社団法人　情報処理学会　研究報告
IPSJ SIG Technical Reports

2005－ＳＬＰ－59（16）
2005／12／22

超並列計算機を用いた入力音声の変動に頑健な音声対話システムの検討

中川　竜太†　　岩野　公司†　　古井　貞熙†

† 東京工業大学　〒152-8552 東京都目黒区大岡山 2-12-1-W8-81
E-mail: †{rtag,iwano,furui}@furui.cs.titech.ac.jp

あらまし　入力音声の変動のうち，事前に予測できる変動にはそれらに適したモデルを予め用意し，予測が困難な変動にはモデルを逐次適応化することで，音声認識の頑健性が向上する．計算量が莫大となるこれらの手法を組み合わせて，実時間での処理が要求される音声対話システムに適用するために，超並列計算機を用いることを検討する．本稿では，複数の音声認識器を同時並行に駆動して得られた複数の認識仮説を尤度を基準に選択し，適応処理をバックグラウンドで行うアーキテクチャを GRID 上に実装した．飲食店舗検索タスクにおいて，発話内容 (話題・発話カテゴリ) による入力音声の言語的変動を表す複数の言語モデルと，話者の違いによる音響的変動を表す複数の特定話者音響モデルを用いた．事前に収録された対話音声による評価実験を行ったところ，単一の音響モデルと言語モデルによる従来のシステムと比べ，75 台の音声認識ノードと 15 台の話者適応ノードを駆動することで，構築したシステムではキーワード認識誤り率を 25.5% 削減することができた．

キーワード　対話システム, 逐次話者適応, 超並列計算機

Spoken dialogue system robust against speech variations based on massively parallel computing

Ryuta NAKAGAWA†, Koji IWANO†, and Sadaoki FURUI†

† Tokyo Institute of Technology, Ookayama 2-12-1-W8-81, Meguro-ku, Tokyo, 152-8552 Japan
E-mail: †{rtag,iwano,furui}@furui.cs.titech.ac.jp

Abstract　Robustness of speech recognition increases by preparing models suitable to acoustic and linguistic variations when they can be predicted. This also increases by incrementally adapting the models when the variation is difficult to predict. In order to combine these methods which need huge amount of computation, and implement them in spoken dialogue systems which need real time processing, this paper investigates using a massively parallel computer. Architecture of selecting a recognition result having the maximum likelihood from the results obtained by multiple speech recognizers driven in parallel and running adaptation processes in the background has been implemented on a GRID computing system. In a restaurant information retrieval task, multiple language models representing linguistic variations of input speech according to utterance contents (topics/utterance categories) and multiple speaker-dependent acoustic models representing speaker variations have been used. Results of evaluation experiments using pre-recorded dialogue utterances show that the proposed system achieves 25.5% reduction in the keyword recognition error rate in comparison with a conventional system using a single acoustic as well as language model, when 75 recognition nodes and 15 speaker-adaptation nodes are driven.

Key words　dialogue system, incremental adaptation, massively parallel computer

1. まえがき

近年，様々な音声認識アプリケーションが実用化されているが，その中でも音声対話システムは，話者や背景雑音などの音響的特徴や，話題や発話スタイルなどの言語的特徴の変動により，認識性能が劣化するという問題を抱えている．つまり，認識対象音声の音響的・言語的特徴に適した音響・言語モデルを速やかに獲得することが課題となっている．

そのための有用な手法として，予測される変動に適したモデルを予め複数用意し，それらで認識した仮説の中から最適なものを選択する手法[1]や，変動が予測困難な入力音声に逐次適応化する手法[2]がある．

