Vol.2009-SLP-77 No.19
2009/7/18

安心ウェブの実現に向けた 大人・子ども発話のネット収集実験

西村竜一[†1] 宮森翔子[†1] 鈴田健太郎[†1]
河原英紀[†1] 入野俊夫[†1]

本研究では，利用者の年齢層を発話音声から自動推定し，子どものアクセスを制限するウェブフィルタリングサービスの開発を目指す．今回，提案システムの実現に向けて，(1) 音声ウェブシステム w3voice を用いた大人・子ども発話のネットワーク収集実験，(2) GMM 音響モデルを用いた若年者自動判別の予備的実験を行った．発話収集の実験では，389 名の被験者の実環境発話 1,109 を集めることに成功した．発話を分析した結果，大人と子どもで，発話内容に異なる言語的傾向があることを確認した．また，GMM 音響モデルを用いた 14 歳以下の子どもの検出実験では正解率 65.9%を得た（大人の検出も含めると正解率 82.6%）．

Web-based adult and child voice collection to develop a voice-oriented web filtering service

Ryuichi Nisimura,[†1] Shoko Miyamori,[†1]
Kentaro Suzuta,[†1] Hideki Kawahara[†1]
and Toshio Irino[†1]

This study aims at developing a voice-based web filtering service to restrict children from the harmful websites. It is based on an automatic estimation of an age group from their voices. To realize it, we have performed (1) a collection of adult and child voices using voice-enabled web system "w3voice", and (2) an experiment of young voice detection on the basis of GMM-based acoustic recognition. In the experiment of the utterance collection, we succeeded in the collection of the 389 testees' real environmental 1,109 utterances. It was confirmed that there was the difference of language tendencies between adults and children as a result of analyzing the utterances. In the experiment on 14-years-old or younger child detection, 65.9% correct rate was obtained.

1. はじめに

インターネットの普及により，近年では，若年者のアンダーグランドなウェブサイト（いわゆるアングラサイト）の利用が問題になっている．特に，ネットワークを離れた実社会でのいじめの原因となる「学校裏サイト」の存在は，今や社会問題である．家庭でインターネットを使う子どもたちを，危険なアングラサイトから隔離することが求められている．その問題に対し，ブラックリストに掲載されたサイトへのアクセスを遮断するウェブフィルタリングの導入が進んでいる．

そのようなウェブフィルタリングには，アクセス時に利用者の年齢確認を行うものがある．つまり，設定された年齢以上の利用者には，自由にすべてのサイトへのアクセスを許可する．一方，設定年齢に満たないときにはアクセスを制限する．しかし，利用者の年齢層を自動確認する手法の確立は遅れており，フィルタリングが有効に利用されないケースが発生している．例えば，利用者が年齢を申告（システムに年齢を設定）する際に，自ら積極的に年齢を偽れば，危険なウェブサイトにもアクセスできてしまう．

最近，タバコの自動販売機では，顔画像を用いて年齢層を自動推定し，未成年者の利用を拒絶するシステムの導入が話題になった[⋆1]．今後，さまざまな生体情報を用いて利用者の年齢層を判別する技術は，そのアプリケーションとともに重要度を増すことになる．それに伴って，判別性能の不安定さ等に指摘があり，技術的な完成度の向上が求められている．

本研究で検討するウェブフィルタリングでは，年齢層判別の生体情報として，一般の PC に接続されたマイクで集音した利用者発話を利用する．一般家庭での利用を想定している．しかし，家庭等のプライベートな空間で，PC のマイクを用いて発話を収集した事例は極めて少ない．判別システムを構築するのに必要なデータは不足している．そこで，我々は，インターネットアンケート業者（楽天リサーチ社）の協力下，家庭等でウェブを利用するモニタ被験者の発話を収集し，データの整備を行っている．

本研究と同様に，ウェブを用いて発話を収集した先行研究に原らの実験がある[1)]．しかし，原らの実験は，タスクを音声認識による音楽検索としており複雑であった．タスク達成率が低いことが問題として指摘されている．加えて，本研究の目的を達成するには，ウェブインタフェースの操作や実験協力等に不馴れな子どもの協力を集める必要がある．そのため，実

