J-006

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

An investigation of child user identification based on speech recognition of a short sentence

宮森 翔子 †　西村 竜一 †　栗原 理沙 †　入野 俊夫 †　河原 英紀 †
Shoko Miyamori　Ryuichi Nisimura　Risa Kurihara　Toshio Irino　Hideki Kawahara

1　はじめに

従来より，あらゆる場面において子ども利用者の判別が行われてきた．たばこや酒の購入ではもちろんのこと，青少年に悪影響を与える可能性のあるウェブページでは，子どもを保護する目的で年齢確認が行われている．また，対話システムでは利用者の年齢層を判別することで，より親切な対応を実現することができる．

現在，ユーザに負担をかけずに年齢確認を行うために，生態情報を用いた判別技術の開発が進んでいる．2007 年には，たばこの自動販売機において顔画像を用いた年齢確認 [1] が話題となった．しかし，生態情報による年齢確認の技術は確立されていない．

そこで本研究では，ちょっとした一言の音声認識による子ども利用者判別システムを提案する．発話情報には，音響的特徴と言語的特徴の 2 種類を得られるという利点がある．

今回はそのうち音響的特徴についてのみを扱ったプロトタイプシステムを作成した．また，作成の過程で発話の収集実験および人間と機械の子ども判別能力の比較実験を行ったため，結果を報告する．

2　プロトタイプシステムの構成

発話の音響的特徴に基づいて子ども判別を行うプロトタイプシステムを作成した．図 1 にプロトタイプシステムのスクリーンショットを，図 2 にシステムの構成を示す．ユーザは音声ウェブシステム w3voice [2][*1] のインタフェースを操作し，発話を入力する．発話の入力は Java アプレットによって実装されている．

サーバでは入力発話が録音され，CGI を通して，音声認識システムを応用した自動判別が行われる．このシステムを動作させるプログラムのソースは 101 行である．また，音声認識はパターン認識の一つであり，識別辞書として言語モデルと音響モデルが必要である．

今回は，音響モデルとして HMM（Hidden Markov Model：隠れマルコフモデル）を用いた．これは音声認識を実現するための一般的なモデルである．図 3 に HMM による尤度比較の過程を示す．

学習の段階では，まず，収集発話を発話者の年齢と性別に基づき，「大人男性」，「大人女性」，「子ども」の 3 つのクラスに分ける．次にそれぞれのクラスについて，各発話サンプルから音響的特徴を抽出し，3 状態の GMM（Gaussian Mixture Model：混合正規分布）[3] からなる HMM を構築する．音響特徴量には 12 次元の

