6F-7

音声認識を応用した大人・こども話者識別法における言語的特徴の利用

西村 竜一
和歌山大学 システム工学部

1 はじめに

我々は，音声認識技術の応用として，発話を入力とする大人・子ども自動識別法の検討を行っている．大人・子どもの自動識別は，利用者の属性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，かつユーザに適したインタラクションの提供を可能とする．生体情報の一つである発話を入力とする場合，音響的特徴と言語的特徴の2つの側面から大人・子どもの識別を実現することが可能である．また，音声対話インタフェースと組み合わせることで，自然な対話を繰り返し，利用者に負担を与えることなく，提案法を実現できるメリットがある．

これまでの報告 [1, 2] では，収録音声信号に周波数分析等を適用して抽出した音響的特徴のみを識別器の素性としており，使用語彙や言い回しの傾向を起源とする言語的特徴については検討が不十分であった．また，発話を用いる場合，これまでの実験では，大人と子どもを区別する年齢境界の上昇に伴う精度低下を確認している．特に，変声期に当たる10代後半の発話に対しては，人間の耳でも大人と子どもを判別することは難しく，自動識別も容易ではない．言語的特徴を加えることができれば，より高精度な識別を実現できる可能性がある．そこで，収集した大人・子ども発話を多角的に分析し，識別に有効な言語的特徴を得ることを目指す．

2 音声ウェブシステムを用いた発話の収集

本研究で分析の対象とした発話のデータについて述べる．これまでに，我々は，音声ウェブシステム w3voice[3] を用いてインターネットを介した発話の収集を行った．本研究用に作成したウェブサイトでは，音声入力に対応している．利用者が発話を行う過程として,「練習」「本番1」「本番2」があり，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の設問内容は,「好きな食べ物は何ですか?」「好きな言葉を教えてください.」である．発話者は各ステップにおいて設問への回答を発話し，ブラウザ上のプログラム（Java アプレット）が声を録音する．収録信号は，我々のウェブサーバに自動的にアップロードされる．

本研究では,「好きな言葉を教えてください.」という質問に対する回答を対象として扱う．これは，この質問が，他の質問よりも個人差に起因する多様性が大きい自由な発話が回答として期待できるものであり，今回の分析に適していると判断したためである．意味のある発話ができていない0，1歳の収集データはあらかじめ除外し，2歳以降の収集発話のうち，正しく音声が含まれている発話を分析した．以下では，これらの発話の内容をテキストとして人手で書き起こしたものを使用する．

Development of Linguistic Features for the ASR-based Child Speaker Detection Method, Ryuichi NISIMURA (Faculty of Systems Engineering, Wakayama University)

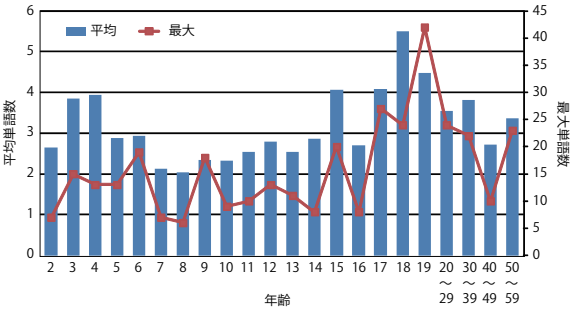

図 1: 収集発話に含まれる単語数の平均・最大

3 ウェブ収集発話の言語的特徴に関する検討

3.1 単語数

形態素解析ツール Mecab[4] を用いて収集発話の書き起こしを形態素解析し，単語数を求めた．2歳から19歳までは各年齢ごとに，20歳以上は10歳ごとに一つの発話に含まれる単語数の平均値と最大値を求めた．

結果を図1に示す．図の横軸は年齢，縦軸は単語数である．棒グラフは左縦軸の平均単語数を示し，折れ線グラフは右縦軸の最大単語数を示している．平均数は2.03から5.50であり，年齢ごとの違いに有意な差を認めることはできなかった．

最大単語数に着目すると，単語数が20以上あった発話は17歳から52歳を発話者とする7つの発話のみであった．この結果からは20単語以上の長いものが大人の発話であると判断することはできない．しかし，年齢が高いほど長く発話する傾向があることは確認できた．

