# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| 論題(和文) | 逐次話者適応を用いた並列処理型会議音声認識システムの検討 |
| Title(English) | |
| 著者(和文) | 田熊 竜太, 岩野 公司, 古井 貞熙 |
| Authors(English) | Koji Iwano, SADAOKI FURUI |
| 出典(和文) | 日本音響学会 2002年春季講演論文集, Vol. , No. 2-5-16, pp. 105-106 |
| Citation(English) | , Vol. , No. 2-5-16, pp. 105-106 |
| 発行日 / Pub. date | 2002, 3 |

# 逐次話者適応を用いた並列処理型会議音声認識システムの検討*

◎田熊 竜太，岩野 公司，古井 貞熙 （東工大）

## 1．はじめに

複数話者による会議音声をオンラインで音声認識するシステムを構築している．話者交代を検出しながらオンラインで教師なし話者適応を行う．並列型計算機によって特定話者モデルをもつ複数の認識器を同時並行に駆動し，尤度を基準に認識結果を選択する．入力音声が未知話者の音声の場合にはその音声を用いて話者適応を行い，新たに特定話者モデルを生成する．

本稿では逐次話者適応化法を組み込んだ並列処理型音声認識システムの概要を説明し，実際の会議音声を用いて評価した結果について論ずる．本システムは，オンライン話者適応法[1]を並列処理型音声認識システム[2]に融合することにより，計算時間の大幅削減を実現している．

## 2．システムの概要

本システムは，音声入力受け付け・認識結果集計・結果出力・話者適応を行うコアサーバーと，音声認識を行う複数の音声認識モジュールからなる．複数の特定話者モデルでの認識を，複数の認識モジュールで並列処理することにより認識時間を短縮している．

### 2.1．コアサーバー

コアサーバーでは発話ごとに区切られた音声入力を受け付け，その音声を複数の認識モジュールに渡す．全ての認識モジュールからの出力を受け取り，尤度が最大になる認識結果を出力する．特定話者用音響モデルを持つ認識モジュールの出力の尤度が最大だった場合はその特定話者モデルをインクリメンタルに適応化する．不特定話者用音響モデル(SIモデル)をもつ認識モジュールの尤度が最大だった場合は，新たな話者の登場と考え，SIモデルを適応化することにより特定話者モデルを生成し，新たな認識モジュールを起動する．話者適応は教師なしであり，MLLR適応を行った後，MAP適応を行う．話者適応とモジュールの増加の様子を図1に示す．

### 2.2．認識モジュール

認識モジュールではコアサーバーから音声データと音声認識用のモデルを受け取る．音声データから特徴量を抽出し，与えられたモデルで音声認識をし，対数尤度スコアと認識結果をコアサーバーに返す．特徴量としてはMFCC12次元,その微分12次元,対数パワーの1次微分の計25次元を利用した．音声認識にはjulius3.1を用いた．

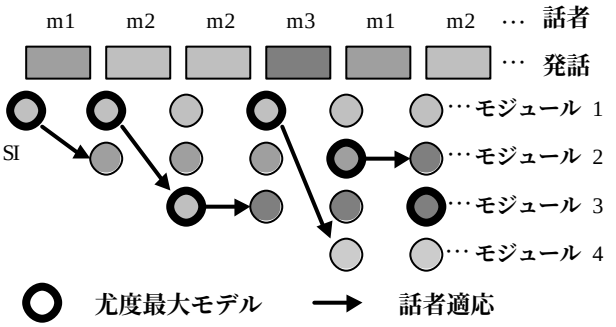


図1．話者適応とモジュールの増加の様子

## 3．認識実験

### 3.1．認識タスク

実験には1999年6月6日放送のNHK「日曜討論」約1時間の音声を利用した．発話はあらかじめ文単位で切り出した．この番組は司会も含め男性6名,女性1名による討論番組である．このうち女性の話者と男性話者1名は発話数が他の5人に比べて非常に少ないため，学習および認識の対象から外した．テストセットパープレキシティは292.2であった．

### 3.2．言語モデル・音響モデル

言語モデルはWWW上で公開されている講演書き起こしテキスト[3]から作成し，全認識モジュールで同一の物を利用した．語彙数は2万語であり，書き起こしテキストに含まれていない未知語(259語)は全て辞書に登録した．本稿では一つの不特定話者用の音響モデル(SIモデル)を初期モデルとして利用した．SIモデルは『日本語話し言葉工学』プロジェクトで作成された音声コーパスから作成した[4].

