LG-003

携帯電話用プロセッサで動作する大語彙連続音声認識の並列処理

Parallelization of an LVCSR algorithm for cellphone-oriented processors

石川 晋也† 山端 潔† 磯谷 亮輔† 奥村 明俊†
Shin-ya ISHIKAWA Kiyoshi YAMABANA Ryosuke ISOTANI Akitoshi OKUMURA

1. まえがき

近年、携帯電話や携帯機器は日増しに複雑化しており、例えばキーワードを入力することにより望む機能を検索する機能があれば便利だと思われる。しかし限られたキーによる入力は煩雑であり、代わりに音声認識を用いることが考えられる。同様に e-メールも良く利用される機能であるが、メール文のキーによる入力は煩雑であり、音声認識による代替が期待される。これらは単語発声に制約されない大語彙連続音声認識によって実現されうるが、大語彙連続音声認識は電池容量の制約や携帯電話向けプロセッサの処理能力の制限により携帯電話に搭載されてこなかった。最近、複数のプロセッサコアを搭載し、省電力・高性能を両立させた携帯電話向けプロセッサが発表されている[1]。我々は、3 つのプロセッサコアを使用して大語彙連続音声認識の実時間動作を実現する並列処理を検討し、認識率及び速度評価を行ったので報告する。また実際に携帯電話のマニュアルを自由文発声で検索するシステムを試作した。

2. ベースの大語彙連続音声認識システム

ベースシステムの全体構成

並列化のベースとした音声認識システムは、携帯機器向けに構築された省メモリ・省演算量の大語彙連続音声認識システムである[2]。デコーダは木構造辞書を用いた 1 パスのフレーム同期ビームサーチで、木構造辞書は音素表現で 1 音素 1byte で表現されている。各フレームで単語終端に到達した仮説を単語終端テーブルに書き出し、最終的に発声終端においてワードグラフとして結果を出力する。木構造辞書途中に位置する仮説については近似値として単語 unigram 確率を factoring 値として付与している。発声最後でワードグラフをバックトラックして一位認識結果を得る。ワードグラフを用いた 2 パス処理を追加することも可能である。なお、単語終端テーブルは一定フレーム間隔毎にガベージコレクションすることにより、発声長に依存して増加するメモリ使用量を抑えている。

音響モデル

音響モデルとして triphone の混合ガウス分布 HMM を使用している。状態ごとにガウス分布の木構造を作成し、MDL 基準を最小にする分布集合を選択することで、性能劣化を抑えて効率的に混合数を削減している[3]。また、ガウス分布の分散を全ガウス分布にわたって共通化することにより、HMM のメモリサイズを約半分に低減し、さらにガウス分布確率値計算の演算量も低減している。加えて、ガウス分布を類似度に基づいて木構造化し、認識時に尤度の高い分布についてのみ精密な計算を行う[4]ことで確率値計算量を大幅に削減している。

†NEC メディア情報研究所

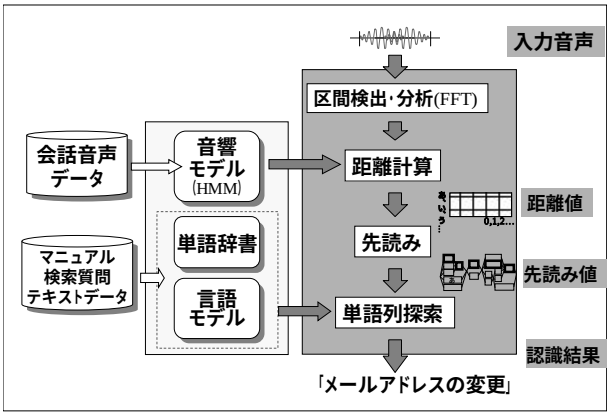


図1： 先読みを導入した大語彙連続音声認識

言語モデル

言語モデルは対象となるタスクにおける数千～数万文規模のコーパスから学習された bigram および品詞をベースにしたクラス bigram から構成される。また確率値は量子化されてコンパクトに格納される。

