RJ-005

対話型音声インタフェースのための大人・子ども判別技術の改良

An improvement of an adult and child identification method for spoken dialog systems

宮森 翔子 † **西村 竜一** † **入野 俊夫** † **河原 英紀** †
Shoko Miyamori Ryuichi Nisimura Toshio Irino Hideki Kawahara

1 はじめに

本稿では，従来から検討を進めている音声認識の応用による大人・子ども判別技術の改良について述べる．ICT システムが社会基盤として普及するに従い，利用者の年齢を確認する技術に対するニーズは高まっている．例えば，たばこや酒，切符の自動販売機，他にも青少年に対し悪影響を及ぼす可能性があるウェブサイトをフィルタリングする技術などが挙げられる．また，これからの普及が予期されるロボット等の対話型音声インタフェースにおいても，要素技術として年齢確認は有効である．具体的には，利用者の年齢層に応じてシステムの反応を切り替えることで，より柔軟で親切な対話処理を実現できると考えられている．

ユーザに過度な負担を与えることなく年齢確認を実現するには，生体情報を入力とする認識技術の応用が有効である．その一つの例として，2007 年には，顔認識で成人判別を行う機能を備えたたばこの自動販売機が登場し，話題となった [1]．しかし，現実の実用化例では，子どもを大人と誤認識することを防ぐため，システムが大人とみなす年齢は高く設定されているなど，10 代から 20 歳前後を境界とした大人・子ども判別の技術はまだ確立されていない．

生体情報として顔画像や体型，動作パターン等，様々な信号を利用することが考えられるが，本研究では，その中でも音声信号に着目した．自然な対話の中で生じる会話音声を入力とすることで，利用者に負担を与えずに，自然な会話を通じて判断することが可能になると考えている．音声による年齢推定に，和田ら [2] の研究がある．ただし，学会などの講演音声を対象とした推定を行っており，子ども話者を対象とした検討としては不十分である．そこで，本研究では多量に集めた子ども発話を用いて，子ども利用者に特に注目し実験を行った．また，google 音声認識などで注目を集めている音声ウェブインタフェースでの利用を想定した実環境データを用いた検討を行った．

2 音声認識を転用した大人・子ども判別手法

本研究では，既存の音声認識システムのアルゴリズムを一部転用することで，発話による大人・子ども判別を実現することを目指している．現在の音声認識は，音響的な特徴を統計化した音響モデルとして HMM（Hidden Markov Model）を用いることが一般的となっている．我々は以前の研究で，HMM のみを用いた識別アルゴリズムによる大人・子ども判別を行った [3]．ここでは，音声認識で用いる音素クラスを，大人と子どもの発話を集合とする 2 つのクラスに置き換え，音声認識アルゴリズムを 2 クラス識別に転用した．まず，大人と子どもの二つのクラスに分けた発話に周波数分析等を適用し，音響特徴量を得る．その特徴量ベクトルを混合ガウス分布を状態とする 3 状態 HMM としてモデル化し，入力に対する各クラスの尤度を比較して，識別結果を得た．その結果として，年齢閾値 16 歳以上においても，70% を超える子ども識別を得ることに成功した．しかし，この方法では，変声期に当たる 10 代若年者全体の識別精度が低く，本研究の目的から考えて実用的では無かった．これは二つのクラス間の単純な尤度比較のみで識別結果を得ていたことに一つの原因があると考えた．

そこで我々は，HMM に加えて SVM(Support Vector Machine) を用いる 2 層アルゴリズムを開発した [4]．HMM のクラスを，これまで用いた 2 から 24 に増やすことで，HMM が出力する尤度を 24 個に増やした．ここで得られた尤度を単純に大小比較するのではなく，24 次元の素性とみなすことで SVM による 2 クラス識別を実施し，発話者を大人・子どもに判別した．

