階層化された複数の音声認識器を選択的に利用する音声理解手法

横山　貴彦†　嶋田　和孝†　遠藤　勉†
九州工業大学情報工学部†

1　はじめに

近年では，誰しもが直観的かつ効率的に扱えるようなインタフェースの研究が広まっている．特にマイクやカメラなどの複数の媒体を取り入れたマルチモーダルインタフェースの開発に関する研究が盛んに行われている．

マイクを用いた音声インタフェースでは，高精度な音声認識が要求される．一般にキーワードスポッティングによる音声認識が広く用いられている．音声認識の結果を踏まえて何らかの処理を行うアプリケーションにおいては，システムへ入力される音声はその全てが命令ではなく，多くの非命令(雑談)を含む．スポッティングの手法を用いた場合，雑談の入力によって誤動作する場合[1]がある．したがって，命令と雑談を高精度に分別する必要がある．

嶋田らは，小語彙認識器と大語彙認識器を並列に動作させ，その出力の類似性を信頼度として用いて命令と雑談の分別問題を解決している [1]．また，並列化された複数の認識器の認識結果を信頼度によって判別することができる．語彙が機能別に分けられた複数の認識器を用いることで，機能の追加や削除といったメンテナンスを行う際の利便性を高められるという利点がある．

一般に音声認識は認識器の語彙が少ないほど高精度で行うことができる．そのため，入力する発話に依存関係があった場合は，多種類の発話を1つの認識器で認識するよりも，語彙数の少ない複数の認識器を使い分けて認識した方が認識率が向上する．例えば，あるシステムに「拡大」を入力したとする．この後に必ずその倍率が入力されるならば，2度目の認識時は倍率を認識する認識器のみを用いればよい．そこで，我々は認識器を階層化し選択的に用いる手法を提案した [2]．

認識器の数が複数ある場合は，複数の結果から判別を行う必要がある．判別候補が多くなると判別が難しくなり，判別を間違える可能性が高くなる．従来の階層化手法では，多くの認識器が動作する階層で精度が低下するという問題があった．本稿では，この点を改善するために新たな階層化手法を提案し，従来手法と比較を行った．

Speech Understanding with a Hierarchical Multiple Recognizer.
† Yokoyama Takahiko, Kazutaka Shimada and Tsutomu Endo, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology.

[1]例えば，システムが「拡大」という命令をスポッティングの手法によって認識するとする．このとき「ちゃんと拡大した?」と言うような雑談が入力された場合に，「拡大」を認識してアプリケーションが誤動作する問題がある．

2　従来手法

本研究では，音声を入力して動作させる写真管理アプリケーションを実験対象としている．入力語彙としては「拡大」や「検索」といった約30種類の命令と，それに対応する最大4桁の単位付き数値や日付などがある．階層化を施す前の非階層型手法では，これらの語彙は1つの認識器に登録されている．

階層型手法では，語彙を音声入力手順に踏まえて分類しておき，階層構造を作って音声入力時に動作する語彙の認識器を切り替えて用いる．すなわち，拡大命令を入力した後は対応する値を認識できる認識器のみを動作する．但し，階層構造によって複数の認識器の結果が同時に得られた場合，嶋田らの手法で候補を選ぶ．

我々は図1の階層構造のように命令の階層に細分化された認識器を用いる手法を提案した [2]．このような階層化手法を細分化階層法と呼ぶ．細分化とは拡大命令や回転命令といった単位で分割することを表す．命令の階層の認識器が細かく分かれていることにより各認識器が高い認識率となる．その反面，判別候補が多くなるため雑談分別率[2]が低下する．

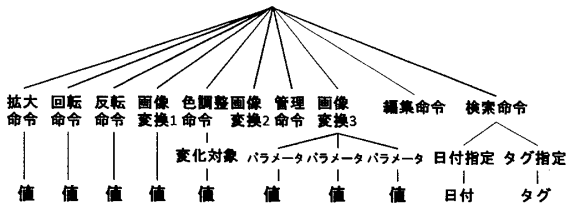

図 1: 細分化階層法における認識器の階層構造

3　提案手法

本節では，従来の細分化階層法における問題点を改善した統合型階層法を提案する．また，その手法と細分化階層法を組み合わせて認識器を選択的に利用する複合型階層法についても提案する．

3.1　統合型階層法

統合型階層法は，図2の階層構造のように命令の階層に全ての命令の語彙を登録した認識器を用いる手法である．すなわち，細分化階層法における拡大命令や回転命令といったものは，すべて1つの認識器にまとめられる．この結果，複数認識器による判別候補の乱立を避けることができるので，雑談分別率が低下しないことが期待できる．その反面，1つの命令の認識器の辞書に全命令が混在することになり認識率が落ちる可能性がある．

[2]雑談の入力を正しく雑談であると分別できる割合．

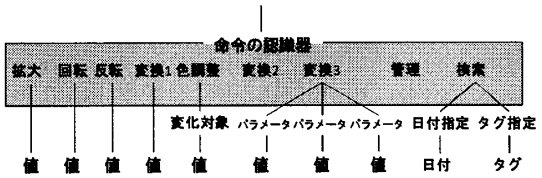

図 2: 統合型階層法における認識器の階層構造

3.2 複合型階層法

複合型階層法は，細分化階層法における各認識器の高い認識率と統合型階層法の高い雑談分別率に注目して，それらを組み合わせた手法である．両手法の認識器を段階的に動作させて，図 3 のフローチャートに示す処理を行う．