### 3.3．予備実験

まず，SIモデルのみを用いて全評価データを認識した(SI)．これは本システムのベースラインとなる結果である．次に，評価データの前半と後半を分け，一方の発話を全て用いて5人の特定話者モデル作成し，他方を評価した．特定話者モデル作成時の話者適応化手法は教師ありで，MLLRおよびMAPを併用した．これら5つの特定話者モデ

---

* Parallel computing-based meeting speech recognition system with incremental on-line speaker adaptation.
By Ryuta Taguma, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

ルとSIモデルを持つ認識モジュール計6つを同時並行に駆動した．入力音声の発話者は未知なので，尤度を基準に最適な認識結果を選択し，出力した(SA)．認識時には逐次話者適応は行っていない．

SI，SAの条件での各文の音素数とその正解精度を図2にまとめた．特に音素数が少ない発話(30音素以下 図中網掛け)で認識率が低いことがわかった．本稿で利用する逐次話者適応法では，この音素列を教師なし適応時に音素書き起こしとして利用するので，音素数が少なく，音素誤り率が高い場合，逐次話者適応により適切な特定話者モデルが得られない．よって今回は，認識結果の音素数が30以下の発話は逐次適応しなかった．SIモデルのみの予備実験で，音素数30以下の発話は85文で全評価データ(574文)の14.8%であった．

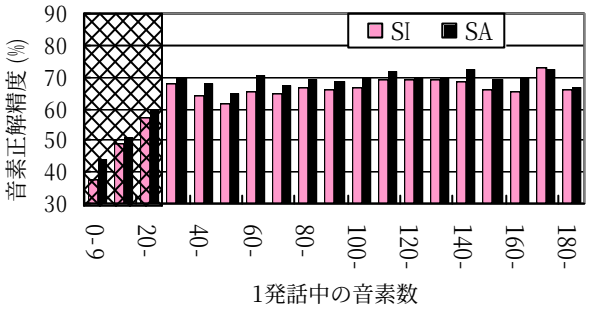


図2．1発話中の音素数と音素正解精度

## 3.4．実験

SIモデルを持つ認識モジュールのみから開始して発話ごとに逐次話者適応を行った実験の結果(SI+OA)を図3に示す．ベースラインとなるSIモデルのみで逐次適応を行わない結果(SI)，および会議に参加する話者それぞれに適応化したモデルを事前に用意し逐次適応を行わない結果(SA)も併記した．重み係数などは経験的に最適なものを利用した．

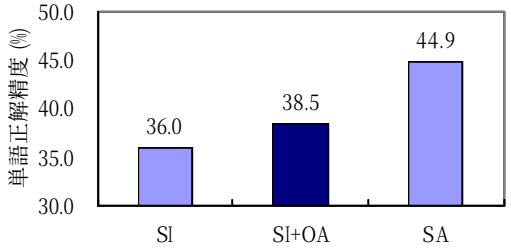


図3. 提案システムの認識性能

SAではベースラインより正解精度が絶対値で8.9%上昇したが，提案手法であるSI+OAでは2.5%の上昇にとどまった．これはSIモデルの認識率が低く，多くの誤りが含まれる音素書き起こしで逐次話者適応を行ったため，十分な精度の特定話者モデルが得られなかったためと考えられる．

提案システム(SI+OA)により，評価データ全てを認識した結果として，適応化されたモデルを持つモジュールが13個生成された．モジュールの増加の様子を図4に示す．

全評価データに対し各モジュールでの音声認識回数は6619回であった．音声認識時間を1発話あたり平均で$T$(秒)と仮定すると，単一プロセッサで処理した場合，総認識時間は6619$T$(秒)であり，並列処理した場合は総発話数×$T$=574$T$(秒)なので，処理時間が約1/11に削減されたことになる．音声認識がリアルタイムで処理できると仮定すると，提案システムは実時間で処理を終えることができるが，従来法ではその約10倍の時間がかかる．

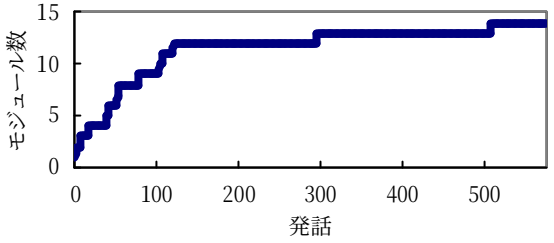


図4．発話ごとのモジュールの増加

## 4． まとめ

逐次話者適応法を用いた並列処理型音声認識システムを構築し,その有効性を示した．単一のモデルを利用したシステム(SI)よりも正解精度が絶対値で2.5%向上した．特定話者モデルを用いた場合 (SA)よりも性能は劣るものの，本システムは事前に特定話者モデルを用意できない場合でも，話者に適応したモデルで認識することが可能である．また並列処理化により，計算時間を単一プロセッサの場合の約1/10に削減できた．

今後は，複数の言語モデルを持つ音声認識モジュールも組み合わせることにより，話題を含む言語特徴の変化や多様性にも対応したシステムの構築を目指す．


## 謝辞

討論番組音声を提供していただいたNHK放送技術研究所の関係諸氏に感謝する．


## 参考文献


[1] Z.P.Zhang and S.Furui : ”On-line Incremental Speaker Adaptation for Broadcast News Transcription”, ICASSP, pp.961-964 (2000)

[2] 田熊，岩野，古井：”並列処理型会議音声認識システムの検討”，秋季音学講論, 3-1-11 (2001)

[3] 篠崎，細川，古井：”話し言葉コーパスを用いた音声認識の検討”，春季音学講論, 1-3-14 (2001)

[4] 篠崎，斎藤，堀，古井：”話し言葉音声の認識を目指して”，信学技報, SP2000-96 (2000)