3. 認識処理の並列化

3.1 先読み処理の導入

処理全体において支配的時間をとる単語列探索は、複雑なグラフ処理であるため簡単に並列処理に分割できない。そこでこの単語列探索処理を、言語モデルを用いない簡単な音響レベルの前処理(先読み処理)とそれを用いて仮説数制限を行う単語列探索処理へと分割・置換する[5]。先読み処理では一定区間の音声を用いて発声とは逆方向に予測音響スコア(先読み値)を作成する。単語列探索ではそれを利用して仮説数を効率的に削減する。この構成により大語彙連続音声認識が分析+距離計算、先読み処理、単語列探索の３段処理で実現される(図１)。

3.2 ３段処理の並列化

前述した３段処理をそれぞれ一つの CPU でフレーム同期処理させることができれば、1 フレームずつの遅れで 3 処理を同時に実行することができる。しかし先読み処理が時間方向逆向きの処理であるため、これは不可能である。

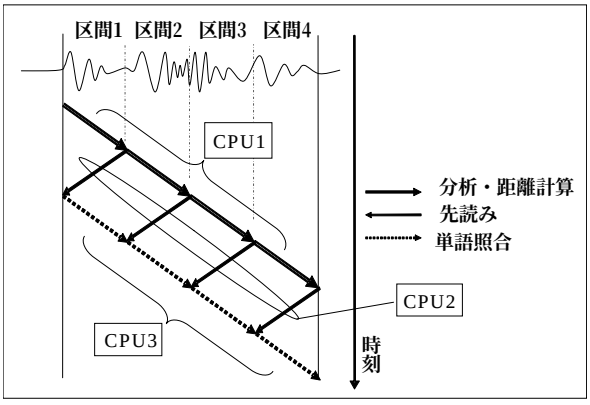


図2： 3段処理の区間単位の並列化

そこで、図２に示すように発声を一定間隔の区間に分け、区間遅れで３処理を並列処理することにした。区間中は３処理とも独立して処理を行い、区間の境界で分析+距離計算処理は先読み処理にその区間の距離値を、先読み処理は単語列探索処理にその区間の距離値と先読み値を、それぞれ渡す。単語列探索処理は発声全体の先読み値を参照しながら区間継ぎ目に関係なくフレーム同期に単語列をサーチし、発声全体に対する認識結果を生成する。3CPU 間での上記データのやり取りは 3CPU で共有しているメモリ上に値を書き出し、そのメモリエリアを発声区間の境界で受け渡すことで行う。
なお、メモリ上の固定リソースである音響モデルと、言語モデル及び対応する辞書は、それぞれ分析+距離計算処理、単語列探索処理のみで使用される。このため、これらはそれぞれ特定の CPU からのみ参照できればよく、重複して複数の CPU のメモリ空間を占有することはない。

4. 認識速度と精度の評価

ここでは上述した方式を実際に実装して行った大語彙連続音声認識の認識速度と精度の評価について説明する。

評価条件

実装は ARM9 コアを３つもつプロセッサに対して行った。動作周波数は 150MHz に設定した。認識速度評価用発声は通訳システム[6]で対象とした旅行会話の日本語読み上げ文発声 50 発声である。認識精度評価用発声もまた、男性 5 名による合計 1000 発声の通訳旅行会話の読み上げ文発声である。ともに音響モデルは男性話者不特定モデルである。辞書、言語モデルは通訳システム[6]のものを用いており、認識単語数は約 50000 語である。評価用音声は wave ファイルでシステムに格納され、マイク入力と同様にタイマ処理でタイマ間隔分だけ取り込まれて認識処理される。なおメモリに関してはシステム各部が固定量を確保する場合が多く、後述するマニュアル検索システムについてこれらを合計した数値で比較を行った。