文献[1]では，変動を予測した複数の言語モデルと音響モデルを用意し，それらの組み合わせ数のデコーダを並列で駆動するという手法について論じており，この手法では，入力の変動が予測したものであれば最適なモデルでの認識が可能であるという特徴を持つ．タスクは学会講演音声の自動書き起こしである．言語モデルは，大規模コーパスをクラスタリングしてそれぞれのクラスタから作成している．また，音響モデルは，ベースとなるモデルを話者ごとに適応化することで，複数の話者依存モデルを作成している．複数のデコーダからの認識仮説は，尤度を基準に選択することで，書き起こし精度の向上を確認している．また，文献[1]では高精度の自動書き起こしを目的として，全評価データを用いて音響モデルと言語モデルを教師なしバッチ適応を行い，更なる精度向上を確認しているが，これは対話システムのような実時間での応答を求めるタスクには適用できない．

文献[2]で提案されている手法は，教師なしで逐次適応化した音響モデルを用いるもので，未知の変動にも対応できるという特徴を持つ．逐次適応化した音響モデルと適応化しない音響モデルを並列して持つが，認識に用いるモデルはGMMによる近似計算で選択する．適応化されていない不特定話者モデルが選ばれると新たな話者が登場したとして，その話者のためのモデルを作成する．GMMを用いることでモデル選択の最適性は保証されないが，計算量を大幅に削減しつつ認識性能の向上がほぼ同等であることを確認している．

これらの手法を音声対話システムのような応答の即時性を要求するタスクに適応するためには，複数の計算機を用いて音声認識を同時並行に行い，モデル適応をバックグラウンド処理する必要がある．しかし，変動を想定したモデルの増加に伴い，必要となる計算機の数も増加するため，これまで大規模な実装例はほとんど見られなかった．

一方，マルチコアプロセッサの登場や，GRID[3]のような数百から数千を超える計算ノードを持つシステムが次々と登場しており[4]，今後もさらに大規模な並列計算機が増えると予想される．このように，プロセッサ単体の性能向上もさることながら，計算機システム全体の性能向上が期待できる．

以上のような背景から，音声による飲食店舗検索システム[5],[6]を超並列計算機上に構築することを検討する．[5],[6]では，発話内容(話題・発話カテゴリ)ごとの言語モデルで認識して得られる仮説を尤度を基準に選択することで，ユーザからの様々な発話を受け付けている．本稿では，これに複数の話者依存音響モデルを組み合わせ，逐次適応も行うことで頑健性の向上を目指す．実時間での対話応答を可能にするため，言語モデルと音響モデルの組み合わせ数の音声認識器を，超並列計算機上の別々の計算ノードで駆動し，それらの出力する認識仮説を尤度を基準に選択する．また逐次適応をバックグラウンド処理させ，適応後のモデルを持つ認識器もバックグラウンドで起動しておく．この構成は，次のようなメリットを持つ．

即時性 適応モデルの生成およびそのモデルを持つ認識器の起動を認識や対話制御とは別の計算ノードで行うことで，対話の進行と独立させている．新しい認識器は，起動後の発話から用いられる．

最適性 従来のモデル選択手法では，近似計算などで認識するモデルを予め限定する．よって，モデル選択誤りにより認識仮説の尤度が最大ではない可能性がある．全てのモデルでの認識を行うことで，最尤の認識仮説を得ることができる．

拡張性 新たな変動に対応したモデルの追加が容易である．また，モデルの数を増加させても，システム全体の認識時間には影響がない．

耐障害性 認識仮説のうちのいくつかが得られなかったとしても，他の認識器からの仮説が利用できるので，認識仮説選択の最適性を保証する必要がなければ，再認識せずとも良い．また，故障したノードで駆動していた認識器を別の計算ノードに移すことも容易である．

移植性 構築したシステムを他のタスクに移植する場合，共通で利用できるモデルはそのままで，タスクに依存したモデルのみを置き換えることで実現できる．

実装した音声対話システムの頑健性向上を確認するため，[5],[6]で収録した，飲食店舗検索システムを使用したときの音声を用いてシミュレーション実験を行った．以降，2.ではGRID上に実装したシステムの構成について述べ，3.で認識仮説の選択方法について述べる．4.ではシステムの対象とするタスクについて説明し，5.で評価実験について述べ，6.で本稿をまとめる．

