論文 / 著書情報
Article / Book Information

論題(和文)	超並列デコーダによる話し言葉音声認識
Title	
著者(和文)	篠崎 隆宏, 古井 貞熙
Author	Takahiro Shinozaki, SADAOKI FURUI
出典(和文)	第3回話し言葉の科学と工学ワークショップ 講演予稿集, Vol. , No. , pp. 67-72
Journal/Book name	, Vol. , No. , pp. 67-72
発行日 / Issue date	2004, 2

超並列デコーダによる話し言葉音声認識

篠崎 隆宏† 古井 貞熙†

† 東京工業大学 大学院情報理工学研究科 計算工学専攻
〒 152-8552 東京都目黒区大岡山 2-12-1

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

あらまし 話し言葉音声は話者間および話者内において多様に変化するため、単一の不特定話者モデルを用いて認識を行おうとすると、低い認識率となってしまう。そこで、どのような音声に対しても頑健かつ高い精度で認識を行うために、様々な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、超並列デコーダを提案する。各発話に対して最も高い適合度をもつモデルを用いた認識結果を選択することで、認識率の向上を図る。超並列計算機を用いることにより、認識処理に必要な時間は従来のデコーダとほぼ同程度とすることが出来る。日本語話し言葉コーパスの講演音声を用い、およそ400のデコーダを並列実行する認識実験を行った。提案手法は単独で用いることも、音響モデルの教師なし適応と組み合わせて用いることも出来、どちらの場合においても認識率の向上に有効である。教師なし適応と組み合わせて用いた場合には、CSJテストセットの10講演に対し、平均で80%近い認識率が得られた。

キーワード 話し言葉音声認識, 音響モデル, 言語モデル, 超並列計算機, マルチプロセッサ, 日本語話し言葉コーパス

Spontaneous speech recognition using a Massively Parallel Decoder

Takahiro SHINOZAKI† and Sadaoki FURUI†

† Department of Computer Science, Graduate School of Information Science and Engineering
Tokyo Institute of Technology
Ookayama 2-12-1, Meguro-ku, Tokyo, 152-8552 Japan

E-mail: †{staka,furui}@furui.cs.titech.ac.jp

Abstract One of the main difficulties of spontaneous speech recognition comes from diversity nature of spontaneous speech. Acoustic and language models made by simply mixing many speakers' utterances do not work well since many utterance dependent characteristics are lost. One solution is to prepare a large set of models including a suitable model for every input utterance. We propose Massively Parallel Decoder that consists of a large number of decoding units and an integrator. Each decoding unit uses one of the models in the collection. By using a massively parallel computer, the increase of the recognition time is small compared to conventional decoders using single model and processor. Recognition experiments are conducted using the Massively Parallel Decoder with around 400 decoding units. The proposed Massively Parallel Decoder works by itself or in combination with unsupervised acoustic model adaptation. When combined with the MLLR unsupervised adaptation, averaged word accuracy of almost 80% was obtained for the CSJ test-set lectures.

Key words spontaneous speech recognition, acoustic model, language model, massively parallel computer, multi-processor, Corpus of Spontaneous Japanese

1. はじめに

従来、認識システムで使用される音響モデルや言語モデルは読み上げ音声を用いて作成され、ニュースのアナウンサーなどに対しては90%以上の高い認識率が達成されたものの、一般の話し言葉を対象とすると認識率が著しく低下してしまう問題があった。これは読み上げ音声と話し言葉音声では、音声の性質が大きく異るためである。近年大規模な『日本語話し言葉コーパス』(Corpus of Spontaneous Japanese : CSJ) が構築され、実際の話し言葉データを用いてモデルを学習することで、日本語話し言葉に対する音声認識性能が大きく向上した。しかし不特定話者モデルを用いた認識率は70%程度と [1]、依然として多くのアプリケーションにとって不充分な状況である。

大量の話し言葉データを用いて認識に使用するモデルを学習しているにもかかわらず読み上げ音声と比較して認識率が低い理由としては、話し言葉音声の多様性が挙げられる。これまでのCSJを用いた研究においても、話し言葉音声では話者間で発話スタイルが大きく異ることに加え [2]、同じ話者内でも様々な要因により音声の性質が多様に変化することが示されている [3]。このような多様な音声を1つに集めて作成した音響モデルや言語モデルでは、個々の音声の特徴が平均化されることでぼやけてしまい、高い認識性能が得られないと考えられる。