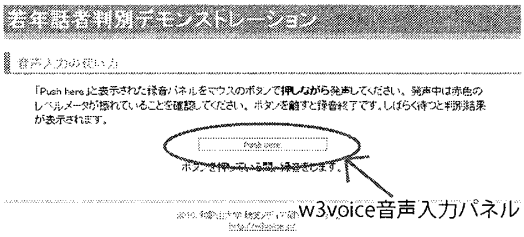

図 1　プロトタイプシステムのスクリーンショット

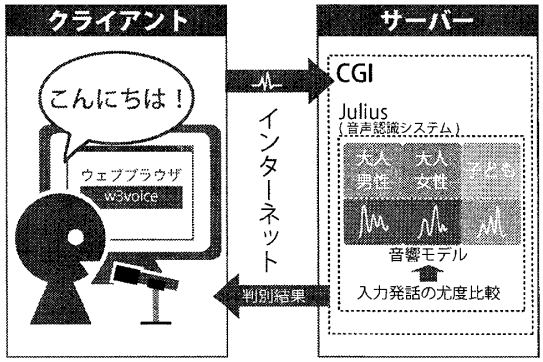

図 2　プロトタイプシステムの構成

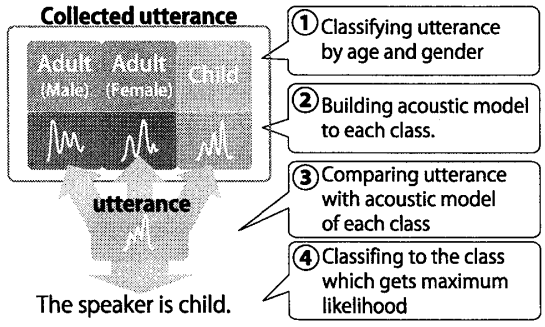

図 3　子ども自動判別処理の概要

MFCC，ΔMFCC，ΔPower を用いた．GMM の混合数は 128 である．構築には HTK3.4.1 [4] を用いた．

判別段階では，HMM の尤度比較により，より高い尤度を得たクラスを判別結果とする．判別デコーダには音声認識システム Julius4.1.4 [5] を用いた．

3　音声ウェブシステムによる発話収集実験

本研究で提案する子ども判別システムは，家庭などでの使用が想定される．実際の使用環境に即した判別実験を行うために，使用する音声をインターネットを介

† 和歌山大学，Wakayama University

*1 http://w3voice.jp/

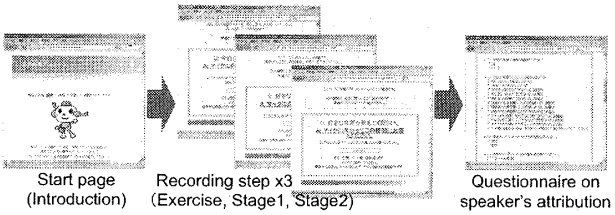

図 4　実験用ウェブサイトの構成

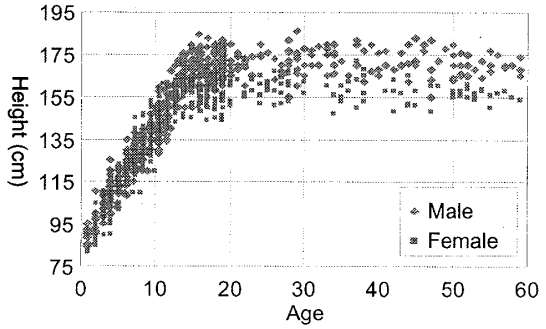

図 5　発話者の身長・年齢分布

して収集した．録音には，音声ウェブシステム w3voice を用いた．w3voice により，録音された発話は自動的にサーバーにアップロードされる．

図 4 に実験用ウェブサイトの構成を示す．被験者は，「練習」「本番 1」「本番 2」の合計 3 回の発話を録音した．各録音ステップにおいて，発話者には簡単な質問が提示される．本番 1，本番 2 の質問を以下に示す．

- 本番 1: 好きな食べ物を教えてください．
- 本番 2: 好きな言葉を教えてください．

発話者は，全ての録音ステップが完了した後，発話者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，発話者が低年齢の際は，録音時の PC 操作およびアンケートの回答は，付添いの保護者が代行するように事前に要請した．アンケート項目を表 1 に示す．

3.1　収集実験結果

収集実験の結果，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 5 に，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を，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 6 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．よって今後，10 代の発話サンプルを追加収集する予定である．

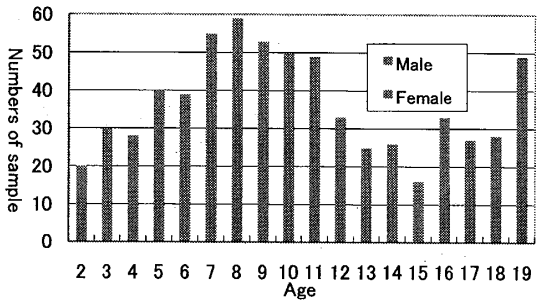

図 6　*2～19* 歳の発話者の分布

4　人間と機械の子ども判別能力の比較

子ども自動判別における目標値と，自動判別に用いる年齢閾値の検討のため，人間と機械の子ども判別能力を比較した．

なお，本研究では，大人と子どもの境目となる年齢を年齢閾値という概念で表わした．例えば，年齢閾値 15 歳の場合，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．判別では，年齢閾値を 9 歳から 18 歳まで 1 歳ごとに変化させて検討した．

4.1　人間の主観による子ども判別実験 実験条件

収集発話を用いて，人間の主観による子ども判別実験を行った．被験者は 5 名 (男性 2 名，女性 3 名) であった．判別対象の発話は収集発話のうち本番 2（質問「好きな言葉を教えてください」）で録音された 260 発話（うち男声 146 発話，女声 114 発話）とした．被験者はスピーカーから音声を聞き，次の質問に答えた．

- 大人か子どものどちらに聞こえるか
- およそ何歳に聞こえるか
- 男女どちらの性別に聞こえるか

4.2　機械による子ども判別実験 実験条件

本研究では，音声認識システムを応用して大人・子ども自動判別を実現した．前述の HMM 音響モデルの学習には，収集実験で録音された 2,361 発話を用いる．

評価では，収集発話から評価用データを抜き出し，10 分割交差検定を行った．評価用データでは，学習段階で使用した発話者の発話を除いている（話者オープン）．これは，実際のアプリケーションでの使用を想定した条件となっている．

4.3　正解率による結果の比較

図 7 は子ども発話の正解率を比較したグラフである．正解率は，年齢閾値に満たない年齢の発話者の音声が，子どもであると正しく判別された割合を示す．図の青線は人間による判別の正解率，緑線は自動判別の正解率を示す．横軸は年齢閾値である．全体的に，自動判別は人間による判別よりも正解率が低い．自動判別の正解率の最高値は，年齢閾値 13 歳における 68.7% であり，人間による判別の最高値は年齢閾値 12 歳における 87.7% であった．