†1 和歌山大学システム工学部
Faculty of Systems Engineering, Wakayama University
⋆1 http://www.fujitaka.com/kao/index.html

Vol.2009-SLP-77 No.19
2009/7/18

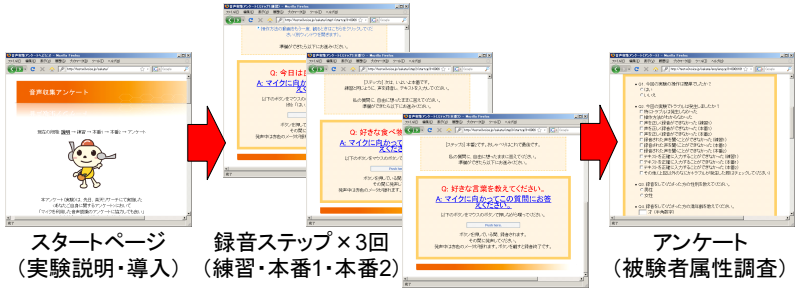

図 1 実験サイトの全体構成
Fig. 1 Overview of the Web site for a voice collection.

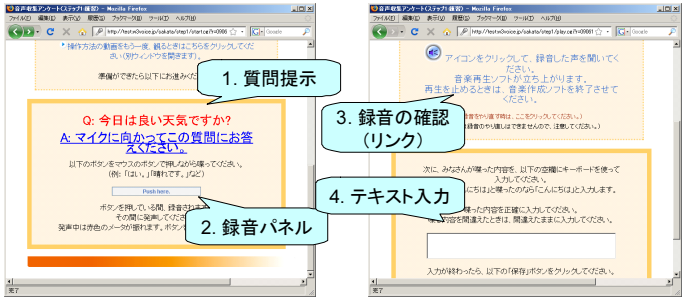

図 2 録音ステップにおけるウェブインタフェースの画面構成
Fig. 2 Snapshot of the Web interface for the voice recording steps.

験内容やインタフェース等のデザインを，子どもの被験者にも対応できるように検討した．

今回，提案システムの実現に向けて，(1) 音声ウェブシステム w3voice を用いた大人・子ども発話のネットワーク収集実験，(2) GMM 音響モデルを用いた若年者自動判別の予備的実験 の二つを行ったので，その結果を報告する．

2. 大人・子ども発話のネットワーク収集

本研究では，発話を収集するため，音声ウェブシステム w3voice[2),3)]⋆1による録音インタフェースを持ったウェブサイトをインターネット上に開設した．作成した実験サイトは，練習 1 つ及び本番 2 つの合計 3 回，被験者に発話の録音を促す録音ステップを持つ構成となっている．また，被験者の属性を調査するために，録音終了後，引き続き，ウェブを用いたアンケートを実施した．**図 1** に実験サイトの全体構成を示す．

実験に際して，ウェブサイトをインターネットで公開するだけでは，被験者の数を集めることは難しい．このため，楽天リサーチ社のモニタ誘因サービスを利用し，大規模に被験者を集めることにした．モニタ登録をしている被験者候補に，同社は，実験サイトの URL が書かれたメールを送付する．それを読んだ被験者が，家庭の PC から実験サイトにアクセスすることで，本実験に参加することができる．すべての録音ステップ及びアンケートを完遂した被験者には，報酬として同社よりポイントが付与される仕組みである．

子どもの発話を集めるために，モニタの事前スクリーニングを同社に依頼し，幅広い年齢層の被験者を集めるように努めた．スクリーニングは，同社の有する大人のモニタに対して行い，一緒に実験に参加できる子どもはいるかを尋ねるようにした．モニタから「実験に協力できる子どもがいる」と回答があった場合，操作は付き添いの大人が行い，発話行為は子どもがするように，メールを通じて説明した上で実験への参加を促した．その結果，子どもによる録音やアンケートの操作のミスを回避でき，子ども発話の有効回答を増やすことができた．