多様に変化する話し言葉音声に対して高い認識性能を得るためには、ただ1つの不特定話者モデルをもとに認識を行うのではなく、種々の特徴を持った多数のモデルを用いることが有効と考えられる。そこで、様々な音声モデルをもとに多数のデコーダを超並列計算機上で実行する、超並列デコーダの提案を行う。各発話に対して最も高い適合度をもつモデルを用いた認識結果を選択することにより、認識率の向上を図る。

提案手法では多種類のモデルを用いて認識処理を行う必要があるが、超並列計算機を用いることで、認識処理に必要となる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 (DU) およびそれら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。入力された音声は全てのデコーディングユニットに配られ、各デコーディングユニットはそれぞれの音声モデルを基に同じ音声を並列処理する。各デコーディングユニットからの認識結果は統合器において統合され、最終的な認識結果が出力される。

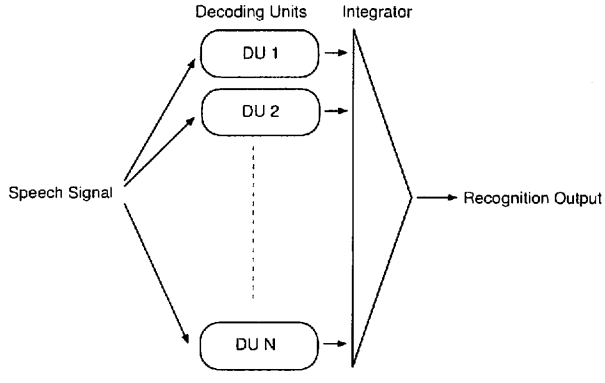

図1 Architecture of the Massively Parallel Decoder.

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、図2に示すように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シングユニット (PU) 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。

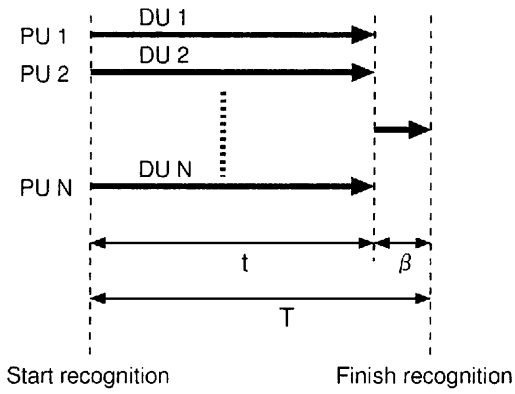

図 2 Processing time of the Massively Parallel Decoder.

$$T = t + \beta \tag{3}$$

$$\approx t \tag{4}$$

現状では数百のプロセッサを搭載した超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トランジスタの動作周波数の向上や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 [4]。また、将来的にはGRID [5] のようなシステムもパッケージ上に構築されるようになると考えられる [6]。超並列デコーダはデコーディングユニット間の依存度が低く、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。発話単位は CSJ の書き起こしに含まれる時間情報を基に、およそ 500ms 以上の無音区間を基準として切り出して用いた。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。各講演 5 分間の中に含まれる発話数を表 1 に示す。言語モデルの学習セットとして、CSJ の学会/模擬講演を含む 2485 講演の書き起こし約 6.1M 形態素を用いた。音響モデルの学習セットとして、CSJ の男性話者による学会講演約 186 時間を用いた。

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 [7] 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。

表 1 Test-set

Conference name	# of utterances used
A01M0097	58
A04M0051	77
A04M0121	73
A03M0156	88
A03M0112	43
A01M0110	65
A05M0011	31
A03M0106	27
A01M0137	45
A04M0123	23

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 を各講演に教師あり 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、発話毎に認識仮説を選択することにより行った。具体的には図 3 に示す手順で、以下の 4 通りの認識実験を行った。

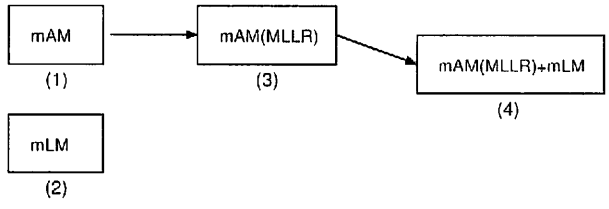

図 3 Dependency of the recognition experiments.