2. システム構成

構築した対話システムの構成を図1に示す．本システムは，インタフェース部(I/O manager)，対話制御部(Dialogue manager)，認識ノード管理部(Recognition node manager)，適応ノード管理部(Adaptation node manager)，データベース検索部(Database retriever)からなる．各部はTCPのソケット通信によりメッセージや認識データを送受信している．

認識ノード管理部と適応ノード管理部は，モデル管理リスト(Model management list)とノード管理リスト(Node management list)を共有している．モデル管理リストは，システムが利用可能な言語モデルおよび音響モデルの情報を保持しており，適応により新たに作成されたモデルには，適応のもととなったモデルの情報も記録している．ノード管理リストは計算ノード(CPU node)の利用状況を反映させたもので，空き状況，認識ノード(Recognition node)として使用，適応ノード(Adaptation node)として使用などのステータスと，使用中の言語・音響モデルが記録されている．

インタフェース部 ユーザからの音声入力を受け付けて対話制御部に送る．対話制御部から受け取ったプロンプトや検索結果をテキストで出力する．

対話制御部　インタフェース部から受け取った音声を認識ノード管理部に送り，複数の認識仮説を受け取る．それらの中から最適な認識仮説を選択し，それに応じた応答を生成してインタフェース部に送る．このとき必要であればデータベース検索部に検索要求を出し，その結果をインタフェース部に送る．また，音声とその認識仮説などのモデルの適応に必要なデータを適応ノード管理部に送る．

データベース検索部　対話制御部から検索用キーワードが送られてくると，該当する店舗を検索し，店舗数と店舗名を対話制御部に送る．店舗名が送られてくると，その店舗の詳細情報を検索して送る．

適応ノード管理部　対話制御部から適応用データを受け取ると，モデル管理リストを参照し，全てのモデルの適応用に計算ノードを確保してノード管理リストの当該ノードのステータスを適応ノードに変更する．適応が終わるとノードのステータスを計算ノードに変更し，認識器を起動するため計算ノードを確保する．認識器が起動するとステータスを認識ノードに変更する．また，適応化したモデルの情報をモデル管理リストに追加する．適応ノード管理部は対話制御部や認識ノード管理部とは独立に適応および認識器の起動を行っている．つまり，認識ノード管理部や対話制御部は適応や認識器の起動を待つことはない．

認識ノード管理部　対話制御部から音声が送られてくると，ノード管理リストとモデル管理リストを参照し，最新のモデルを持つ認識ノード全てに音声データを送る．全認識ノードから認識仮説を受け取るとそれを対話制御部に送る．

なお，本稿ではモデル適応を行ったモデルと適応前のモデルを置き換えることとする．このため，モデル適応が終わると，適応前のモデルは古いモデルとしてモデル管理リストから削除される．また，実装したシステムでは，インタフェース部と対話制御部とデータベース検索部は同一のノードに実装しており，認識ノード管理部と適応ノード管理部も同一ノードに実装している．

3. 認識仮説選択

対話制御部では，認識ノード管理部から得られる複数の認識仮説から尤度最大の認識結果を選択する [5]．つまり，x を入力特徴量，w を仮説単語列，c を発話内容として，

$$P(c,w|x) = \frac{P(x|c,w)P(c,w)}{P(x)} = \frac{P(x|c,w)P(w|c)P(c)}{P(x)} \quad (1)$$

であり，

$$P = P(x|w)P(w|c)P(c) \approx P(x|w)P(w|c)^{\alpha}P(c)^{\beta} \quad (2)$$

の P が最大となる単語列 w を認識仮説とする．ここで $P(x|w)$ は音響尤度，$P(w|c)$ は言語尤度であり，$P(c)$ は発話内容尤度である．また，α は言語重み，β は発話内容重みである．

発話内容尤度 $P(c)$ は，予めシステム設計者が対話戦略に基づいて設定したルールや，大規模な対話コーパスから統計的に作成したモデルからの尤度として求められるが，本稿では簡単のため，全ての発話内容の生起が一様であると仮定して，発話内容重み β を含めた $P(c)^{\beta}$ を一定値とした．言語重み α は，5.1 で述べる認識パラメータ推定用のデータを用いた予備実験で認識性能が最大となる値を用いた．