2.1 音声ウェブシステム w3voice

著者らが開発し，フリーソフトウェアとして公開している音声ウェブシステム w3voice は，通常のウェブシステムに音声入力インタフェースを付加するフレームワークである．ユーザのローカル PC 上で動作する Java アプレットと，ウェブサーバ上で動作する CGI プログラムにより構成される．

本研究は，大人・子どもを問わず幅の広い年齢層の発話を収集する必要がある．このため，発話収集の際には，できる限り被験者に煩雑な手間を掛けないようにし，実験に参加するにあたってのハードルを下げることが重要となる．その点，Java が動作する PC からの利用であれば，利用者に特別なソフトウェアのインストールや設定を要求しないことが，w3voice システムの優れた特徴となる．また，Windows や Mac，Linux 等の主要 OS 及び IE や Firefox 等ブラウザの動作環境を選ばずに利用できる利便性においても優れている．

加えて，w3voice システムは，収録発話をウェブサーバ上に一括保存する能力を持ち，発話をコレクションするのに適したプラットホームであると言える．

⋆1 http://w3voice.jp/

2.2 録音ステップにおける実験手順

録音ステップにおけるウェブページの遷移を**図 2** に示す．練習 1 回，本番 2 回の計 3 回ある各録音ステップでは,「(1) 質問文を確認」「(2) 録音」「(3) 録音音声の再生確認」「(4) 発話内容のテキスト入力」の 4 つの操作を被験者に求める．以下に，被験者側の視点から録音ステップの実験手順を説明する．

(1) 質問文を確認

録音ステップに突入した状態のウェブページの例が図 2 の a) である．このページには，被験者に向けた質問文章がテキストで提示されている．被験者は，これを目視することで内容を確認する．掲示する質問の内容については，2.3 で後述する．

(2) 録音

次に，提示された質問に対し，被験者は発話で回答をする．w3voice システムの仕様に従い，ウェブページ上に配置された録音パネルをマウスボタンで長押ししながら発声する．ボタンを離したら録音が終わり，発話は実験サイトのサーバに自動的にアップロードされる．

(3) 録音音声の再生確認

録音終了後，ブラウザに表示されるのが図 2 の b) 画面である．ここには，アップロードデータへのハイパーリンクが設置されている．被験者がリンクをクリックすることで，PC のミュージックプレイヤーが起動し，収録音を再生することで状態を確認することができる．なお,「練習」では，直前の録音のページへ戻るハイパーリンクを設置し，再生確認の後に録音のやり直しができるようにした．一方,「本番」でも，ブラウザの戻るボタンで前のページに戻ることは可能だが，録音やり直しは基本的に禁止とした（文章として戻る操作の禁止を記載）．

(4) 発話内容のテキスト入力

本実験では，書き起こしデータの整備も同時に進める為，被験者自らによる発話内容のテキスト入力を必須とした．実験の過程において，ウェブページ上にテキストボックスを設置し，被験者自身がテキストを記入し，登録するようにした．この際，発話した内容を曲解したりせず，一言一句そのまま正確に入力するように，被験者に文章で注意を促した．テキストの入力が完了したら，一連の録音ステップは終了である．

2.3 質問及びアンケートの詳細

「練習」「本番 1」「本番 2」の各録音ステップにおいて提示する質問は，**表 1** の通りである．最初の「練習」の段階では，回答内容の自由度が少なくなると想定した上で簡単な質問を設定した．そして，質問を徐々に難しくし，最後の質問では，回答の中に，被験者の好みを表現できるようにした．

表 1 各録音ステップにおいて被験者に提示した質問
Table 1 Question presented to test subjects in each recording step.

練習:	「今日は良い天気ですか?」
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.」

表 2 アンケートの設問
Table 2 List of questionnaires.

