高耐雑音 音声認識用アクセラレータの開発

5P−8

坪井 宏之　金沢 博史　竹林 洋一
(株) 東芝　総合研究所

1. まえがき

大語彙や連続音声認識システムのリアルタイム処理を目的として高速ハードウエアの開発が報告されている[1][2]。雑音の少ない環境下の音声を対象とし高い認識率を得ているが、実際に使用される環境下では雑音の影響により認識性能は大幅に下がる。

筆者らはワークステーション(WS)上で音声認識の研究開発に利用するため、DSPを並列化し最大性能132MFLOPSで、高雑音下でも高性能な不特定話者単語認識をリアルタイムに処理できる能力を持つアクセラレータを開発した。本論文ではアクセラレータの構成、性能について述べる。

2. 高耐雑音音声認識

雑音にロバストな単語認識のために雑音免疫学習法とワードスポッティング法を開発し報告した[3]。ワードスポッティング法は入力の分析周期(8msec)ごとに、雑音免疫学習法で学習した認識辞書との類似度演算を逐次行い、入力中の認識対象単語を検出を行う認識する方式である(図1)。この方式により高雑音下でも高性能な音声認識が可能となるが、高速なWSを使用しても認識は実時間の約90倍の処理時間がかかる。

3. 認識処理の分割

認識処理の約90%は入力と認識辞書との積和を基本とする類似度演算であり、この膨大な処理をワークステーション(WS)上でリアルタイムに行えるアクセラレータはなかった。そこで、既存のDSPを複数使用したアクセラレータを開発することにした。認識処理を個々のDSPに分割する必要があるが、分割の方向として処理の流れ、認識語彙などが考えられる(図2)。主要な処理は類似度演算であり、分析専用のDSPは設けずに各DSPを同一構成とした。また、認識語彙を分割してDSPに割り当て、それぞれの認識辞書のみを持つようにした。これにより、共通メモリを介して同一構成のDSPが並列に接続され柔軟な構成となり、認識処理の変更などが容易となる。

4. アクセラレータの構成

アクセラレータのハードウェア構成を図3に示す。DSP(TMS320C30)を4個使用し、各DSPに256kbyteのローカルメモリと32kbyteのプログラムメモリを用意し、WSや各DSP間のデータの転送は512kbyteの共有メモリを使う。ホストと共有メモリ間のデータ転送のためにDMAを使用する。バスインターフェイスはVMEを採用した。

共有メモリのバス アービトレーションはDMA、ホスト、DSPについて行い、バスのリクエスト、リリースはDSPからはファームウエアのインターロック オペレーション、ホストからはコマンド、DMAはDMA自身が行なう。

共有メモリ、DSPのローカルメモリはホストのメモリ空間にマップでき、各DSPのファームウエアはホストからロードする。割込みはホストから

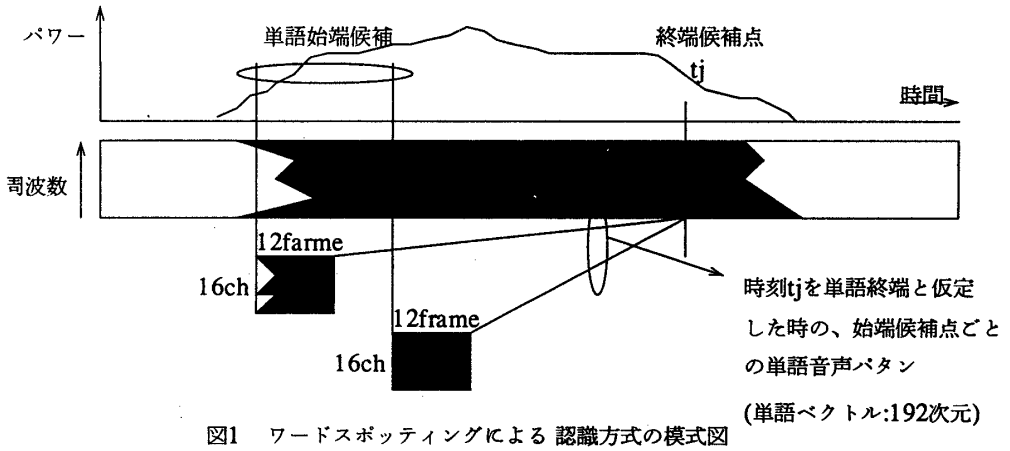

図1　ワードスポッティングによる認識方式の模式図

Development of an Accelerator for Speech Recognition in Noisy Environments
Hiroyuki Tsuboi,Hiroshi Kanazawa, Yoichi Takebayashi
R & D Center, Toshiba Corp.

