論文 / 著書情報
Article / Book Information

論題(和文)	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 2001年秋季講演論文集, Vol. , No. 3-1-11, pp. 113-114
Citation(English)	, Vol. , No. 3-1-11, pp. 113-114
発行日 / Pub. date	2001, 10

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並列型計算機によってオンラインで音声認識するシステムを構築している. 異なるモデルを持った音声認識システムを同時並列に駆動することにより発話を認識し, 尤度を基準に最適な認識結果を選択することで, 単一の音声認識システムよりも頑健で高速なシステムとなる. 本稿では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ROVERシステム[1]の発展形と位置付けられる.

2. システム

並列型計算機による音声認識システムは, コアサーバーと認識モジュールの二つからなる. コアサーバーは発話者・発話環境の多様な音声を受け付け, 得られた音声データを複数のモジュールに受け渡す. 各モジュールはコアサーバーから音声データや各種パラメータによる指示を受け取り, 音声認識を行い, その結果やスコアなどをコアサーバーに返す. コアサーバーは複数のモジュールから受け取った結果を統合し最終結果を出力する．図１はその概念図である.

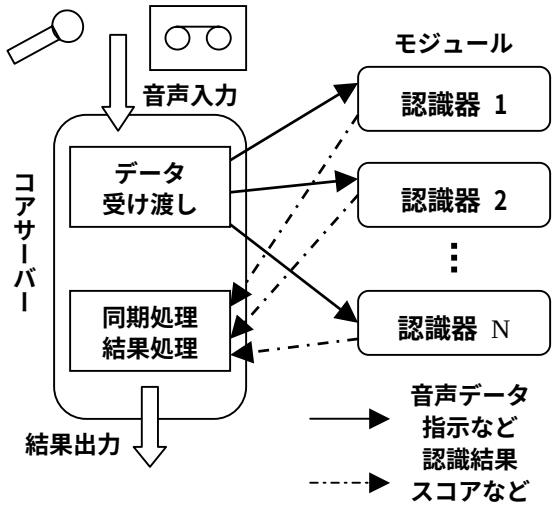

図１:提案システムの概念図

2.1. 構成

本システムは一台のコアサーバーと複数台のモジュールでネットワークを組み，コアサーバー・モジュール間ではTCP/IPを利用してデータや指示などのやり取りをする．全てのマシンはPC Linuxで構成されている.

2.2. 特徴

本システムのモジュールの数は必要に応じて自由に増やすことができるよう柔軟に設計されている. また状況に応じて各モジュールが使用するモデルやパラメータをコアサーバーが設定することができる. また処理速度や計算量の違いによるモジュールの処理時間の違いはコアサーバーが同期を取ることで解決している.

3. 実験

3.1. 使用データ

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. 実験に利用した残りの各話者の発話数を表１に示す.

表１ 番組中の話者別の発話数

話者	M1	M2	M3	M4	M5	All
発話数	217	93	80	91	93	574

各話者の発話文のうち前半部分をFセット,後半部分をSセットとした．Fセットの発話を評価データとして認識する場合はSセットを学習データとして用いた．Sセットの発話を評価データとして認識する場合はFセットを学習データとして用いた．Fセット, Sセットそれぞれの認識率の平均を評価に用いた．これにより番組中の全発話574文を評価した.

3.2. モジュール

モジュールはコアサーバーから送られてきた音声データから特徴量を抽出して音声認識をし, その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.