Q1. 今回の実験の操作は簡単でしたか?（YES/NO）
Q2. 今回の実験でトラブルは発生しましたか? （選択）
Q3. 録音をしてくださった方の性別を教えてください.（男性/女性）
Q4. 録音をしてくださった方の満年齢を教えてください.（数値入力）
Q5. 録音をしてくださった方の身長を教えてください．おおよそで結構です.（数値入力）
Q6. 録音をしてくださった方の職業を教えてください.（選択）
Q7. 録音をしてくだった方のご出身の都道府県を教えてください.（選択）
Q8. 録音をした場所はどこですか?（選択）
Q9. 使用したパソコンの種類を教えてください.（選択）
Q10. 使用したマイクの種類を教えてください.（選択）
Q11. 本実験に関して感想，御意見，トラブルの内容などご自由にご記入ください.（自由記述）

続いて，**表 2** に，録音ステップ終了後に実施するアンケートの設問を示す．被験者はウェブインタフェースを通じてアンケートに回答する．本稿では詳述しないが，被験者への負担を減らすため，選択肢をメニューから選ぶ形式のアンケートを重点的に設定した．

2.4 収集実験結果

準備した実験サイトを用いて，発話の収集実験を行った．実施期間は，2009 年 2 月 25 日から 3 月 30 日までである．合計 3,331 のアクセスを確認し，ユニーク IP 数で 2,011 箇所からのアクセスを得ることができた．そのうち，3 つの録音ステップ及びアンケートのタスクを完遂できたものは 432 であった（回答率 12.7%）．

ただし，この中には無効な録音データ及びアンケートの入力ミスが含まれるので，収集発話を作業者一名の人手で確認した．まずは，発話が含まれない（無音，雑音のみ）の収録ファイルを除外扱いにした．次に，自己申告による被験者の年齢・性別が，耳で聞いた限り，明らかに録音音声のそれと異なる時，その被験者の発話はすべて無効とした．同時に，各被験者の入力による書き起こしテキストが，発話通りに入力されているか確認した（異なる時は，手動で書き起こしテキストの方を修正した）．

Vol.2009-SLP-77 No.19
2009/7/18

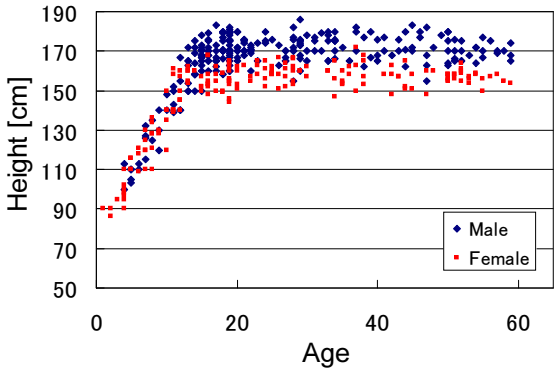

図 3 被験者の年齢・身長分布
Fig. 3 Distribution of test subject's age and height

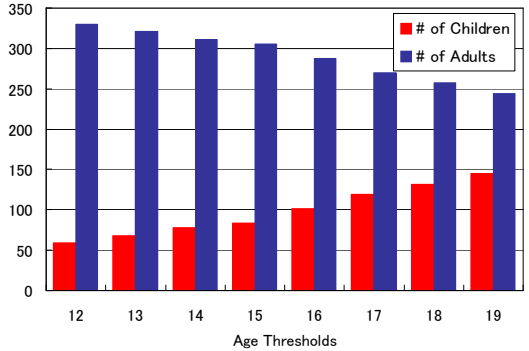

図 4 年齢閾値別大人・子ども発話数
Fig. 4 The numbers of adult / child utterances according to age threshold.

以上の作業の結果，有効な発話数は，1,109 であった．記録された被験者数は，389（ユニーク IP 数 384）であった．

2.5 年齢・身長分布

図 3 に，アンケートの自己申告によって得られた各被験者の年齢と身長の散布図を示す．横軸は年齢，縦軸は身長（cm）であり，各点は，赤は女性，青は男性の被験者を示す．