2.1 HMM+SVM の 2 層アルゴリズムによる大人・子ども判別手法

本節では，HMM+SVM の 2 層アルゴリズムによる大人・子ども判別手法を処理の順を追って説明する．また，図 1 に本手法の概要を示す．

[第 1 層] 本手法は，1 層目で HMM の構築と尤度の算出を行い，2 層目で用いる素性を構成する．

1. 発話者の年齢と性別に基づき，発話を男女各 5 歳ごとの 24 クラス (女:0〜5 歳, 男:0〜5 歳, 女:6〜10 歳, 男:6〜10 歳, $\cdots$, 女:56〜60 歳, 男:56〜60 歳) に分類
2. 発話データに周波数分析等を適用し，音響特徴量を抽出
3. 抽出した音響特徴量から 3 状態の混合ガウス分布からなる HMM（24 クラス）を構築
4. すべての発話を用いて 24 クラス HMM に対する尤度 ($AM_c, c = 1, 2, \cdots, 24$) を算出
5. 尤度 (AM_c) を下式を用いて発話のフレーム数で正規化

$$\hat{AP}_c = \frac{AM_c}{\#of\ input\ speech\ frames} \qquad (1)$$

6. 正規化音響尤度 (AP_c) と HMM のクラス番号をセットにした SVM の素性を構成

[第 2 層] 2 層目では，1 層目で作成した素性を入力にし

† 和歌山大学，Wakayama University

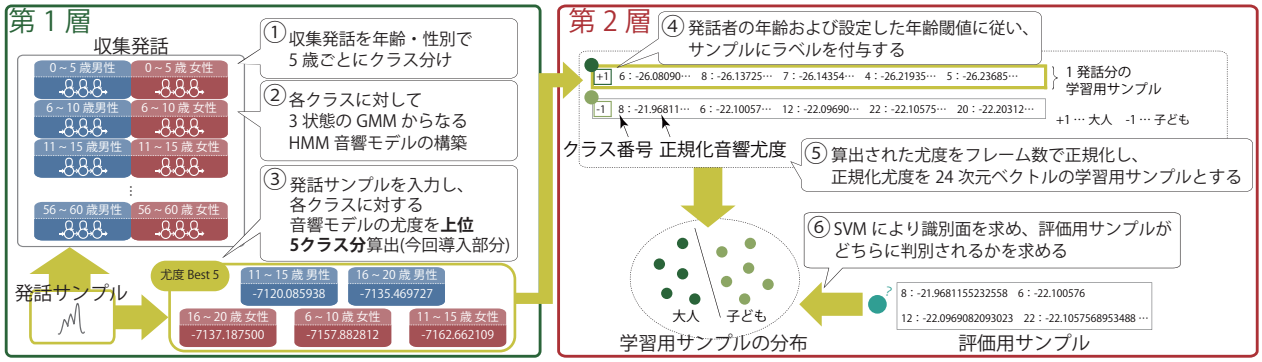

図 1　提案手法の概要

た SVM による 2 クラス識別を行う．

1. 1 層目で得た素性を組み合わせ，SVM の入力データを作成
2. 入力データに対し，発話者の年齢と年齢閾値に基づいた 2 クラスラベル（年齢閾値以上 (大人) をポジティブ，年齢閾値未満 (子ども) をネガティブ）を付与
3. 作成した入力データとラベルに基づき SVM による 2 クラス識別を適用

2.2　尤度上位に基づく素性選択の導入

これまでに，我々は，24 クラス HMM から算出した 24 個の尤度すべてを SVM の素性としてきた [4]．しかし，24 クラスのうち，値の小さい下位クラスの尤度は，その算出された値そのものの信頼性が低いため，SVM の素性からは除外することが適当であると考えた．このため，本稿において，第 1 層と第 2 層の間に，HMM で求めた尤度による素性選択を導入することにした．具体的には，24 クラスの尤度のうち，値が高い上位 5 クラス分のみを選択し，SVM の素性（5 次元）とする．