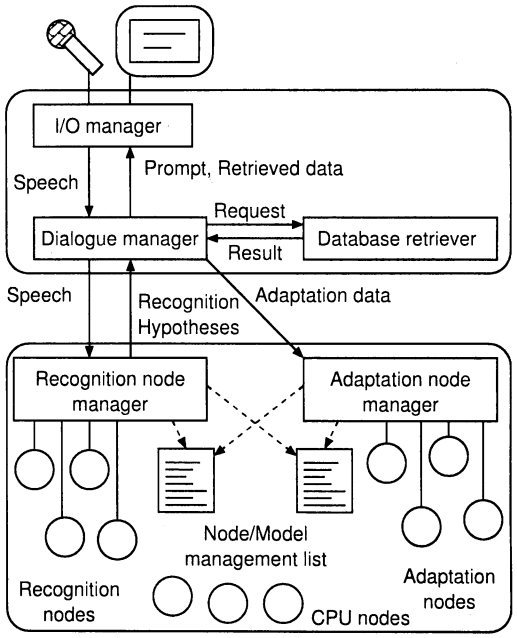

図 1　システム構成図
Fig. 1　System architecture.

4. 飲食店舗検索システム

4.1 対話タスク

本稿で対象とする対話タスクは，音声入力と画面出力による飲食店舗検索である．ユーザは，場所，料理種，価格帯やサービスや設備などの条件により候補となる店舗を絞り込み，最後に希望の店舗の詳細情報を画面に表示させる．場所としては東京圏の駅名 1,070 を，料理種としては 80 業種 307 キーワードを認識対象とする．店舗情報データベースは「インターネットタウンページ [7]」からの約 4,400 店舗分のデータをもとに作成した．

4.2 対話戦略とシステムの内部状態

本稿での対話戦略を図 2 を用いて説明する．図中の System State はシステムの内部状態であり，システムが画面出力するメッセージはこの内部状態ごとに異なる．また，内部状態 A を対話開始の状態とする．

内部状態 A　検索の基本的な条件として，場所と料理種を受け付け，内部状態 B に遷移する．画面には，場所と料理種の発声を促すメッセージが出力される．

内部状態 B　内部状態 A での発話を確認する．画面上に認識

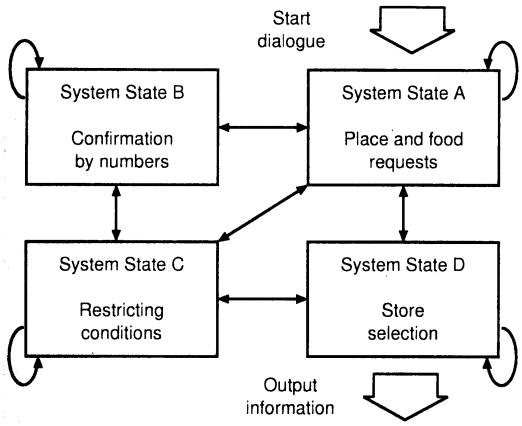

図 2　システムの内部状態の遷移図
Fig. 2　System state transition.

した場所と料理種を表示し，「両方正しい」「場所のみ正しい」「料理のみ正しい」「両方間違っている」の 4 つから番号発声により選択してもらう．「両方正しい」が選ばれると内部状態 C に遷移し，それ以外は内部状態 A に戻る．

内部状態 C　場所と料理種のみの条件で候補となる店舗数が閾値よりも多い場合，その他の絞り込み条件を発声してもらう．店舗の候補数が閾値以下になると内部状態 D に遷移する．本稿ではこの閾値を 10 としている．

内部状態 D　候補の店舗名を番号を付けて表示する．その番号や店舗名が発声されると，選ばれた店舗の詳細情報を画面に表示する．ここではさらに絞り込み条件を受け付けることができる．

