AS4260 + Accelerator	12ms

While the above experiments were performed for 13-word vocabulary, real-time word-spotting for increased vocabulary size can be easily realized by using more accelerators. For example, seven accelerators (28 DSPs) enable real-time word-spotting for a 100-word vocabulary. This expandability of the accelerator is useful to apply various speech recognition systems.

VI. CONCLUSION

The accelerator has been successfully implemented on a high-speed workstation to facilitate real-time robust speech recognition based on Noise Immunity Learning. The results indicate that the accelerator effectively speeds up both the recognition and learning processes.

Furthermore, good software environments on the workstation with the accelerator facilitate the development of various real-time speech recognition research systems. It also serves as a valuable tool for research in the area of human-computer interaction.

ACKNOWLEDGEMENTS

The authors wish to thank Mr. Hiroyuki Chimoto for his help in the development of the accelerator and Mrs. Kris Maeda for her assistance in the preparation of this paper.

REFERENCES

[1] R. Bisiani, T. Anantharaman, and L. Butcher, "BEAM: An Accelerator for Speech Recognition," Proc. ICASSP, pp.782-784, May 1989.

[2] A. Sutherland, M. Campbell, Y. Ariki, and M. Jack, "OSPREY: A Transputer Based Continuous Speech Recognition System," Proc. ICASSP, pp.949-952, Apr. 1990.

[3] D. Roe, A. Gorin, and P. Ramesh, "Incorporating Syntax into the Level-Building Algorithm on a Tree-Structured Parallel Computer," Proc. ICASSP, pp.778-781, May 1989.

[4] S. Chatterjee, and P. Agrawal, "Connected Speech Recognition on a Multiple Processor Pipeline," Proc. ICASSP, pp.774-777, May 1989.

[5] H. Kanazawa, and Y. Takebayashi, "A Learning Word-Spotting Method for Speaker Independent Word Recognition Under Noisy Environment," IEICE Tech. Rep., SP89-19, pp.51-58, Jun. 1989 (In Japanese).