なお，以下では，素性選択を適用したものを提案手法，素性選択をせず 24 次元すべて利用する方法を，比較のための従来手法として議論を進める．

3　評価実験

本研究では，音声ウェブシステムによって集めた実環境発話を用いて，提案手法及び従来手法を評価した．

3.1　実験に用いた実環境発話

大人・子ども判別の技術を応用する際，家庭の利用がまず最初に想定されるため，家庭において PC に向かって発声した音声を使って実験を行うことが望ましい．本当の家庭環境で PC に向かって発声した音声を集めた事例はこれまで限られており，特に，子どもによる発話を含んだデータベースが整備されたことは皆無である．そこで，本研究では，音声ウェブシステム w3voice [5][*1]を用いて，インターネットを介した発話の収集を試みた．音声ウェブシステム w3voice は，利用者によって録音された音声をサーバに自動送信する機構を持つため，サーバ上で音声を収集することができる．発話者は楽天リサーチ社に依頼して募集した．

[*1] http://w3voice.jp/

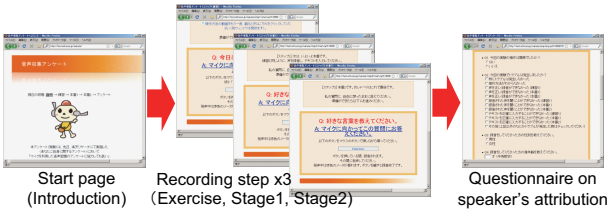

図 2　発話者に提示する音声収集ウェブサイトの全体構成

図 2 に音声収集用に構築したウェブサイトの構成を示す．このウェブサイトでは，利用者が発話を行う過程として，「練習」「本番 1」「本番 2」の 3 つのステップが用意されており，各ステップには簡単な設問が用意されている．本番 1，本番 2 の設問内容を以下に示す．

- 本番 1:好きな食べ物は何ですか．
- 本番 2:好きな言葉を教えてください．

発話者は各ステップにおいて設問への解答を発話し，録音する．全ての録音ステップが完了した後，発話者は，自身の属性および使用した機材に関するアンケートに回答する．なお，低年齢の発話者には保護者が付き添い操作を行うように要請した．

この収集の結果，用意した収集用ウェブサイトに，ユニーク IP アドレスで 5,778 のアクセスを得た．そのうち，3 つの録音ステップを完遂した発話者は 1,152 名であり，回答率は 19.9% であった．収集された発話の中には無効な録音データやアンケートの入力ミスなどが含まれるため，大学生 2 名が人手で内容を確認した [6]．その結果，発話者 1,050 名分の 3,053 発話が有効であった．(1,037 ユニーク IP アドレス)

図 3 に，収集した発話の発話者の年齢と身長の散布図を示す．図の赤点は女性の発話者，青点は男性の発話者を示す．全ての発話サンプルのうち，15 歳以下の子どもの発話サンプルは 1,533 発話であり，全体の 59.7% を占めた．

図 4 に，2 歳から 19 歳までの発話数を示す．図より，10 歳未満の発話者に対して，10 代の発話数が少ないことがわかる．特に，15 歳によって発話されたサンプルは 26 発話であった．よって，10 代の発話を追加収集する予定ではあるが，今のところ，収集発話における発話者の年齢には偏りが存在する．

ここから述べる判別実験の結果は，ここで収集した有効発話のうちの 2,359 発話を用いた結果である．ウェブベースのインタフェースを用いて収集した発話のデータベースにより，家庭等の実環境を反映した条件