（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) の認識結果を用い、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) の認識結果を用い、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。

3.4 選択基準

デコーディングユニット群から出力される認識結果を統合する方法としては、認識結果の尤度や信頼度を用いることが考えれる。認識結果の尤度を用いる尤度法として、音響尤度と言語尤度の和を最大とする仮説を選択する方法を用いた。

認識結果の信頼度を用いる信頼度法として、式 (5) に示すように認識仮説中の単語信頼度を平均した値を用いた選択を行った。

$$C(Hyp_i) = \frac{\sum_{w=1}^{W_i} c_i(w)}{W_i} \qquad (5)$$

ここで $c_i(w)$ は i 番目のデコーディングユニットの認識仮説において、文頭から w 番目の単語の信頼度である。単語信頼度は、各デコーディングユニット内でのトレリス上の単語集合から、その単語の事後確率を求めることにより行った [8]。これは、認識仮説の認識率の推定を行い、推定値を最大とするように認識仮説の選択を行う基準である。

また、式 (6) に示すように、各仮説毎にその他の仮説への距離の和 $SD(Hyp)$ を計算し、$SD(Hyp)$ を最小とする仮説を選択する距離和法を試みた。

$$SD(Hyp_i) = \sum_{j=0}^{N} dist(Hyp_i, Hyp_j) \qquad (6)$$

ここで $dist(Hyp_i, Hyp_j)$ は 2 つの仮説間の距離で、仮説 Hyp_i を基準として Hyp_j との相違を DP マッチングにより求めたものである。置換、削除および挿入による相違のコストをそれぞれ 1 とし、足し合わせた値を用いた。$dist(Hyp_i, Hyp_j) = dist(Hyp_j, Hyp_i)$ および、$dist(Hyp_i, Hyp_i) = 0$ が成り立つ。$SD(Hyp)$ はデコーディングユニット群が出力する仮説の分布中で中心的な仮説に対して小さな値となり、他のどの仮説からも相違の大きい仮説に対して大きな値となる。すなわち、認識仮説の分布中で一番平均的な仮説を選択する基準である。

4. 実験結果

4.1 尤度基準選択による認識率

超並列デコーダのデコーディングユニット群から得られる認識仮説を、音響尤度と言語尤度の和に基づき選択する尤度法を用いた実験を行った。使用した言語尤度には言語重みや挿入ペナルティも含まれている。

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 4 に示す。単語正解精度は 10 講演の結果を平均した値である。ベースライン認識率 71.8% に対して、超並列デコーダ mAM では 73.7% と、認識率が 1.9% 向上した。超並列デコーダ mLM では認識率が 2.4% 向上し、74.2% となった。

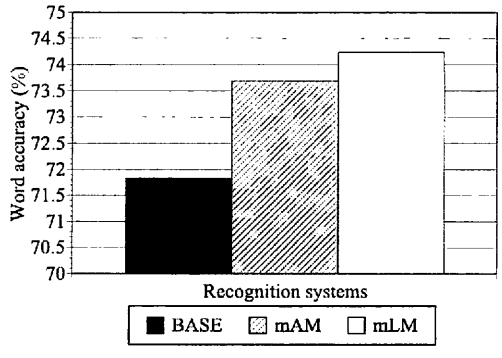

図 4 Word accuracy using the Massively Parallel Decoder.