本研究では，被験者の大人と子どもの判別を目指すが，今のところ，大人と子どもの境界は明確には定義していない．安全ウェブの実現という目的を考えると，年齢で言うところの 10 代の半ばに境界年齢を設定すべきだと考えている．今回は，満年齢 12 歳から 19 歳まで 1 歳単位で，年齢の閾値を変化させ，検討を進めることにする．**図 4** は，年齢閾値を変化させた場合の，収集データにおける大人と子どもの被験者数を示したグラフである．例えば，年齢閾値 15 歳の場合は，アンケートの自己申告により満年齢 14 歳と答えた被験者までを子ども，15 歳以上を大人と考える．この場合，今回の収集発話数は大人 305，子ども 84 となる．年齢閾値を引き上げれば，子どもとみなすデータは多くなる．しかし，それでも大人の数は，子どもより 1.6〜5.6 倍多い結果となっている．

2.6 発話内容に関する分類

「本番 2」（最後の録音ステップ）で被験者に提示した質問「好きな言葉を教えてください.」に対して得られた回答の書き起こしを，その内容に応じて以下の 8 種類に分類した．作業は一人の人間（大学生）が人手で行った．**表 3** に，実際の回答（書き起こしテキスト）の一部を抜粋する．

- 単語: 文章ではなく一般的な単語のみの回答
- フレーズ: あいさつ等，日常的な生活で使う文章の回答．キャッチコピーのような文章．
- ことわざ・格言: 一般的に使われることわざや格言の回答．漫画・書籍から引用した有名な文章も含む．
- 四字熟語: 一般的な四字熟語を含む回答
- 人・場所:特定の人物や地名等を含む回答
- 「特になし」: 好きな言葉が無いと答えた回答
- 英語: 英語による回答
- 意味不明: 作業員が聞いて意味がわからなかった回答

図 5 に分類結果を示す．大人と子どもの各グラフ上の割合は，収集発話を大人と子どもに分割した後の割合を示す．なお，グラフの横軸は年齢閾値である．年齢閾値 15 歳の時は，満年齢 14 歳までを子ども，15 歳以上を大人と見なしている．

このグラフより，発話内容に大人と子どもで異なる傾向があることがわかる．子どもは「単語」の発話が多い．一方，大人は「ことわざ・格言」や「四字熟語」が多くなる．具体的には，年齢閾値 15 歳の時の大人発話における「ことわざ・格言」の割合は 23.6%に対し，子どもは 6.2%であった．「単語」に関しては大人 33.4%，45.4%となった．

Vol.2009-SLP-77 No.19
2009/7/18

表 3 回答の内容例
Table 3 Examples of sentences of collected utterances.

（単語）	「ポケモン」「根性」「チャレンジ」「愛」
（フレーズ）	「ありがとう」「今日はとても寒いです」「ラーメンセット」「好きな言葉は、大好きです」
（ことわざ・格言）	「情けは人のためならず」「あきらめたらそこで試合終了」
（四字熟語）	「一期一会」「温故知新です」「一石二鳥」
（人・場所）	「ひばりさん」「ママです、ママです」
（「特に無し」）	「特にありません」「考えたけど、思いつきませんでした」
（英語）	「Yes, We Can.」「私の好きな言葉は ”You can do it!” です」
（意味不明）	「あー、あー、うー、えー、あー、け」

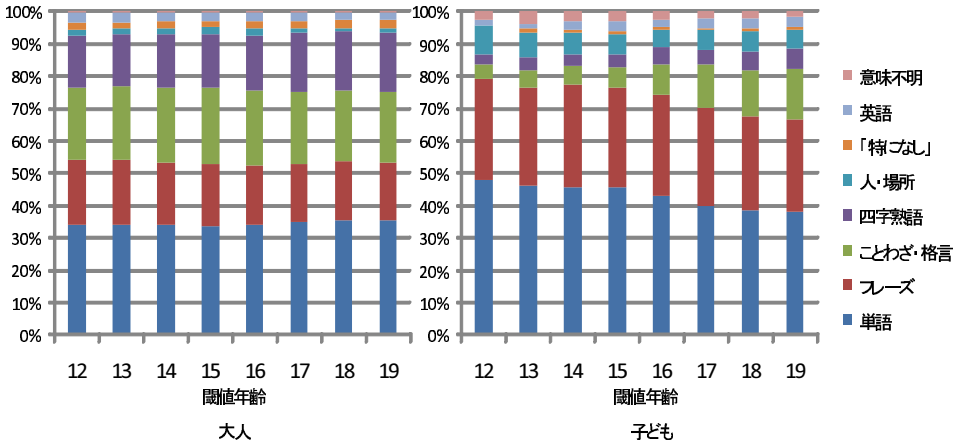

図 5 発話内容に関する分類結果
Fig. 5 Classification results concerning contents of collected utterances.