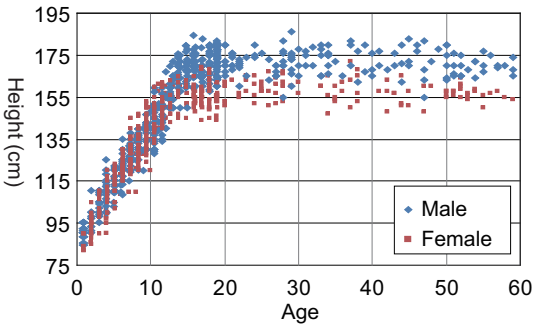

図 3　発話者の身長・年齢分布

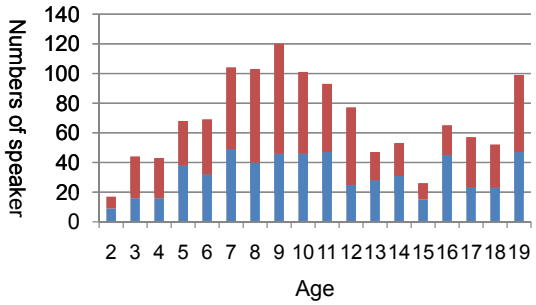

図 4　2〜19 歳の発話者の分布

での実験を実現することができた．

加えて，検定方法に 10 分割交差検定法を用いて，HMM や SVM の学習段階で，評価に用いる被験者発話が含まれない条件 (話者オープン) を構築し，極めて実際の利用に近い環境を再現している．なお，判別実験に使用した発話や検定方法は，先行研究と同様である．

3.2　実験条件

提案手法の第 1 層において，HMM の学習に用いた音響特徴量は，12 次元の MFCC, Δ MFCC, Δ Power であり，これは一般的な音声認識に用いるパラメタと同様である．HMM の各状態の混合ガウス分布混合数は 64（提案手法）, 128（従来手法）とした．また，HMM の構築には，音声認識の音響モデルを構築する際に広く用いられるツールである HTK3.4.1 [7] を用いた．

第 2 層の SVM の実装には，提案手法では SVM-Light6.02[8], 従来手法では TinySVM0.09[9] を用いた．カーネル関数には，3 次の多項式カーネル（提案手法），ANOVA カーネル（従来手法）を用いた．

なお，各実験条件は，複数の条件において評価実験を実施した中で，最も良い結果が得られたものを選んだ．

4　評価実験結果

本節では，評価実験の結果を述べる．なお，本研究では，大人と子どもの境界となる年齢を示す値として年齢閾値を定義して用いる．例えば，年齢閾値 15 歳の場合は，15 歳未満の発話者を子ども，15 歳以上の発話者を大人とみなすことになる．以下で述べる実験では，年齢閾値を 9 歳から 20 歳まで 1 歳ごとに変化させて，年齢閾値の変化によって生じる精度の違いを検討する．

4.1　提案手法における大人・子ども判別結果の比較

今回，正解率と F 値の 2 つの評価尺度を用いた評価結果を示す．F 値は，情報検索システムの性能を表わす

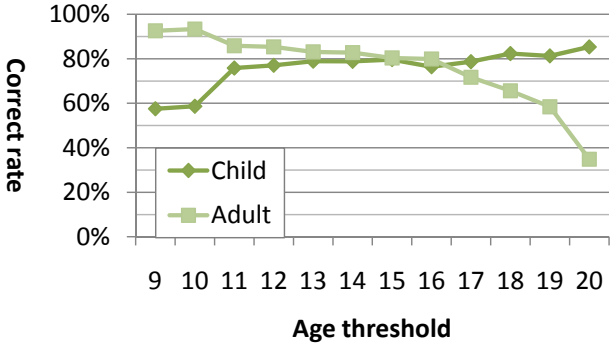

図 5　評価実験結果（正解率）

総合的な評価尺度であり，適合率及び再現率（正解率）の調和平均で求めることができる．

提案手法による大人・子ども判別の結果として，図 5 に正解率，図 6 に F 値の比較を示す．子ども判別の正解率は，11 歳以上の年齢閾値において，常に 60% 以上を示した．大人を判別した場合においても，年齢閾値 17 歳まで 60% 以上の正解率を示す結果が得られた．F 値は，年齢閾値 11 歳から 17 歳までの区間では，0.7 以上を示す結果となった．