年齢閾値が 15 歳以上の場合では，どちらの判別においても正解率が減少する傾向がみられた．この傾向の一因には，変声期の影響が考えられる．変声期における音声は音響的に大きな変動がある．そのため，人間

表 1　アンケートの設問

Q1.	今回の実験の説明はわかりやすかったですか？（選択・5 段階）
Q2.	今回の実験の操作は簡単でしたか？（選択・5 段階）
Q3.	今回の実験でトラブルは発生しましたか？（選択）
Q4.	録音をしてくださったお子様の性別を教えてください（男性/女性）
Q5.	録音をしてくださったお子様の満年齢を教えてください（数値入力）
Q6.	録音をしてくださったお子様の身長を教えてください．おおよそで結構です．（数値入力）
Q7.	録音をしてくだったお子様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）

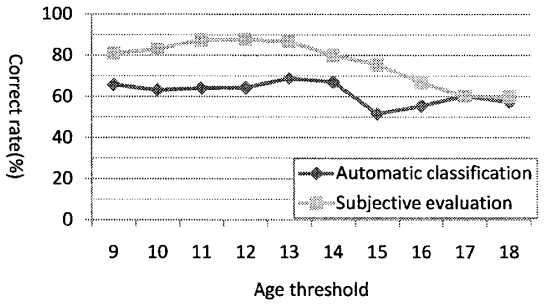

図 7　子ども判別における正解率の比較

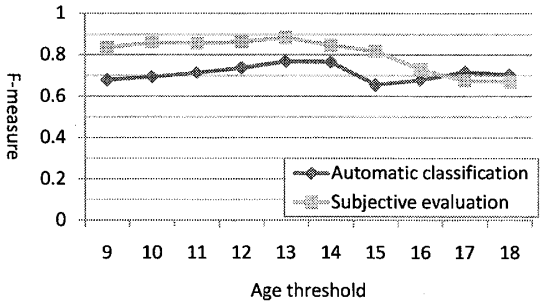

図 9　子ども判別の F 値による比較

にとっても年齢の判別が難しいことが予想される．

さらに，図 8 では子どもの発話サンプルがどのクラスに判別されたかを年齢閾値ごとに示している．左は自動判別の結果を示すグラフであり，右は人間による主観評価の結果を示す．グラフの緑の部分は発話サンプルが子どもであると判別された割合を示す．同様に，赤い部分は大人女性，青い部分は大人男性と判別された割合である．グラフから，人間による判別と自動判別では誤判別の傾向が違うことがわかる．人間は，年齢閾値の上昇に伴って，子どもを大人男性と間違えた．一方，自動判別では，年齢閾値に関わらず一定して子どもと大人女性を誤判別する傾向にあった．

4.4　F 値による結果の比較

子ども判別を F 値によって比較した．F 値は情報検索システムの性能を表わす総合的な評価尺度であり，次式で求められる．式中の *precision* は適合率，*recall* は再現率である．

$$F-measure = \frac{2 \times precision \times recall}{(precision + recall)} \qquad (1)$$

適合率は正確性を示す指標であり，式 (2) によって求まる．再現率は網羅性を示す指標であり，式 (3) で求まる．式中の R は適合されたサンプルの数，N は検索結果のサンプルの数，C は全対象サンプル中の正解サンプルの数である．なお，今回は子どものデータに対して評価を行うため，R は子ども発話として判別された数，N は子ども発話を子どもとして判別した数，C は全ての子ども発話の数である．

$$precision = \frac{R}{N} \qquad (2)$$

表 2　変声期を考慮して設定した 6 クラスの一覧

		年齢
子ども	低年齢層	2 ~ 10 歳
	高年齢層	11 ~ y_{child} 歳
大人		y_{child} 歳以上

(性別：女性 / 男性)

$$recall = \frac{R}{C} \qquad (3)$$

図 9 に F 値での比較結果を示す．年齢閾値が 13 歳のとき，どちらの判別においても最高値であり，自動判別では 0.77，人間による主観評価では 0.88 を得た．

5　変声期の音声が子ども判別に与える影響の調査

変声期における音声の影響を除き，より高い年齢閾値での判別を実現するため，再び自動判別実験を行った．表 2 に，本実験で用いた HMM 音響モデルのクラスをまとめる．ここで，各年齢層ごとに男女を別クラスとし，合計 6 クラスを新たに定義した．

今回は，高年齢層の子どもと大人の境界となる新たな年齢閾値 y_{child} を用いて実験を行った．この年齢閾値 y_{child} を，16~20 歳まで 1 歳刻みで変化させて検討を進める．また，子どもの低年齢層と高年齢層の境界となる年齢は，10 歳で固定した．10 代を子どもの高年齢層のクラスとすることで，変声期における音声を子どもクラスの学習から分離できた．

