論文 / 著書情報
Article / Book Information

論題(和文)	超並列デコーダを用いた話し言葉音声認識
Title(English)	
著者(和文)	篠崎 隆宏, 古井 貞熙
Authors(English)	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	日本音響学会 2004年春季講演論文集, Vol. , No. 2-11-6, pp. 111-112
Citation(English)	, Vol. , No. 2-11-6, pp. 111-112
発行日 / Pub. date	2004, 3

2—11—6 超並列デコーダを用いた話し言葉音声認識*

◎篠崎 隆宏　古井 貞熙　(東工大)

1 はじめに

多様に変化する話し言葉音声に対応することを目的として、種々の特徴を持った音声モデルをもとに多数のデコーダを並列計算機上で実行する超並列デコーダの提案を行う。各発話に対してたまたま高い適合度をもつモデルを、音響尤度および言語尤度両方を用いて選択することにより認識率の向上を図る。並列計算機を用いることで、認識処理に必要な時間は単一のモデルを用いた従来のデコーダと比較して僅かに増えるのみである。モデルの並列化は音響モデルまたは言語モデル、あるいは両方の組み合わせに対して行うことが出来る。また、教師なし話者適応化と組み合わせることも可能である。約 400 のデコーダを並列実行した結果について報告する。

2 超並列デコーダ

超並列デコーダは、多数のデコーディングユニットおよび個々のデコーディングユニットの結果を統合する統合器から構成される。各デコーディングユニットはそれぞれ異る音声モデルを用いた通常のデコーダで、全体としてあらゆる音声に対応できるように構成するのが望ましい。図 1 に超並列デコーダのアーキテクチャを示す。

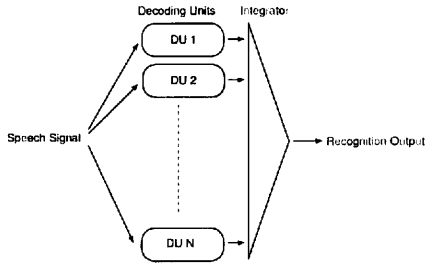

図 1. Architecture of the Massively Parallel Decoder.

超並列デコーダにおいて 1 発話を処理するために必要な計算量 Q は、デコーディングユニットの計算量を q 、並列数を N 、統合器の計算量を α とすると、式 (1) のように表される。

$$Q = q \times N + \alpha \quad (1)$$
$$\approx q \times N \quad (2)$$

デコーディングユニットの計算量は通常のデコーダと全く同じである。統合器の計算量 α はデコーディングユニットの計算量と比較して小さく、無視できる程度である。すなわち、式 (2) に示すように通常のデコーダと比較しておよそ N 倍の計算が必要となる。しかし処理時間 T (認識処理を開始してから、認識結果が得られるまでの時間) に関しては、並列計算機を用い各デコーディングユニットを別個のプロセッサに割り当てることで、式 (3), (4) に示すようにデコーディングユニットが必要とする時間 t とほぼ同じ時間とすることが出来る。ここで β は統合器が必要とする計算時間で、t と比較して無視できる。

$$T = t + \beta \quad (3)$$
$$\approx t \quad (4)$$

現状では数百のプロセッサを搭載した並列計算機は高価であり、広い場所と大きな電力を必要とする。しかし半導体技術の動向として、プロセスルールの微細化に伴う動作周波数の向上と配線抵抗の増大から配線遅延が大きな問題となりつつあり、今後は複数のプロセシングユニットがワンチップ上に実装される並列アーキテクチャに移行すると予想される [1]。将来的には GRID [2] のようなシステムがパッケージ上に構築されると考えられる。超並列デコーダはデコーディングユニット間の結合度が低く、このような並列アーキテクチャにおいて各プロセシングユニットを効率的に活用できる利点がある。

3 実験条件

3.1 タスクおよび学習セット

認識タスクは男性話者による学会 10 講演からなる、CSJ テストセット 1 である。実験では各講演毎に約 5 分間の発話をランダム抽出したサブセットを用いた。言語モデルの学習セットは CSJ の学会/模擬講演を含む 2485 講演の書き起こし、約 6.1M 形態素である。音響モデルの学習セットは CSJ の男性話者による学会講演約 186 時間である。

3.2 ベースラインシステム

音響モデルとして 3k 状態 16 混合の Tri-phone モデルを HTK を用いて作成した。HMM には MLLR 適応用に、64 の葉を持つ回帰木を付加した。言語モデルとして 30k 語彙の Tri-gram モデルを用いた。デコーダとして Julius を用いた。音響モデル/言語モデルとも不特定話者モデルを用いた実験、および不特定話者モデルによる認識結果を用いた MLLR による教師なし話者適応化音響モデルを用いた実験を行った。以下では不特定話者モデルを用いた認識システムを BASE、教師なし話者適応モデルを用いたシステムを BASE(MLLR) とする。

3.3 超並列システム

超並列デコーダでは様々な音声をカバーするような音響モデル/言語モデルのセットを用いる。今回は学習セット中の男性による 402 学会講演にそれぞれ適応化させたモデルを作成し、モデルセットとした。音響モデルセットはベースラインシステムで使用する HMM を各講演に MAP 適応させることで作成した。言語モデルセットは各講演のテキストを講演数で全体の 5%となるように繰り返し学習セットに加えた後に、Tri-gram を学習することで作成した。なお、バックオフスムージングには Witten-Bell 法を用いた。