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 5 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、認識率はそれぞれ 78.5%と 79.4%となった。なお、mAM(MLLR) と mAM(MLLR)+mLM はどちらも教師なし話者適応音響モデルを用いているが、mAM(MLLR) では多重化したモデル群を用いているのに対し、mAM(MLLR)+mLM では mAM(MLLR) の結果を基に単一の音響モデルを用いており、言語モデル多重化の効果の観点からはこれらの認

識率を単純に比較することは出来ない。しかし、mAM(MLLR)+mLM において使用した音響モデルと単一の不特定話者言語モデルを組み合わせた実験における認識率は mAM(MLLR) と比較してほぼ同じで向上は見られなかったことから、mAM(MLLR) と mAM(MLLR)+mLM の認識率の差はほぼ言語モデルを多重化した効果といえる。

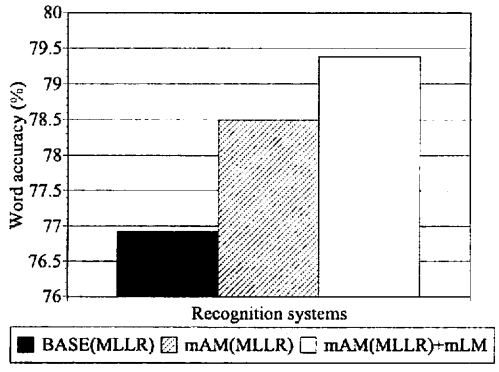

図 5 Word accuracy using the Massively Parallel Decoder with unsupervised acoustic model adaptation.

ベースライン認識器 BASE および、超並列デコーダ mAM(MLLR)+mLM を用いた場合の、講演毎の認識率を図 6 に示す。mAM(MLLR)+mLM を用いた実験では、ベースラインと比較して認識率が大きく向上し、ほとんどの講演で 75%以上の単語正解精度が得られた。しかし、講演間のばらつきが大きく、90%以上と高い認識率を示す講演がある一方で、70%未満の認識率となる講演も存在することが分かる。

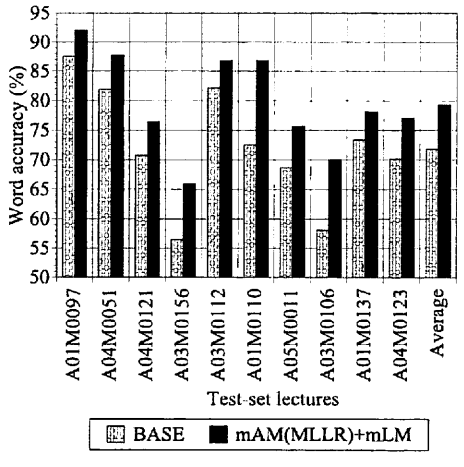

図 6 Word accuracy of the lectures using mAM(MLLR)+mLM.

4.2 選択基準と認識率

音響モデルを多重化した超並列デコーダ 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) および言語尤度 (LML) をそれぞれ単独で用いた場合の単語正解精度を表 2 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 2 Word accuracy vs. hypothesis selection criterion

	AML+LML	AML	LML	GMML
ACC(%)	73.7	71.4	72.6	71.3

また、超並列デコーダ mAM(MLLR)+mLM の仮説選択において、信頼度および距離和を用いた実験を行った。信頼度をもちいた実験では式 (7) に示すように、信頼度と認識仮説の尤度を線形補間した値を用いて選択を行った。

$$score = (AML + LML) + \alpha CM \qquad (7)$$

ここで $score$ は仮説選択に用いるスコア、$AML + LML$ は認識仮説の尤度、CM は信頼度である。α は線形補間の際の重みで、$\alpha = 0$ のときは尤度のみを用いた選択となる。図 7 に α を変化させたときの認識率を示す。α の値を適切に設定することで (およそ 50〜100)、僅かであるが尤度のみを用いる場合と比較して認識率が向上した。最適値を超えて α の値を大きくしていくと認識率は徐々に減少していき、信頼度のみにより選択を行った場合の認識率は 78.6% であった。