まず，細分化階層法の判別候補のうち，高信頼度[3]の出力があれば優先して用いる．この出力により，統合型階層法の出力が誤っている場合でも，正解が得られることを期待する．次に，低信頼度の出力は細分化階層法において雑談分別率の低下の原因となっているため，雑談分別率に優れる統合型階層法の出力と一致した場合にのみ出力として用いる．この結果，複合型階層法は両手法と比べ，正解率と雑談分別率のバランスがとれた手法となる．

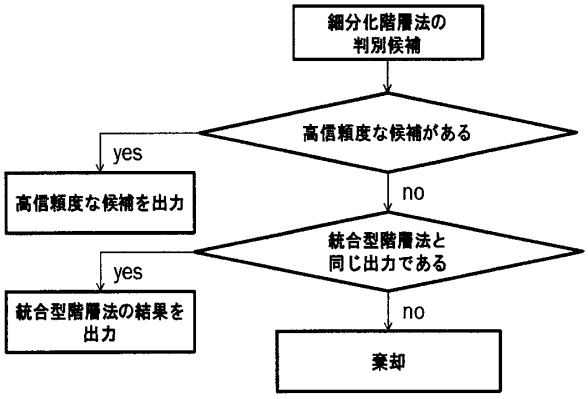

図 3: 複合型階層法における処理のフローチャート

4 比較実験

本節では，本稿で提案した統合型階層法と複合型階層法の有効性を検証するため，従来手法との比較実験を行った．

4.1 収録データ

各手法の比較実験に用いる音声として，男性 4 名と女性 2 名の音声を収録した．各被験者の収録内容は，「拡大」などを含む命令の 50 発話，「2 倍」などを含む値とパラメータの 38 発話，「おはよう」などを含む雑談 60 発話の各 5 回分とした．

[3]本研究では判別時の信頼度は大語彙認識器の認識結果との音素の編集距離による．この編集距離が 0.26 未満であれば命令と分別される候補となる．編集距離が 0.14 未満のものを高信頼度とする．この閾値は 4 節の実験で最適な結果が得られる値である．なお，低信頼度とは編集距離が 0.14 以上で 0.26 未満のものを指す．

4.2 実験結果

階層化を施す前の非階層型手法[4] と，各階層化手法で収録データの音声理解を行った．正解率[5]と雑談分別率を比較したグラフを図 4 に示す．

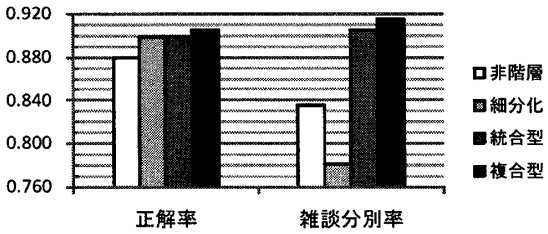

図 4: 命令と雑談のデータセットの結果

図 4 に示したように，従来の細分化型の正解率は，非階層型と比べて 1.8% 向上しているが，雑談分別率は大幅に劣り 5.5% 低下した．この雑談分別率の低下が，従来手法の問題点である．本研究で提案する統合型の正解率も，非階層と比べて 1.8% 向上した．雑談分別率は大幅に優れ 6.8% 向上した．したがって，従来手法の問題点が大幅に改善されたことが分かった．また，統合型と細分化型の正解率は共に 89.8% であり，懸念されていた統合型の正解率の低下はなかった．

階層化型と統合型を組み合わせた複合型は統合型を正解率で 0.6%，雑談分別率で 1.0% 上回り，最も良い結果となった．各手法の出力を詳細に分析した結果，細分化型で高信頼度な結果が得られるケースは全命令入力 1500 件中 1173 件あり，そのうち 11 件は統合型では誤った結果が出力されていたことが分かった．このことから，複合型階層法において正解率が統合型を上回った理由は，細分化型の高信頼度な候補の利用によるものであることが分かった．また，一部の雑談の入力が細分化型の低信頼度な出力と統合型の出力の不一致によって棄却されたため，雑談分別率が統合型を上回っていた．

5 おわりに

本稿では，音声理解における階層化を用いた従来手法の問題点を改善する目的で，新たな階層化手法を提案した．従来手法との比較実験の結果，提案手法は音声理解の精度向上に繋がることが分かった．

今後の課題として，階層構造の自動生成手法の検討や，予期していない入力順序の発話への対応などが挙げられる．さらに，信頼度がある程度高い判別候補が複数得られた場合は，話者に正解の入力文を聞き返すというようなインタフェースへの応用なども挙げられる．

参考文献

[1] Kazutaka Shimada, et al. , “An Effective Speech Understanding Method with a Multiple Speech Recognizer based on Output Selection using Edit Distance”, Proceedings of the PACLIC22, pp.341-349, 2008.

[2] 横山貴彦，嶋田和孝，遠藤 勉，“複数の認識器を選択的に利用する音声理解手法のマルチモーダルインタフェースへの適用”，第 17 回電子情報通信学会九州支部学生会，2009.

[4]全ての要求発話を認識できる 1 つの認識器のみを用いた手法．
[5]全て命令入力のうち，最終的な出力が正解であるものの割合．