なお，ユーザはいつでもコマンドを発声することで直前の対話に戻ったり，絞り込み条件を修正したり，初期画面に戻すことができる．

4.3 発話内容

この対話戦略においてユーザが発声する発話内容 (話題・発話カテゴリ) を次の 5 つに分類した．

PnF　場所と料理種の少なくとも一方を含む発話．
RES　場所と料理種以外の検索用キーワードで，サービス，施設，営業時間，価格帯などを含む発話．
NUM　内部状態 B での場所と料理種の確認における番号選択や内部状態 D での店舗選択での発話．
STO　データベース中の全店舗名を含む発話．
COM　システムの状態を直接操作するコマンド発話．

これら 5 つの発話内容ごとに言語モデルを作成した．システムは各内部状態で様々な発話を受け付けるが，特に内部状態 D では，**RES**, **STO**, **NUM**, **COM** の発話内容を受け付け，それに基づき内部状態を遷移させたり検索結果の表示や店舗の詳細情報を出力する．

表 1　発話内容ごとの語彙数，キーワード数，書き起こし発話数
Table 1　Vocabulary size, number of keywords and number of sentences for each speech content.

	PnF	RES	STO	COM	NUM	1LM
vocaburary	1,586	684	4,442	133	120	6,965
keyword	1,387	322	4,347	23	23	6,102
sentence	489	558	18	76	17	1,158

5. 評価実験

5.1 実験データ

評価に用いる対話音声は，研究室内で [5] の対話システムを用いて収録した．話者は研究室に所属する男性話者 19 名である．場所，料理種，絞り込み条件などは予め設定するが，それ以外には特に制限を設けず自由に発声してもらった．1 話者あたり複数のタスクを連続して行い，評価実験は話者ごとに行った．つまり話者交代はない．評価データはこのうちの 14 名分，2,416 発話，2,757 キーワードとし，残りの 5 名分，570 発話，655 キーワードは言語重みなどの認識パラメータ推定用データとした．

発話内容ごとの 5 つの言語モデルは 3-gram でモデル化した．場所と料理種を含む発話 **PnF** を受け付ける言語モデルは，飲食店舗検索システムを実際に利用したときの音声を書き起こして作成した [5]．それ以外の言語モデルは，システム設計者がシステムの利用を想定して作成した書き起こしを用いて作成した．いずれもフィラーを含み，キーワードはクラス化している．各モデルの語彙数，対話制御に用いられるキーワードの数を表 1 に示す．また，各モデルの作成に用いた書き起こし発話数も併記した．単一のモデルによる従来の認識システムと提案するシステムの比較のために，5 つの発話内容全てを受け付ける単一言語モデルを，5 つの言語モデルを線形補間することで作成した (1LM)．各発話内容の補間係数は認識パラメータ推定用データで最適値を推定して用いた．

ベースラインの音響モデルは，[8] に同梱されている JNAS 読み上げ音声データベースによる性別非依存 2,000 状態 16 混合 triphone HMM を用いた (BS モデル)．さらに，評価データのうち被験者を除く 13 名の音声を用いて，BS モデルを話者適応することで，不特定話者モデル (SI モデル) を作成した．また，被験者を除く 13 名それぞれに適応化した 13 個の特定話者モデル (SD モデル) を作成した．SI モデルは BS モデルとの収録環境等の違いを考慮したモデルで，SD モデルはそれに加えて話者の違いも考慮したモデルとなっている．適応には MLLR [9] および MAP [10] を教師ありで組み合わせて用いた．適応用のデータは 1 話者平均で 172.6 発話であった．

なお，実験では，これら複数の言語モデルと音響モデルの総組み合わせ数の認識ノードを並列に駆動する．例えば，SI モデルのみ，SD モデルのみ，BS, SD, SI モデル併用のそれぞれ実験では 5, 65, 75 ノードを同時に駆動している．

音響特徴量は 12 次元 MFCC とその差分，パワーの差分の計 25 次元とし，認識は Julius 3.4.2 [11]，適応は HTK 3.2 [12] で

表 2　各手法におけるキーワード正解精度 (%)
Table 2　Keyword accuracy(%) of each methods.