提案手法では, 年齢閾値 13 歳のとき, 正解率 78.9% を得ることができた．これは，我々の先行研究である HMM のみを用いた場合より，10.2 ポイントの精度向上である．HMM の方法では，年齢閾値 13 歳のときに最も良い結果を得られたので，HMM と SVM を組み合わせた本手法は，大人・子ども判別における年齢閾値を向上させることに有効であることがわかる．

4.2　提案手法と従来手法の比較

図 7 において，提案手法と従来手法の正解率を比較する．図中の青線は従来手法を，赤線は提案手法の正解率である．各色のうち，濃色の線は子ども判別，淡色の線は大人判別の結果を示す．

子どもを判別した結果では，両者の違いに大きな差は見られなかった．しかし，大人判別では，10 代後半以降の年齢閾値において差が見られた．従来手法は，年齢閾値 16 歳以上のおいて，大人判別の性能に著しい低下が生じていた．一方，提案手法では，年齢閾値 19 歳における大人判別についても，58.4% の正解率を得ることに成功している．この結果から，提案手法によって，子ども判別の性能を高く維持したまま，大人判別の一定の性能を獲得できるようになったと言える．

つまり，24 クラスすべての尤度を素性としてた従来手法では，低い尤度を持つ素性が SVM の識別能力に悪影響を与えていた．下位の尤度を持つクラス間では，統計的に有意な差を持った尤度の分布が得られていなかったことが考えられる．その結果, 同じクラス内での個人差が，学習データの分布の広がりに直接反映され，SVM を正しく学習できなかった可能性がある．高い尤度のみを抽出した提案手法では，個人差の影響を軽減でき，精度の向上に繋がったのではないかと考える．

5　1 歳単位 HMM による 116 次元素性の利用

ここまでで 24 クラス HMM を用いることで，大人・子ども識別性能の向上を確認することができた．本稿

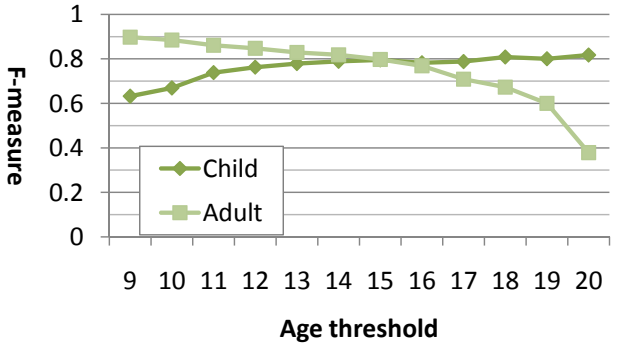

図 6　評価実験結果（F 値）

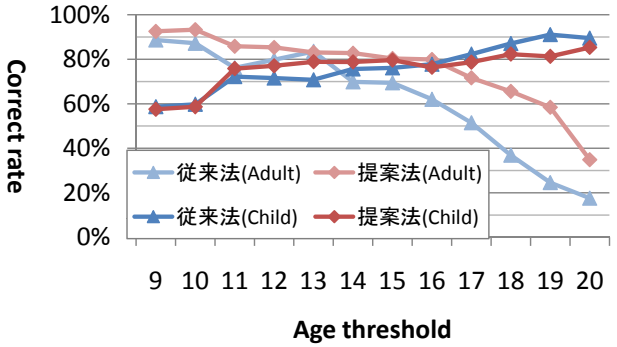

図 7　提案法と従来法の比較（正解率）

では，さらにクラスを増やした場合の検討を行った．ここでは，HMM を 1 歳単位のクラスで構築することを考える．使用した発話の発話者は，2〜59 歳であるため，1 歳単位かつ男女別にすると合計で 116 クラスの HMM を構築することができる．つまり，この HMM で尤度を算出すると，最大で 116 次元の素性を得て SVM を学習することが可能になる．ただし，HMM の学習は，各クラスごと（1 歳ごと）に十分な量の発話データが必要となる．例えば，35 歳男性の発話は，今回用いた収集発話には，3 個しか含まれておらず，学習データの数に不足が生じる．今回，HMM の学習に前後 2 歳の話者の発話をオーバラップさせることで，この問題に対処した．HMM は 1 歳単位で構築する一方で，その学習には前後 2 歳の発話者による発話データも含めることにした．これにより，HMM の学習に用いるデータを，見かけ上，増やすことができた．