実験はモデルセット中の各モデルを用いて Julius により認識処理を行い、発話毎に最尤となる認識仮

* Spontaneous speech recognition using Massively Parallel Decoder
By Takahiro Shinozaki and Sadaoki Furui (Tokyo Institute of Technology)

説を選択することにより行った。具体的には以下の4通りの認識実験を行った。

(1) 音響モデルのみを多重化し、言語モデルには不特定話者モデルを用いる。このシステムを mAM とする。

(2) 音響モデルに不特定話者モデルを用い、言語モデルを多重化する。このシステムを mLM とする。

(3) 実験 (1) の結果を用い、多重化した各音響モデルを初期モデルにして、テストセットの各講演に対して MLLR により教師なし適応する。言語モデルには不特定話者モデルを用いる。このシステムを mAM(MLLR) とする。

(4) 実験 (3) の結果を用い、不特定話者モデルを元に教師なし適応した音響モデルをテスト話者毎に1つ作成する。言語モデルのみを多重化して用いる。このシステムを mAM(MLLR)+mLM とする。

4 実験結果

音響モデル/言語モデルに単一の不特定話者モデルを用いたベースラインデコーダ BASE、音響モデルを多重化した超並列デコーダ mAM、および言語モデルを多重化した超並列デコーダ mLM の単語正解精度を図 2 に示す。ベースライン認識率 71.8% と比較して、超並列デコーダ mAM で 1.9%、mLM で 2.4% 認識率が向上した。

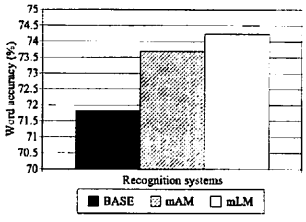

図 2. Word accuracy using the Massively Parallel Decoder.

次に音響モデルの教師なし話者適応を行った場合の実験結果として、教師なし適応化音響モデルを用いたベースラインデコーダ BASE(MLLR) 、多重化した教師なし適応化音響モデルを用いた超並列デコーダ mAM(MLLR)、さらに言語モデルの多重化を組み合わせた超並列デコーダ mAM(MLLR)+mLM の単語正解精度を図 3 に示す。ベースライン認識率 76.9% と比較して mAM(MLLR) で 1.6%、mAM(MLLR)+mLM で 2.5% 認識率が向上した。

5 考察

超並列デコーダ mAM において、統合器における認識仮説選択の基準として音響尤度と言語尤度の和 (AML+LML) の他に、仮説の音響尤度 (AML)、言語尤度 (LML) を用いた場合の単語正解精度を表 1 に示す。言語尤度は言語重みや挿入ペナルティを含めた値を用いている。また認識結果の尤度を用いる代わりに、HMM セットの作成に用いた 402 学会講演それぞれに対して学習した GMM の尤度 (GMML) を用いて選択を行った場合の認識率も合わせて示す。GMM を用いる場合、認識処理は選択されたモデル

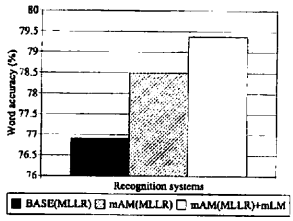

図 3. Word accuracy using the Massively Parallel Decoder with unsupervised acoustic model adaptation.

のみを使用して行えばよいので、計算量を大幅に削減できる利点がある。しかし表より GMM を用いた場合は、認識結果の尤度を用いる場合と比較して認識率が悪いことが分かる。また、認識結果の尤度を用いる場合、音響尤度または言語尤度の何方か片方を用いるよりも、両方の和を用いた方が高い認識率が得られることが分かる。

表 1. Word accuracy vs. hypothesis selection criterion

	AML+LML	AML	LML	GMML
ACC(%)	73.69	71.4	72.6	71.33

超並列デコーダ mAM における、デコーディングユニット数と単語正解精度の関係を図 4 に示す。認識率の向上がユニット数の対数にほぼ比例することが分かる。

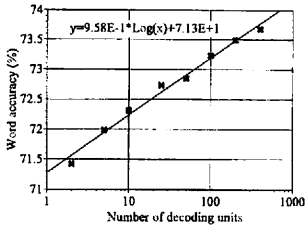

図 4. Relationship between number of decoding units and word accuracy.

6 まとめ

音声の性質が様々な要因により影響を受ける話し言葉音声に対応するために、それぞれ異った特性を持つ多数のデコーディングユニットを並列に用いた上で結果を統合する、超並列デコーダの提案を行った。本稿ではモデルセットとして多数の (およそ 400) 話者適応化モデルを用いたが、おそらく人においてもこの程度の数の話者に対応したモデルは持っていると思われる。提案手法は単独で用いた場合および MLLR による教師なし話者適応と組み合わせた場合どちらにおいても、認識率の向上に有効であることを示した。

参考文献

[1] L. Hammond et.al., "A Single-Chip Multiprocessor," Computer, Vol.30, No.9, pp. 79–85, 1997.

[2] 伊藤智, "グリッドコンピューティングの技術動向," 情報処理, Vol.44, No.6, pp. 576–580, 2003.