Acoustic models	BS	SI	SD	BS +SI	BS +SD	SI +SD	BS +SI +SD	BS	SI	SD	BS +SI	BS +SD	SD +SI	BS +SI +SD
	without incremental adaptation							with incremental adaptation						
1LM	89.4	89.4	89.7	89.7	89.7	90.3	90.2	89.7	90.1	91.7	91.1	91.7	91.8	91.8
Recognition nodes	1	1	13	2	14	14	15	1	1	13	2	14	14	15
5LM	90.9	90.4	90.6	91.0	90.6	91.3	91.4	91.0	91.0	92.0	91.8	92.0	92.0	92.1
Recognition nodes	5	5	65	10	70	70	75	5	5	65	10	70	70	75
Adaptation nodes	—							1	1	13	2	14	14	15

行った．

5.2 逐次適応

前節で述べた複数の音響モデルに対し，MLLR と MAP の併用による教師なし逐次話者適応を行った．本システムはモデル適応をバックグラウンドで行い，適応が終わると認識器を起動するが，モデル適応と認識器の起動は数発話分の時間がかかる．認識ノード管理部は適応ノード管理部での処理を待たずに，そのときに利用できるモデルを用いて認識を行う．しかし，ここではシミュレーション実験として，入力音声 5 発話ごとにそれまでの発話全てを用いて適応を行い，適応後全てのモデルを入れ替えて次の 5 発話を認識することとする．なお，これにより，各発話を認識する認識ノードの数は常に一定となる．

5.3 実験結果

評価データ 2,416 発話に対するキーワード正解精度を表 2 に示す．発話内容ごと 5 つの言語モデルを用いた提案手法の結果は 5LM とし，従来手法である単一言語モデルによる結果は 1LM とした．

単一言語モデル (1LM) で逐次適応なし (without incremental adaptation) の条件で，ベースラインである BS モデルのみの結果と比べ，複数の特定話者モデルによる SD ではキーワード正解精度で 0.3 ポイント改善した．さらに SI モデルや BS モデルを併用したところ (BS+SI+SD)，0.5 ポイントの改善を示した．これはベースラインと比較したキーワード誤り削減率では 7.5%の改善となる．また，音響モデルの逐次適応を組み合わせたところ (with incremental adaptation)，さらに 1.6 ポイントの性能改善を示した．キーワード誤り削減率は 15.1%となる．これらにより，音響モデルを複数用いることで，単一の言語モデルであっても認識性能が向上し，さらに，逐次適応を組み合わせることで，その効果が大きくなることが確認された．

次に，発話内容ごとの 5 つの言語モデル (5LM) を用いたところ，BS モデルのみではベースラインと比較して，キーワード正解精度で 1.5 ポイント，誤り削減率で 14.2%の性能改善となった．複数の音響モデルを用いることで さらに 0.5 ポイントの改善を示した．ベースラインと比較したキーワード誤り削減率は，18.9%となる．言語モデル，音響モデルとも複数用い，逐次話者適応を組み合わせることで 0.7 ポイントの改善を示した．これは，本実験におけるキーワード正解精度では最大であり，ベースラインと比較したキーワード認識誤り削減率では 25.5%の改善である．このとき，認識ノードでは 75 台を用い，適応ノードでは 15 台を用いていた．同条件で本システムを利用して対話を行ったが，単一の認識器のみでのシステムと認識時間の差はほとんどなかったことを確認している．以上より，複数の音響モデルと言語モデルを用い，逐次適応を行うことで，システムの頑健性向上が確認された．

6. おわりに

入力音声の変動を予測した複数のモデルを同時に利用し，予測が困難な変動に対してモデルを逐次適応化する頑健な音声認識システムを検討した．複数の認識器を同時に駆動し，さらにモデルの逐次適応も行うため計算量の問題が生じるが，これを超並列計算機上に実装することで，実時間での応答が可能な対話システムを構築した．