表１

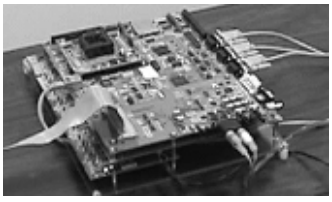
	従来システム (1CPU)	並列システム (3CPU)
速度(傾き)	2.6RT	1.0RT
PC, WA	96.0%, 95.8%	95.6%, 95.4%
使用メモリ	3.5M バイト	5.5M バイト

結果

表１に速度、単語正解率(PC)、単語正解精度(WA)、使用メモリを示す。PC、WA は先読みが導入されたことでともに従来の 1CPU システムよりわずかに落ちるが、ほぼ同等である。速度(傾き)は評価発声の長さの増加に対するシステムの処理時間増加の割合を示したものである。これは、図 1 に示すような発声終了後に行われる認識処理を除いた速度で、本並列システムでは 1 リアルタイム(RT)を実現できていることが分かる。これは今回の測定法では上限の速度であり、まだ CPU パワーに余裕がある可能性がある。メモリ使用量を比較すると、本並列システムは従来の 1CPU システムに比べて 2M バイトほど増加している。これらは 3CPU で独立して区間処理を行うために重複して持つ先読み値、距離値エリア分にほぼ一致する。

5. マニュアル検索システムの試作

上述した 3CPU による大語彙連続音声認識を用いて、音声で入力した質問文に対応する携帯電話マニュアルのページを検索するシステムを試作ボード上に構築した。ここでは、大語彙連続音声認識だけでなく、検索部もボード上で動作させている。なお言語モデルや検索部の使用するモデルは基本的に既存の電話認識システム[7][8]と同一であるが、本システム用にコンパクトに再構築されている。このシステムでは全てをボード上で動作させることで従来の電話認識システムと比べ快適なレスポンスが実現された。



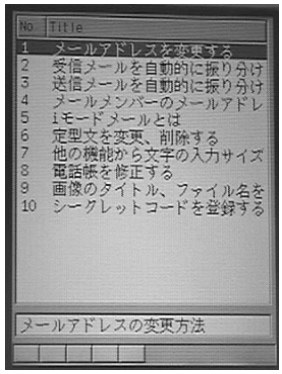


図３
上：評価ボード
右：質問発声に対する
10 位結果表示の様子

6. 結論

プロセッサコアを複数持つ携帯電話用プロセッサ向けに大語彙連続音声認識処理を並列化して評価ボード上に実装し、認識単語数約 50000 の大語彙連続音声認識が 1RT で動作することを確認した。またこれを用いて、従来電話認識システムで実現されていた携帯電話マニュアルの音声検索システムを試作し、将来の携帯電話でマニュアル等の文書の自由発話文による音声検索が動作する可能性を示した。

7. 謝辞

並列化実装にご協力・ご助言頂きました NEC システムデバイス研究所の枝廣主席研究員，酒井主任に感謝致します。

8. 参考文献

[1]鳥居淳他, ”A 600MIPS 120mW 70μA Leakage Triple-CPU Mobile Application Processor Chip”, p136, ISSCC, 2005 年
[2]石川晋也他, “コンパクトなディクテーションの開発”, 音講論集, 3-5-12, 2002 年 3 月
[3] 篠田浩一他, “Efficient reduction of gaussian components using MDL criterion for speech recognition" , 信学技報 , SP-83, p.69, 2001 年
[4]渡辺隆夫他, ”High speed speech recognition using tree-structured probability density function", p.556, ICASSP, 1995 年
[5]堀貴明他, ”大語彙連続音声認識のための音素グラフに基づく仮説制限法の検討”,情処論文誌,Vol.40,No.4,1999 年
[6]山端潔他, “PDA で動作する旅行会話向け日英双方向音声翻訳システム”, 情処研報, 2002-NL-150-9
[7]石川晋也他, “Speech-activated text retrieval system for multi-modal cellular phones”, SP-P4.12, ICASSP, 2004 年
[8]安達史博他, “携帯電話向け音声/WEB 連動型検索システム”, 音講論集, 1-8-20, 2004 年 3 月