実験に用いた条件は，前述の提案手法のものと同様である．ただし，HMM のクラス数は，116 に増加している．前述したように，尤度に基づき第 1 層で出力した上位 5 クラス分の素性のみを選択し，第 2 層の SVM で入力として用いた．

実験の結果を図 8 に示す．この手法でも，従来手法と同様に年齢閾値の上昇に伴い大人正解率が低下している．今回，20 歳以上のクラスでは発話数の少ないクラスが生じている．例えば，50 歳女性のクラスには，11 名分の 27 発話しか発話が存在しない．このようなクラスにおいて構築した HMM が，発話者の個人性を強く反映してしまった可能性がある．

6　まとめ

本研究では，ユーザに親切な対応を行う対話インタフェースのための，大人・子ども判別を提案した．従来

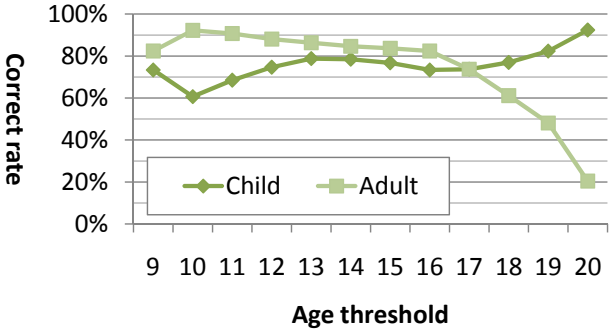

図 8　1 歳単位 HMM の上位 5 尤度を用いた判別 結果

法では，年齢閾値の上昇に伴い，大人判別の正解率が低下するといった問題があった．そこで，24 クラス中上位 5 クラスの音響尤度のみを抽出し，SVM の特徴量とする判別手法を提案し，24 クラス全ての音響尤度を算出する従来手法との比較実験を行った．従来手法では，高い年齢閾値における大人判別に問題があったが，つまり，グラフの数は学習に利用できるデータ数に応じて適切に設定する必要がある．大人の判別性能を維持したまま，十代以降の子ども判別性能を上げることができた．

今後は，自動判別の性能をより向上させるために，発話の言語情報を組み込んだ判別法 [10] を導入する予定である．並行して，発話の収集および分析を進める．

謝辞 本研究の一部は科学研究費補助金の支援を受けた．

参考文献

[1] 株式会社フジタカ, ”成人識別装置「こどもチェックシステム」”, http://www.fujitaka.com/kao/, 2007.
[2] Toshiya Wada, et al., ”Investigations of Features and Estimators for Speech-based Age Estimation ”, Proc. APSIPA, 2010.
[3] 宮森 他, ”ちょっとした一言の音声認識による子ども利用者判別法の検討,” FIT2010 第 9 回情報科学技術フォーラム, pp.469-472, 2010.
[4] 宮森 他, ”実環境発話を用いた子ども判別アルゴリズムの検討”, 日本音響学会：春季研究発表会講演論文集, pp.55-56, 2011.
[5] 西村 他, ”音声入力・認識機能を有する Web システム w3voice の開発と運用”, 情報処理学会研究報告, 2007-SLP-68-3, 2007.
[6] 栗原 他, ”音声ウェブシステムを用いて収集した実環境子供発話に関する調査”, FIT2010 第 9 回情報科学技術フォーラム講演論文集, 2010.
[7] Young, S.J., et al., ”The HTK book version 3.4”, Cambridge University Engineering Department, Cambridge, UK, 2006.
[8] http://svmlight.joachims.org/
[9] http://chasen.org/ taku/software/TinySVM/
[10] Nisimura, R., et al., ”Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability”, Proc. ICASSP2004, Vol.I, pp.433–436, 2004.