ベースラインである単一の言語モデルと音響モデルによる結果と比べ，音響モデルのみ複数用いて並列に駆動することでキーワード誤り率が 7.5%削減され，言語モデルのみ複数用いて並列駆動することで 14.2%削減された．さらに，言語モデル，音響モデルとも複数用いることで 18.9%削減された．また，逐次話者適応を組み合わせることにより 25.5%の誤り削減率となった．このとき用いた計算ノードの数は，認識ノード，適応ノード合わせて最大で 90 台であった．これらの結果により，構築したシステムが入力音声の予測可能な変動および予測困難な変動に対して頑健であることが示された．

今後の課題として，本稿では認識仮説の選択を尤度を基準に行ったが，[13] で提案されているような，複数の認識仮説中から共通して現れる単語を多数決により選択して出力する手法を組み込むことを検討する．また，本稿では認識デコーダとして Julius のみを用いた．しかし，文法ベースの言語モデルを用いる Julian [11] や，SPOJUS [14] といった異なる認識デコーダによる認識結果の統合手法 [15] を組み込むことで，可搬性と性能のさらなる向上が期待できる．また，言語重みなどの認識パラメータを逐次適応に合わせて最適化する枠組みを検討したい．最後に，構築したシステムでの被験者実験を行い，主観評価やタスク達成率，平均対話ターン数などの客観評価も行いたい．

謝辞　飲食店舗データベースを提供して頂いた NTT 番号情報株式会社に感謝する．本研究は科学研究費補助金学術創成研究「言語理解と行動制御」の一部として行われている．

文　献

[1] T. Shinozaki and S. Furui, "Spontaneous speech recognition using a massively parallel decoder," Proc. ICSLP2004, vol.3, pp.1705–1708, Jeju Island, Korea, Oct. 2004.

[2] Z.-P. Zhang, S. Furui and K. Ohtsuki, "On-line incremental speaker adaptation for broadcast news transcription," in Speech Communication, vol.37, nos.3-4, pp.271–281, Jul. 2002.

[3] 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, Jun. 2003.

[4] http://www.top500.org/, TOP500 Supercomputer Sites.

[5] 田熊竜太, 森山達裕, 岩野公司, 古井貞熙, "並列処理型計算機による混合主導型音声対話システムの構築," 音講論, pp.79–80, Sept. 2002.

[6] 田熊竜太, 岩野公司, 古井貞熙, "並列処理型計算機を用いた音声対話システムの検討," 人工知能学会 研報 SLUD-A201-04, pp.21–26, Jun. 2002.

[7] http://itp.ne.jp/ , NTT 番号情報株式会社.

[8] 連続音声認識コンソーシアム 2003 年度版ソフトウェア.

[9] C.J. Leggetter and P.C. Woodland, "Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models," in Computer Speech and Language, pp.171–185, 1995.

[10] J.L. Gauvain and C.H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observation of Markov Chains," in IEEE Transactions on Speech and Audio Processing, vol.2, no.2, pp.291–298, 1994.

[11] http://julius.sourceforge.jp/, 大語彙連続音声認識ソフトウェア Julius.

[12] http://htk.eng.cam.ac.uk/, Hidden Markov Model Toolkit (HTK).

[13] J.G. Fiscus, "A Post-processing System to Yield Reduced Error Word Rates: Recognizer Output Voting Error Reduction (ROVER)," Proc. ASRU, pp.347–354, Santa Barbara, USA, Dec. 1997.

[14] 北岡教英, 高橋伸寿, 中川聖一, "N-best 線形辞書検索と 1-best 近似木構造辞書探索の併用による大語彙連続音声認識," 信学会論文誌, Vol.87-DII, No.3, pp.799–807, Mar. 2004.

[15] 宇津呂武仁, 小玉康広, 渡辺友裕, 西崎博光, 中川聖一, "機械学習を用いた複数の大語彙連続音声認識モデルの出力の混合, " 信学会論文誌, Vol.87-DII, pp.1428-1440, Jul. 2004.