これらの結果は，大人と子どもの判別パラメタに，発話の言語的特徴（単語や言い回し）を利用することが有効であることを示唆している．

3. GMM 音響モデルを用いた若年者自動判別

集めた発話を用いて，簡単な大人・子どもの自動判別実験を実施した．今回は，発話の音響的な特徴のみに着目し，話者認識[4] に用いられる混合正規分布モデル（GMM）を音響モデルにした尤度比較を行った．これは，著者らが以前に開発した奈良県生駒市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」での大人・子ども判別で用いたのと同等の方法である[5),6)]．「たけまるくん」では，収集発話に対する発話者の年齢が未知であった．このため，収集発話に対し，作業者の主観により大人・子どものラベルを付与した．今回の実験は，録音環境（たけまるくんは公共施設に据え置き固定システム）の違いに加えて，被験者自らの自己申告によって，発話者の年齢が既知である点が異なっている．以下に実験条件と結果を述べる．

3.1 実験条件

収集発話を，年齢及び性別を基づき，子ども（女性），子ども（男性），大人（女性），大人（男性）の 4 つにクラスに分類した．学習の段階では，各クラスに対して，音響特徴量を抽出し，GMM 音響モデルを構築した．音響特徴量の抽出及び GMM の構築には，HTK 3.4.1 を用いた．分析に用いた音響特徴量は，音声認識向けに用いられることの多い 12 次元の MFCC と Δ MFCC，Δ Power である．なお，収集発話は 16bit，44.1kHz で収録を行っているが，実験前に 16kHz にダウンサンプリングをし，窓シフト長 10ms で分析を行った．GMM の混合数は 16 である．判別の段階では，入力発話に対する音響モデルの尤度を比較して最も高い尤度を得たクラスに分類する．判別に用いたデコーダは，Julian 3.5.3 である．

評価は，収集発話から評価用データを抜き出し，残りを学習用とする 10 分割の交差検定によって行った．各被験者は 3 回の発話を行っているが[⋆1]，分割のためにデータを抜き出す際，被験者単位で抜きだしを行っている．つまり，学習用データに，評価用データの被験者が含まれない状態（話者オープン）の評価実験となっている．

実験では，2.4 で述べたように，年齢閾値を 12 歳から 19 歳まで変化させた場合の 4 クラス判別の正解率を調査した．ただし，本研究の目的から考えると，男性と女性のクラス判別間違いは無視することができる．今回は，4 クラス判別をしながらも，大人・子どもの 2 つが正しく判別できた時を正解とした．

3.2 実験結果

実験結果として正解率を**図 6** に示す．実線は子ども発話を子どもとして判別した正解率，点線は大人と子どもの全発話の正解率である．横軸は年齢閾値である．全体的に大人を子どもと誤判別した事例は少数であり，逆に，子どもの正解率が低い結果となった．その中で，14 歳以下の子ども検出時（年齢閾値 15 歳）に最も高い正解率 65.9%を得た（全体の正解率は 82.6%）．

⋆1 録音不備により，実験に用いた発話が 1 または 2 個の被験者も存在する．

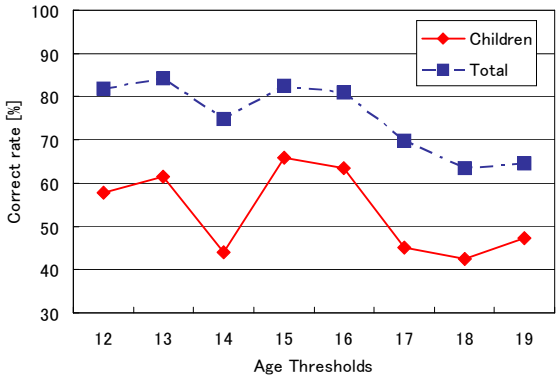

図 6 大人・子ども判別正解率
Fig. 6 Identification correct answer rate of adult and child utterances.