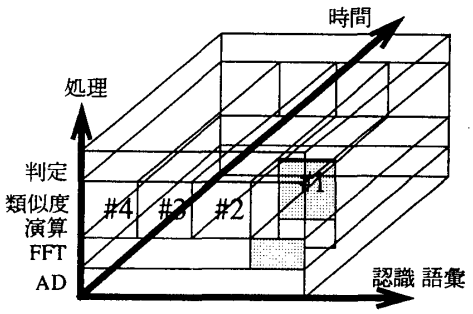

図2 認識処理の分割

DSPへ、DSP、DMAからホストへ可能である。

ファームウエア開発は既存のCコンパイラとアセンブラを利用する。

5. 高耐雑音音声認識による性能評価

5.1 システム構成

評価に使用したシステムの全体構成を図4に示す。ホストはAS4260、バスにA/D,D/A変換器、アクセラレータを接続した。システムの制御はAS4260が行う。

5.2 認識処理の流れ

認識処理の流れは以下のようになる。(図2)

(1) 入力音声をA/D変換し,DMAを用いて共有メモリにA/Dデータ転送する。

(2) DSP(#1)で入力音声をFFTで周波数分析し、共有メモリに転送する。

(3) 共有メモリから、DSP(#2-#4)のローカルメモリに分析結果を転送し、複数の入力候補と各カテゴリの標準パタンのそれぞれの複合類似度演算を行う。結果を共有メモリを通してDMAでホストに転送する。

(4) ホストで類似度に基づき判定処理を行ない、認識結果を出力する。

以上の各処理をパイプライン的に行なう。認識結果を得るまでの遅れは発声者が遅いと感じない程度のものでなければならないので、各処理は80msecごとに行い、認識結果は入力から240msec後に得る。

5.3 処理時間の比較

AS4260をホストとして、アクセラレータを使用した場合としない場合の時間の比較を行った。実験条件を表1に、結果を表2に示す。使用すると2ケタ以上高速化でき、リアルタイムでのワードスポッティングによる認識が可能となった。

6. あとがき

共有メモリを介してDSPを並列に構成したアクセラレータを開発した。これによりワークステーション上で高耐雑音の音声をリアルタイムで認識することが可能となった。本アクセラレータは柔軟な処理を実現できるので、様々な音声研究に利用できる。

今後は、本システムを用いて音韻ベースの認識システム[5]を構築する予定である。

参考文献

[1] S.Chatterjee,Proc.ICASSP-89,pp774-777

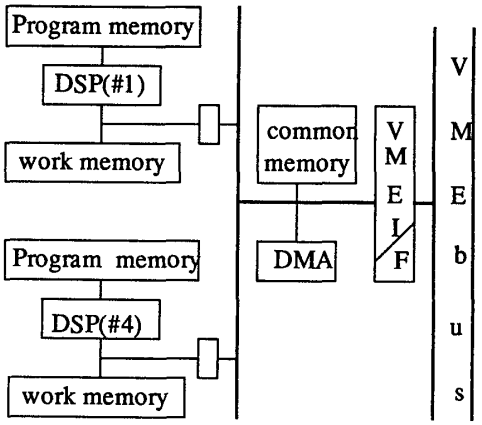

図3 アクセラレータの構成

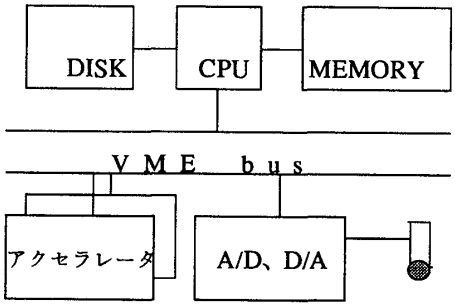

図4 音声認識システム構成

表1 実験条件

フレーム周期	8ms
FFT ポイント数	512 点
次元数	192次元 (16ch x 12frame)
カテゴリ数	13カテゴリ 10数字＋3コマンド
DSP クロック	33 MHz

表2 演算時間の比較
(2ボード,8DSP,13カテゴリ,1フレーム当たり)

システム	演算時間
AS4260 のみ	876ms
AS4260+アクセラレータ	5.5ms

[2] R.Bisiani,Proc.ICASSP89,pp782-784
[3] 金沢、坪井、竹林:音響学会講演論文集, 2-1-12,1989.10
[4] 坪井、竹林:音響学会講演論文集,3-1-6, 1989.10