ユーザの大人・子供を識別する音声情報案内システム

西村 竜一, 中村 敬介, 西原 洋平, 李 晃伸, 猿渡 洋, 鹿野 清宏
奈良先端科学技術大学院大学 情報科学研究科

1 はじめに

ロボットや情報家電等のインタフェースとして，我々の日常生活に音声インタフェースが導入されつつある．本稿では音声インタフェースにおける利用者の年齢層に注目する．システムの家庭等への普及を考えると，今後，子供の存在は無視できない．しかし，従来の音声認識では大人発話から構築した統計モデルが利用されることが多く，子供発話の認識性能に不足が指摘されている [1]．これからは大人のみならず，子供の利用者も意識した改良が求められる．また，大人と子供では発話の表現（使用単語や語尾様式等）や興味対象が異なるため，利用者が大人または子供のどちらかを識別した上で応答した方が，より利用者に適した対話を実現できるだろう．

本研究では，利用者を大人・子供に識別して，その年齢層に順応した応答を生成する音声インタフェースの開発を試みた．主に子供に対する利便性向上を目的とする．

2 大人・子供識別能力を備えた音声インタフェース

開発するシステムは，著者らが開発した生駒市北コミュニティセンターの音声情報案内システム「たけまるくん」[2]（図 1）を改良したものである．図 2 にシステム構成を示す．以下に挙げるのが主な変更点である．

- 大人・子供別の音響モデルと言語モデルを用いた並列音声認識
- 入力音声からの話者の大人・子供自動識別
- 識別結果を考慮した応答生成

並列音声認識は，大人・子供用のモデルを持つ音声認識エンジン Julius[3] を 2 個独立にサーバモードで実行することで実装する．各 Julius には，録音プログラム（adintool）から TCP/IP 経由で入力音声の特徴量が渡され，結果として 2 つの認識出力のテキストを得る．ここで用いる大人・子供別の音響モデルと言語モデルは，2002 年 11 月から継続中のたけまるくんのフィールドテストで収集した利用者の発話から構築した [1]．

続いて，システムは録音音声から話者が大人か子供かを識別し，その結果によって 2 つの認識出力の一方を選択する．話者の年齢層識別手法は本稿で後述する．

たけまるくんの一問一答形式の音声インタフェースでは，あらかじめ用意した応答候補テキストの中から適当なものを一つ選ぶことで応答を生成する．この応答生成過程においても，先ほどの話者年齢層識別結果を利用する．これまでに大人と子供別の応答候補を作成した．大人向けと子供向けに特化した応答内容により，柔軟な情報案内を可能にする．例えば，「タバコを吸うところはありますか?」という問いに対して，利用者が大人なら「喫煙コーナーは，この壁の裏側にございます.」と場所案内を行うが，子供に対しては「タバコなんて吸ったら駄目.」と応答することができる．また，応答の選択基準には，利用者の発話ログを集めた用例のテキストと音

