論文 / 著書情報
Article / Book Information

論題(和文)	超並列デコーダと逐次適応を用いた音声対話システム
Title(English)	
著者(和文)	中川竜太, 岩野公司, 古井貞熙
Authors(English)	Ryuta Nakagawa, Koji Iwano, SADAOKI FURUI
出典(和文)	日本音響学会2005年秋季講演論文集, Vol. 〔1-7-8〕, No. , pp. 7-8
Citation(English)	, Vol. 〔1-7-8〕, No. , pp. 7-8
発行日 / Pub. date	2005, 9

超並列デコーダと逐次適応を用いた音声対話システム*

◎中川竜太, 岩野公司, 古井貞熙 (東工大)

1 はじめに

入力音声の種々の変動に頑健な音声認識の研究として，入力音声に対し逐次適応するもの [1] や，予測される変動に適したモデルを予め複数用意し，それらで認識した結果から最適なものを選択するもの [2] がある．前者は未知の変動にも対応可能であり，後者は入力の変動が予測したものであれば最適なモデルで認識が可能であるという特徴を持つ．[2] ではまた，複数の認識を行うことによる認識時間の問題を，超並列デコーダを用いて解決している．

そこで，飲食店舗検索をタスクとする音声対話システム [3],[4] に対し，これらの手法を組み合わせて用いることを検討する．[3],[4] 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，その結果を選択することで，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，対話システムを超並列計算機上に実装した．そして頑健性の向上を確認するためのシミュレーション実験を行った．

2 システム構成

構築した対話システムの構成を Fig. 1 に示す．本システムは，インタフェース部，対話制御部，認識ノード管理部，適応ノード管理部，データベース検索部からなる．認識ノード管理部と適応ノード管理部は，ノード管理リストを共有している．ノード管理リストは計算ノードの利用状況を反映させたもので，空き状況，認識ノードとして使用，適応ノードとして使用などのステータスが記録されている．

インタフェース部は，ユーザからの音声入力を受け付け，プロンプトや検索結果をテキストで出力する．

対話制御部ではインタフェース部からの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，対話制御部は音声とその認識仮説の音素列などの適応データを適応ノード管理部に送る．

適応ノード管理部は，対話制御部から適応データを受け取ると，ノード管理リストから利用可能な計算ノードを見つけ，そのノードを適応ノードとする．適応ノードは受け取った適応データでモデルの逐次適応を行う．適応が終わったノードを認識ノードとして起動し，ノード管理リストのステータスを変更する．

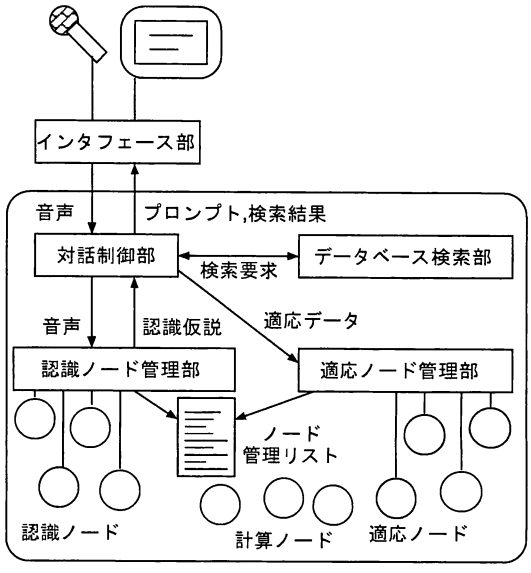

Fig. 1 System architecture.

適応ノード管理部では他とは独立に適応，認識ノード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ノードの起動を待つことはない．

認識ノード管理部は，対話制御部から音声が送られてくると，ノード管理リストを参照し，認識ノードとして起動している全ノード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．また不要となったモデルを持つ認識ノードを停止し，ノード管理リストのステータスを変更する．

対話制御部での認識仮説選択は次式に従う [3]．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，$P(c)$ 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．なお，パラメータ $\alpha, \beta, P(c)$ はそれぞれ後述の学習データを用いた予備実験で認識精度が最大となる値を用いた．

*A massively parallel decoder-based spoken dialogue system with incremental adaptation. By NAKAGAWA Ryuta, IWANO Koji, and FURUI Sadaoki (Tokyo Institute of Technology)

Table 1 Keyword accuracy of the proposed method.