各モジュールは話者に対応した音響モデルを持つ．各話者用の音響モデルは不特定話者モデル(SIモデル)を初期モデルとして各話者の学習セットを用いてMLLR法により適応化した．これらに

---

* Meeting speech recognition system using parallel computing.
By Ryuta Taguma, Koji Iwano and Sadaoki Furui (Tokyo Institute of Technology)

利用されるSIモデルは『日本語話し言葉コーパス』プロジェクトで作成された音声コーパスから作成したものである[2]．このコーパスは男性話者による講演音声を元に作成されたもので, 今回の認識対象である会議音声を認識するのに適している．

5人の話者用モデルおよびSIモデルの計6種類の音響モデルを持つ認識モジュールを同時並行に駆動した. 各モジュールは話者に対応した話者ラベルが振られている(M1,M2,…,M5)．SIモデルにはSIというラベルを付与した．このラベルは各モジュールが認識結果をコアサーバーに返すときに同時に送られ, コアサーバーでの話者識別に利用される．

各モジュールでは共通の言語モデルを利用した．この言語モデルはWorld Wide Web上で公開されている講演書き起こしテキストから作成した[3]．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は全て辞書に登録した．

3.3．コアサーバー

マイク入力あるいはファイルから得られる音声データは発話者が不明なので, 発話ごとに全てのモジュールに渡す．音声は16kHz , 16bitでサンプリングする．その後コアサーバーは，全モジュールの出力同期を取り, 結果がそろうとスコア最大となる認識文をシステムの認識結果とし, その認識結果を出力したモジュールのラベルとともに出力する. 話者ラベルを出力することにより会議の流れをトレースしたり, 会議音声の書き起こしに発話者タグ付けすることができる．

4． 実験結果

本システムによる各話者の評価データごとおよび番組全体(All)の単語正解精度を図２に示す．比較として不特定話者モデルのみによる単語正解精度も併記した. 各話者用に適応化した音響モデルを同時並列で駆動し, その認識結果のスコアから尤度最大となる認識を選択した. これにより, 不特定話者モデルのみを利用した場合よりも正解精度が絶対値で6.3%向上した．

また, 本システムにより選択された話者ラベルと実際に入力された音声の話者ラベルを比較した．各話者における識別誤り数を表２に示す．話者識別の正解率は全体で95.6%である．この結果は尤度を基準に各話者用のモデルが正しく選択されていることを示している. これにより本システムが発話者特定にも有効であることが分かった．なお発話者識別誤りのほとんどは，短い発話の時に発生していた．

表２ ： 話者識別の誤り

話者	M1	M2	M3	M4	M5	All
発話数	217	93	80	91	93	574
識別誤り数	4	3	1	2	15	25

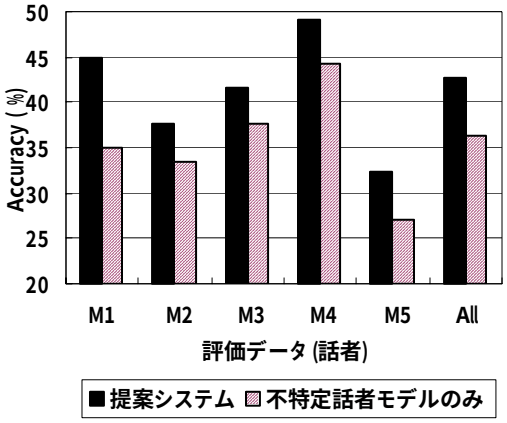

図２ ： 提案システムの認識性能

5． まとめ

並列型計算機による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とほぼ同等の認識時間で正解精度が絶対値で6.3%向上したことを示した. また会議を書き起こしする上で発話者を特定することは非常に重要であるが, 本システムでは, 複数の発話者からの音声の入力を95.6%の精度で正しく話者識別することができた．

本システムはコアサーバーからの指示でタスクにあわせて認識モジュールを変更できるよう設計されている. 本研究ではそれを行わなかったが, 発話内容によりコアサーバーが認識モジュールに指示を出して言語モデルを選択するような対話システムにも応用が可能である．

今後は複数の言語モデルを持つ音声認識モジュールを並行に駆動し音響モデルと組み合わせることで，より頑健な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] Jonathan G. Fiscus : ”A post-processing system to yield reduced word error rates : Recognizer output voting error reduction (ROVER)”, IEEE Workshop on Automatic Speech Recognition and Understanding , 1997

[2] 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)

[3] 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)