距離和を基準とした実験では式 (8) に示すように、認識仮説尤度と組み合わせた値を用いた。$\alpha = 0$ のとき、尤度のみを用いた選択となる。

$$score = (AML + LML) + \alpha \frac{1}{SD + 1} \qquad (8)$$

図 8 に α を変化させたときの認識率を示す。距離和を用いた場合は、α の値にかかわらず、尤度のみを用いた場合と比較して認識率の向上は見られなかった。

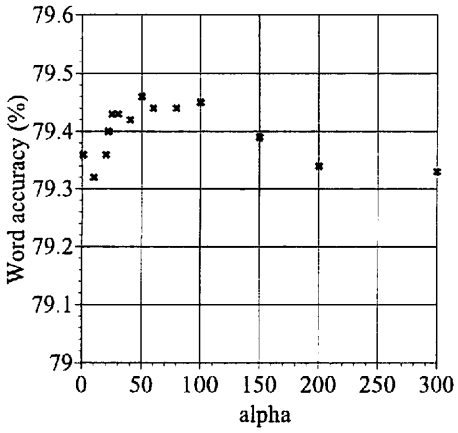

図 7 Word accuracy using confidence measure as a selection criterion.

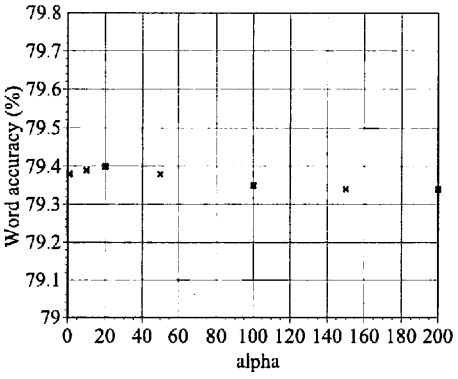

図 8 Word accuracy using distance as a selection criterion.

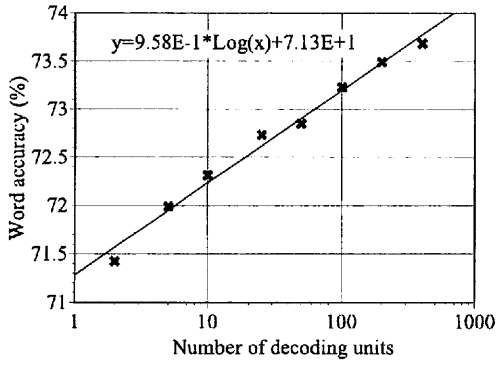

図 9 Relationship between number of decoding units and word accuracy.

5. 考　　察

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 9 に示す。認識率の向上がユニット数の対数にほぼ比例し、ユニット数が 10 倍になると認識率が 0.958% 増加することが分かる。ユニット数を増やすことで、さらに認識率が向上すると考えられる。

超並列デコーダではデコーディングユニットからの認識仮説の選択を行う。理想的な選択が行われた場合の認識率の上限を求めるために、認識率を用いて仮説文選択を行った。この場合のテストセットの認識率は、超並列デコーダ mAM において 82.5%、mLM において 80.7%、mAM(MLLR)+mLM において 83.5%であった。尤度などを用いた場合の認識率と比べて 4%から 9%ほど高く、仮説文の選択方法を改良することで認識率の向上が期待される。

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。また、超並列計算機を用いることで、従来のデコーダと比較した認識処理時間の増加は僅かに抑えることが出来る。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。提案手法と MLLR による教師なし話者適応化を組み合わせることで、10 講演の平均で 79.4%の認識率が得られた。

今後は、仮説選択方法の改良に加え、音響モデルや言語モデルのモデルセット作成法の改良、音響モデルおよび言語モデルの同時多重化などを行う予定である。

文　　献

[1] T. Kawahara et.al., "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp. 135–138, 2003.

[2] T. Shinozaki et.al., "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. ICASSP, Vol.1, pp. 729–732, 2002.

[3] K. Maekawa, "『日本語話し言葉コーパス』の構築," 話し言葉の科学と工学ワークショップ講演予稿集, pp. 7–12, 2001.

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

[5] 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.

[6] R. Tummala et.al., "System on Chip or System on Package?," IEEE Design & Test of Computers, Vol.16, Issue 2, pp 48–56, 1999.

[7] A. Lee, et al., "An efficient two-pass search algorithm using word trellis index," Proc. ICSLP, pp.1831–1834, 1998.

[8] 李 晃伸 他, "単語認識エンジン Julius における単語事後確率を用いた信頼度算出," 2003 年秋期音講論, 3–6–8, pp. 117–118, 2003.