Acoustic Models	BS	SD	SI	SD +SI	BS +SI	BS +SD +SI	BSinc	SDinc	SIinc	SDinc +SIinc	BSinc +SIinc	BSinc +SDinc +SIinc
Accuracy(%)	84.3	84.6	85.5	85.7	85.6	85.8	85.6	85.4	86.9	86.5	87.2	87.2

3 評価実験

3.1 実験データ

研究室内で対話システム [3] を用いて収録した．男性話者 19 名で，設定された店舗検索タスクに対し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った．評価実験は話者ごとに行うとした．つまり話者交代はない．このうちの 14 名分 2,416 発話を評価データとし，5 名分 570 発話を学習データとして言語重みなどの認識パラメータ推定に用いた．

3.2 変動を考慮したモデル

[3],[4] では，入力音声の変動要因のうちの発話内容に注目し，予測される発話内容をカテゴリにわけ，それぞれの 3-gram モデルを作成している．駅名と料理種，価格帯やサービスや施設などの絞り込み条件，店舗名，番号，コマンドの 5 カテゴリである．

ここでは，音響的な変動要因を考慮した複数の音響モデルを組み合わせて用いた．ベースとなる音響モデル (BS モデル) は，[5] 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 triphoneHMM を用いた．

本稿では，話者による違いを考慮した特定話者モデル (SD モデル) と，収録環境の違いや読み上げか対話かという発話スタイルの違いを考慮した不特定話者モデル (SI モデル) を用意した．モデルの適応には評価データを用い，held out 法による open な実験を行った．SD モデルは，評価話者を除く 13 名それぞれの音声を用いて BS モデルを適応化した．評価実験では 13 モデルを並列的に用いて認識する．SI モデルは，評価話者を除く 13 名の音声全てを用いて BS モデルを適応化した．適応化手法はいずれも教師ありで MLLR と MAP を併用した．

3.3 逐次適応 (inc)

前節で述べた複数の音響モデルに対し，入力音声 5 発話ごとに教師なし適応を行った．適応は MLLR と MAP を併用した．なお，適応後はモデルを置き換えた．つまり，入力 5 発話ごとに，それまでの発話に対し適応したモデルで認識している．またこれにより，各発話を認識する認識ノードの数は常に一定数である．

4 実験結果

評価データ 2,416 発話に対する実験結果を Table 1 に示す．逐次適応なしの条件で，ベースラインである BS 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 ポイントの改善に留まった．これは各話者の適応データが平均 172.6 発話と少なく，また，並列的に用いる特定話者モデルが 13 名分と少ないため，評価対象話者と近いモデルが用意できていなかった可能性がある．特定話者モデルの適応データを増やし，さらに多くの特定話者モデルを並列的に用いることで，性能向上が見込めると考えられる．一方，収録環境や発話スタイルに適応した SI モデルでは，ベースラインと比較して 1.2 ポイントの性能改善を示した．また，BS モデル，SD モデルも併用し，15 モデルを並列的に使用した場合 (BS+SD+SI)，さらに 0.3 ポイントの性能改善を確認した．

逐次適応を組み合わせた実験では，逐次適応を用いない場合よりいずれの条件でも性能が改善した．BS, SD, SI モデル全てを用いて逐次適応した場合 (BSinc+SDinc+SIinc)，ベースラインより 2.9 ポイントの改善となった．これは，単一の音響モデルを逐次適応する SIinc と比較しても，0.3 ポイントの改善となっており，入力音声の変動を考慮した複数の音響モデルと逐次適応を組み合わせた本システムの有効性を示すものである．

5 おわりに

複数の適応モデルを同時に利用し，さらに逐次適応も組み合わせることで入力音声の変動に頑健な音声対話システムを構築した．複数の認識を同時に駆動し，さらに適応も行うため，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解決し，実時間での認識を可能にした．

今後は，認識仮説の選択手法として，[6] 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キーワードを出力する手法や，異なる認識デコーダによる認識仮説の統合手法 [7] を組み込むことで，さらなる性能向上を目指す．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．

参考文献

[1] Zhang *et al.*, Speech Communication, vol.37 nos.3-4, 271-281, 2002.
[2] 篠崎，古井，音講論 (春)，111-112，2004.
[3] 田熊 他，音講論 (秋)，79-80，2002.
[4] 田熊 他，人工知能学会 研報 SLUD-A201-04, 21-26，2002.
[5] 連続音声認識コンソーシアム 2003 年度版ソフトウェア
[6] Fiscus, Proc. IEEE ASRU, 347-354, 1997.
[7] 渡辺 他，音講論 (秋)，119-120，2003.