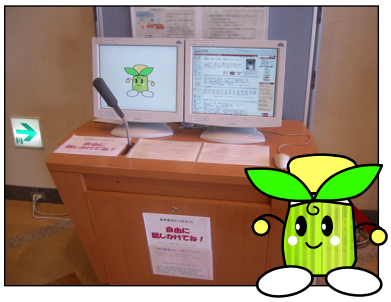

図 1: 音声情報案内システム「たけまるくん」

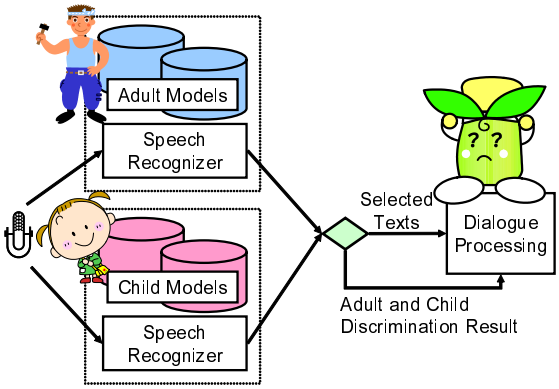

図 2: 大人・子供識別能力を備えた音声インタフェース

声認識結果の形態素一致率を用いるが [2]，その算出の際に識別結果を加味することも検討中である．

3 音声認識スコアに基づく話者年齢層識別手法

話者識別には，音声が持つ音響的特徴をモデル化した GMM（Gaussian Mixture Model）に基づく尤度比較法が広く用いられる．しかし，自由発話では大人と子供で発話の内容に異なる傾向を持つため，言語的特徴の考慮で識別精度の向上が見込まれる．そこで，音声認識スコアから導出される音響的特徴と言語的特徴を併用する話者識別法を提案する．本手法では，音響的特徴（AP）として次式で算出するフレーム平均音響対数尤度，言語的特徴（LP）として単語平均言語対数尤度をパラメータとした機械学習によって識別を行う．

$$AP = \frac{\text{音響対数尤度}}{\text{入力音声のフレーム数}} \quad (1)$$

$$LP = \frac{\text{言語対数尤度}}{\text{出力単語列の単語数}} \quad (2)$$

また，音声認識のモデルや収録系に変更があってもその違いを吸収できるように，大人用の音響・言語モデルを用いた際の認識結果から求めた特徴（AP_{adult}，LP_{adult}）から子供モデル使用時の認識結果による特徴（AP_{child}，LP_{child}）を引いた差をパラメータにする．

音声認識に用いた音響モデルは PTM triphone の性別非依存 HMM モデル，言語モデルは単語数 4 万の 3-gram モデルであり，収集発話から学習した大人・子供別のものである．人の主観によって収録発話を分類（子供:16,892 文，大人:7,606 文），各々から大人用と子供用のモデルを学習した．

Spoken Guidance System with Discrimination of Adult and Child Speech. Ryuichi NISIMURA, Keisuke NAKAMURA, Yohei NISHIHARA, Akinobu LEE, Hiroshi SARUWATARI, Kiyohiro SHIKANO (Graduate School of Information Science, Nara Institute of Science and Technology)

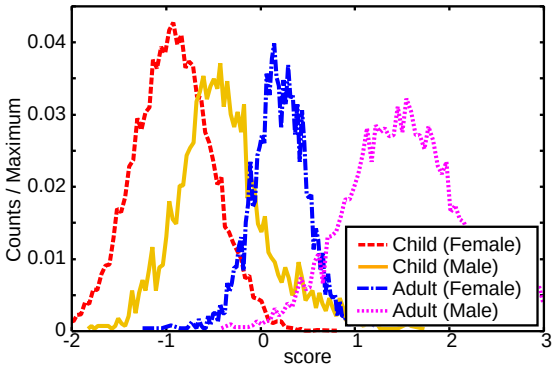

図 3: $AP_{adult} - AP_{child}$ の頻度分布

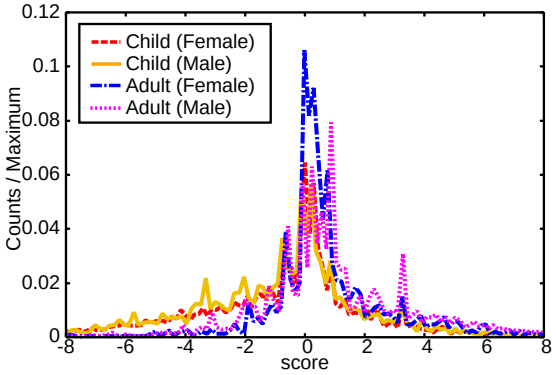

図 4: $LP_{adult} - LP_{child}$ の頻度分布

図 3 は，音響的特徴 $AP_{adult} - AP_{child}$ の頻度分布である．各グラフは，収集発話のうち，子供女性（10,135 個），子供男性（2,831 個），大人女性（2,385 個），大人男性（5,151 個）の発話を認識した際のものである．右に分布するほど大人の発話に特徴が近いことを示す．この結果，大人男性の分布に関しては明確に区別できるが，大人女性は子供に特徴が比較的近いことがわかる．