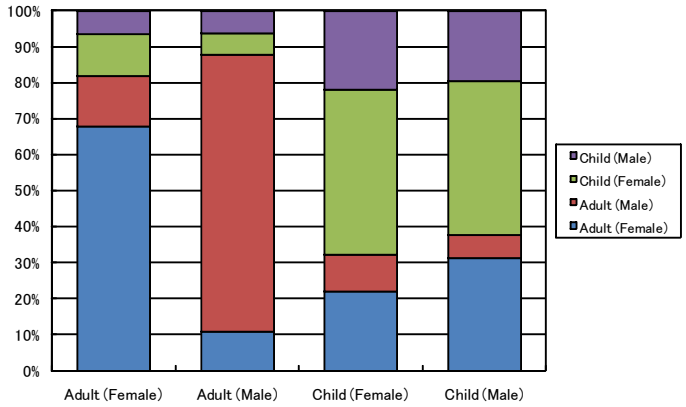

図 7 14 歳以下の子ども検出の結果詳細
Fig. 7 Experimental results on 14-years-old or younger child detection.

年齢閾値 15 歳の詳細な判別結果を**図 7** に示す．この結果から，大人に関しては，男性・女性ともに高い精度で正しいクラスに判別できていることがわかる．一方で，子どもに関しては，大人の女性と誤判別することが多い結果となった（25.3%）．また，子どもと判別ができても，男女間の判別には失敗しており（44.6%），子ども女性と出力する事例が多かった．

4. ま と め

本研究では，発話による若年者自動判別を可能とするウェブフィルタリングサービスを実現することを目的に，(1) 大人・子ども発話のネットワーク収集実験，(2) GMM 音響モデルを用いた若年者判別の予備実験を行った．

発話収集の実験では，389 名の家庭等での利用による実環境発話 1,109 を集めた．発話を分析した結果，大人と子どもで，発話の内容に異なる言語的傾向があることを確認した．

GMM モデルを用いた若年者判別実験では，14 歳以下の子ども検出において正解率 65.9%を確認した（大人も含めると正解率 82.6%）．

4.1 今度の課題

今後は，引き続き，発話の収集と詳細な分析を行う．また，若年者判別の精度が不足しており，精度向上が必要である．今回は，発話に含まれる音響的特徴を抽出した判別を行った．今後は言語的特徴も組み込んだ判別法[5),6)] の導入を検討する．

謝辞 本研究は，科学研究費補助金若手研究（B）及び和歌山大学オンリーワン創成プロジェクトの支援を受けた．

参 考 文 献

1) 原 直，宮島 千代美，伊藤 克亘，武田 一哉，“多様な音響環境下における音声認識システム利用時のデータ収集システム”，電子情報通信学会論文誌，Vol.J90-D，No.10，pp.2807–2816，2007.
2) 西村 竜一，三宅 純平，河原 英紀，入野 俊夫，“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.
3) Ryuichi Nisimura, Jumpei Miyake, Hideki Kawahara, Toshio Irino, “Development of Speech Input Method for Interactive VoiceWeb Systems”, HCI International 2009 (13th International Conference on Human-Computer Interaction), 2009. (発表予定)
4) D.A.Reynolds, R.C.Rose: “Robust text-independent speaker identification using Gaussian mixture speaker models”, *IEEE Trans. on Speech and Audio Processing*, vol.3, no.1, pp.72–83, January 1995.
5) Ryuichi Nisimura, Akinobu Lee, Hiroshi Saruwatari, Kiyohiro Shikano, “Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability”, Proc. ICASSP2004, Vol.I, pp.433–436, 2004.
6) 西村 竜一，中村 敬介，李 晃伸，猿渡 洋，鹿野 清宏，“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”，電子情報通信学会技術研究報告，SP2003-129/NLC2003-66，2003.