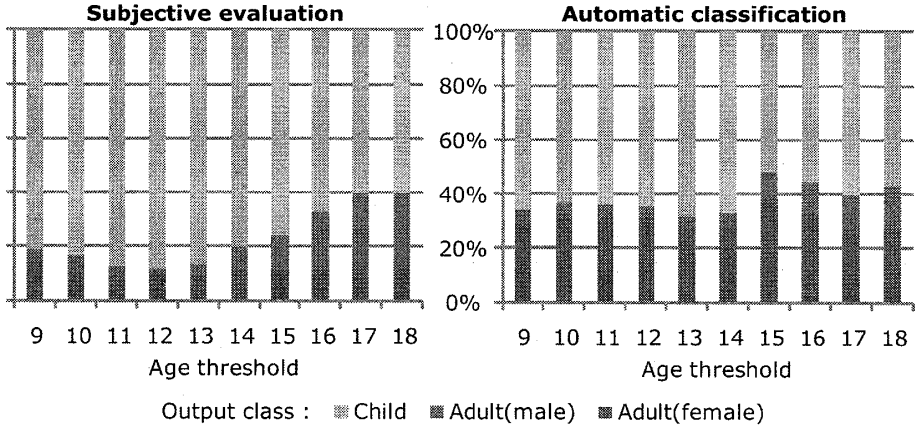

図 8　子ども判別の詳細な結果

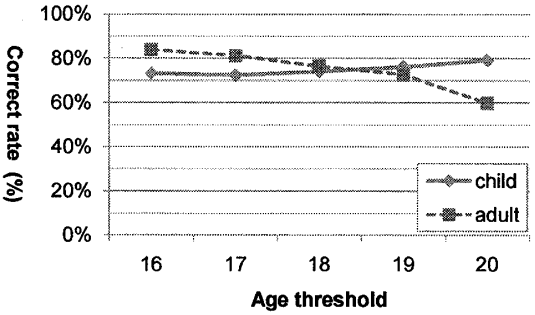

図 10　*6* クラス判別における正解率

5.1　6 クラス判別の結果

図 10 は 6 クラス判別の正解率のグラフである．図中の緑線が子ども，青線が大人の正解率を示す．子どもの正解率では，全ての年齢閾値において 70% 以上の正解率を得た．これは，変声期における発話を子ども発話から除いたことにより，音響特徴量の変動が抑えられたためだと考えられる．しかし，年齢閾値が上昇するに従って，大人の判別精度は低下している．これは大人の発話サンプル数が減少したためだと考えられる．

6　まとめ

本研究では，ちょっとした一言の音声認識による子ども判別システムについて，そのプロトタイプを作成した．また，提案システムを実現するために，発話の収集実験と，収集発話を用いた判別実験を行った．

収集実験では，音声ウェブシステム w3voice を用いることで，3,057 発話を収集した．このうち，15 歳以下の発話者による発話は 59.7% であった．判別実験では，人間と機械の判別能力について比較をし，自動判別の目標値と，自動判別に用いる年齢閾値の検討を行った．比較結果のまとめを表 3 に示す．機械による判別の正解率は，人間による判別よりも 17.8% の低下に収まった．また，どちらの判別においても，年齢閾値 15 歳以上では正解率が低下した．詳細に判別結果を分析したところ，人間と機械では判別の傾向に違いがあった．人間は年齢閾値が高くなるにつれ，子どもを大人男性と間違えることが多く，自動判別は一定して子どもを大人女性と誤判別する傾向にあった．

より高い年齢閾値での判別を実現するため，変声期

表 3　年齢閾値 *13* 歳における判別結果のまとめ

	人間による主観評価	機械による自動判別
正解率	87.7%	68.8%
F 値	0.88	0.77

における音声を子どものサブクラスに分けて，6 クラスでの自動判別実験を行った．その結果，15 歳以上の全ての年齢閾値において 70% 以上の正解率が得られた．

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 [7] を導入する予定である．並行して，プロトタイプシステムによる発話の収集および分析を進めていく．

謝辞　本研究の一部は，科学研究費補助金及び和歌山大学 H22 年度学長裁量経費の支援を受けた．

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] 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.

[3] 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, 1995.

[4] Young, S.J., et al., "The HTK book version 3.4", Cambridge University Engineering Department, Cambridge, UK, 2006.

[5] Lee, A., et al., "Julius — An Open Source Real-Time Large Vocabulary Recognition Engine", Proc. Eurospeech 2001, pp.1691–1694, 2001.

[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.（発表予定）

[7] Nisimura, R., et al., "Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability", Proc. ICASSP2004, Vol.I, pp.433–436, 2004.