同様に言語的特徴 $LP_{adult} - LP_{child}$ の頻度分布を図 4 に示す．音響的特徴ほどには大人と子供で分布傾向の違いは確認できない．しかし，若干ではあるが，分布の中心以外で，大人と子供の分布カーブに特徴的な違いが男女に共通して見られ，先の音響的特徴と組合わせた機械学習による識別精度向上が見込まれる．

機械学習のアルゴリズムには，SVM（Support Vector Machine）[4] を用いた．SVM は二値分類器であり，AP と LP の構成ベクトルをパラメータとして大人・子供識別を行う．SVM のカーネル関数には予備実験で最も良い結果を得た Gaussian 関数を用いた．

4 評価実験

評価のテストセットには大人 500 個，子供 500 個のフィールドテスト収集発話を用いた．これらは音声認識の音響・言語モデルや SVM の学習からは除外している．

表 1 にテストセットに対する大人・子供識別率を示す．この結果，提案手法は 94.6%の識別率を得ることができた．機械学習から言語的特徴を除外すると 1.5%の低下が生じ，音響的特徴と言語的特徴の併用の有効性を確認できる．パラメータに大人・子供用モデル間の差ではなく実測値を用いると若干の精度低下を起こした．

比較のために GMM を使った尤度比較法でも識別を試みた．GMM に使用した音声分析パラメータは，音声

表 1: 大人・子供識別率 [%]

パラメータ	大人	子供	合計
音響的特徴のみ 1. $AP_{adult} - AP_{child}$	95.0	91.2	93.1
音響的特徴と言語的特徴 1. $AP_{adult} - AP_{child}$ 2. $LP_{adult} - LP_{child}$	96.8	92.4	94.6
音響的特徴と言語的特徴 （実測値を使用） 1. AP_{adult} 2.AP_{child} 3. LP_{adult} 4.LP_{child}	92.4	92.0	92.2
GMM 尤度比較法（ベースライン）			86.4

表 2: 単語正解精度 [%]

選択方法	大人	子供	合計
SVM で選択	94.3	76.9	86.2
正解を選択	94.2	77.0	86.2
単一モデル認識	92.0	75.9	84.5

認識のものと同じ 16bit，16kHz 音声を窓シフト長 10ms で分析した 12 次元の MFCC と Δ MFCC，Δ Power であり，正規分布の混合数は 64 である [1]．得られた識別率は 86.4%であり，SVM を用いた提案手法の識別率はこれより 8.2%高かった．

次に，Julius による並列音声認識出力に識別結果を反映して音声認識率を求めた．表 2 に単語正解精度を示す．表中の選択方法の “SVM で選択” が，2 つの認識出力の一方を自動選択する提案法である．“正解を選択” は，人の主観により話者の年齢層を判断して認識出力を選択した結果である．また，“単一モデル認識” は，並列音声認識はせず，大人と子供の両方の発話から年齢層非依存の音響・言語モデルを構築して音声認識した時のものである．この結果から提案法は年齢層非依存の単一モデルを用いる場合よりも高い認識精度を得ることができ，人の主観によって年齢層を選択した場合と比べても同程度の性能を示すことを確認した．

5 まとめ

本稿では，利用者の年齢層に順応した応答生成を実現する音声インタフェースについて検討した．その実装の核となる話者の大人・子供識別法を提案し，評価実験でその有用性を確認した．今後は実装したシステムのフィールドテストを通じた評価を予定している．

謝辞 たけまるくんフィールドテストでの生駒市職員のみなさまの協力に深く感謝します．

参考文献

[1] 西村他: “大人・子供に適応した音声情報案内のためのユーザ自動識別”, 信学技法, SP2003-129/NLC2003-66, 2003
[2] 西村他: “生駒市コミュニティセンター音声情報案内システムの開発と運用”, 情処研報, 2003-SLP-45-6, 2003
[3] A. Lee et al.: “Julius - An Open Source Real-Time Large Vocabulary Recognition Engine”, *EUROSPEECH2001*, pp.1691–1694, 2001
[4] V.N. Vapnik: *The Nature of Statistical Learning Theory*, Springer, 1995