D-32

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

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開発が広まり，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は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，より高精度の音声認識が要求され，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．ここで，認識器を分割してさらに小語彙な複数の認識器を作り，選択的に利用することで認識精度の向上を図る手法が考えられる．このとき，認識器に階層構造をもたせることで，発話の流れに応じてアクティブな認識器を切り替えて，語彙数の削減を図る．

本研究では，1 つの語彙辞書に基づく認識器による音声理解と，それを分割して階層化させた複数の認識器を用いた音声理解において，その成功率を比較する実験を行い，その手法の効果を検証する．

2　提案手法の適用

提案手法をファイル操作，編集，検索機能を有する写真管理アプリケーション [1] の音声入力部に適用する．このアプリケーションの認識器の語彙としては約 30 種類の命令，値として最大 4 桁の単位付き数値や日付，検索用の 300 種のタグなどが存在する．従来では認識器はこの一つであるが，新たに認識器を音声入力手順に踏まえて分割し，図 1 の階層構造で音声入力時にアクティブな認識器を切り替えるようにする．この結果，複数の認識器を用いるので複数の音声認識結果が得られる．このとき，音素を判別材料としてアクティブな認識器の中から正解と思しき認識結果を持つ認識器を判別し，最終結果を得る [2]．この結果と入力音声が等しければ，音声理解の成功とする．

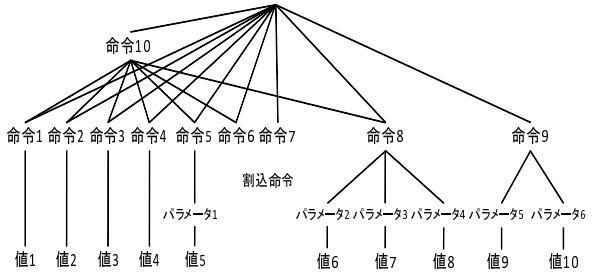

図 1: 提案手法における認識器の階層構造

3　比較実験

3.1　収録データ

男性 4 名と女性 2 名の音声データを収録した．各被験者の収録内容は，拡大や縮小などの語を含む命令の 50 発話と，2 倍などの語を含む値とパラメータの 38 発話の各 5 回分で，計 440 発話である．

3.2　実験結果

収録データの音声理解精度を単一の認識器を用いた従来手法と，複数の認識器を選択的に用いた提案手法で調べた．これら 2 つの手法における被験者毎の音声理解精度のグラフを図 2 に示す．図 2 の結果では提案手法の理解精度が従来手法を上回る傾向が見られ，その有効性が確認できた．

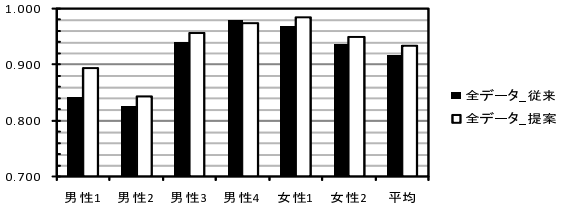

図 2: 全収録データにおける音声理解精度

また，階層化によって認識器の管理面で利便性が高まり，別の階層ならば語彙の音素が同じでも判別可能なので，よりユニークな辞書登録が可能になるという利点が提案手法にはあった．ただし，階層化によって手順を踏んだ発話を行うため，適用させるシステムによっては発話の柔軟性が失われる欠点もあると考えられる．

4　考察

収録データの発話毎に音声理解精度を分析した．その結果，命令の発話群において理解精度の変動がほとんど見られなかったのに対し，値とパラメータの発話群においては以下の図 3 のように著しい結果が見られた．

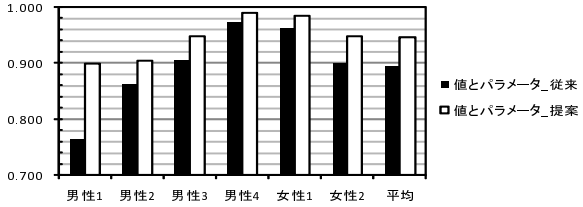

図 3: 値とパラメータの発話群における音声理解精度

このような結果が得られた理由は，値とパラメータの発話群の階層では並列に動作する認識器が少なかったため誤判別がほとんど無く，階層化の語彙数の削減による誤認識の低減効果を直接的に受けられたからであると考えられる．従って，階層構造の中で並列する認識器が多くなるような階層に対して調整を加えることで，より高い効果が期待できると考えられる．

5　おわりに

本研究では，音声理解において認識器を分割し，階層化して利用する手法の有効性を比較実験を行って検証した．その結果，提案手法は音声理解精度の向上に繋がることが確認できた．

参考文献

[1] 武藤 亮介，嶋田 和孝，遠藤 勉，“USB カメラとマーカを用いたハンドジェスチャ認識”，第 16 回電子情報通信学会九州支部学生会，D-51，2008.

[2] 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.