3.2 品詞

形態素解析をした際，同時に得られる品詞情報を用いて，各発話に含まれる品詞の含有率を求めた．使用した品詞は，名詞，助動詞，感動詞，助詞，動詞，形容詞の6種類である．

結果の一部として，図2に感動詞，図3に助詞の結果を示す．他の品詞に比べて感動詞と助詞は，年齢によって異なる傾向がみられることが分かった．感動詞は年齢が低い12歳以下で17%を超えるとなった．また，助詞は年齢が高い15歳以上で11%以上の含有率となった．

3.3 新聞記事掲載単語との比較

新聞記事に掲載されている文章は，正しく整った日本語で，かつ子どもには難しい表現が多く使われている．そこで収集発話に含まれる単語と新聞記事に掲載されている単語を比較して，多く一致すれば大人の発話であり，一致しなければ子どもである可能性が高いと仮説を立て，調査を行った．

新聞記事7年分のテキストを単語に分割し，新聞記事掲載単語のリスト (365,213 単語) を作成した．単語リストと発話を比較し，一致しない単語を抽出した．その結

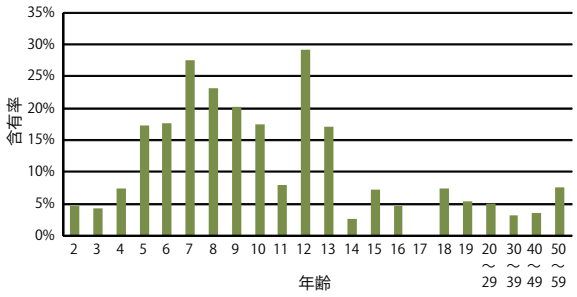

図 2: 収集発話に含まれる感動詞の含有率

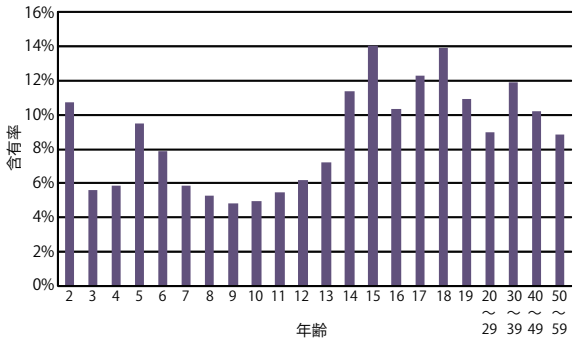

図 3: 収集発話に含まれる助詞の含有率

果，32 単語が抽出され，この 90%である 29 単語の発話者が子どもであった．

また，新聞記事掲載単語を出現頻度順に上位 60,000 単語に限定した比較も行った．この 60,000 単語は，新聞記事 7 年分の単語全体の 99.3%に相当する．収集発話と比較し，一致しない単語には子どもが 1,458 単語，大人が 436 単語が抽出できた．抽出した単語を 16 種類 (アニメ系，ゲーム系，食べ物系，人を表す語，動物，その他固有名詞，挨拶，応答，状態・状況，助詞など，動きを表す語，その他の名詞，四字熟語，流行り言葉，慣用句，意味不明) に分類すると，大人と子どもで種類の異なる単語を好んで利用することがわかった．大人は四字熟語や慣用句，子どもはアニメやゲーム関係などの言葉を使用することが多い結果となった．

3.4 Bag-of-Words（BOW）

次に，語順を無視し，各単語の出現回数で構成したベクトルである Bag-of-Words（BOW）の利用を検討した．自然言語処理の研究では，テキスト分類のタスクにおいて，BOW を素性とした Support Vector Machine（SVM）の利用に高い識別性能を得ている [5]．

今回，収集発話の書き起こしと，音声認識（ASR）の出力結果である単語列からそれぞれ BOW を構成し，比較した．収集発話の書き起こしを形態素解析した後，そこに含まれる単語の出現回数を求め，各単語に割り振られた ID との対からベクトルを作成した．一方，音声認識には，Julius[6] の出力を用いた．言語モデルには，収集発話の書き起こしから作成した単語 3-gram モデルを用いた（登録単語数 981）．音響モデルは，別途用意した子ども発話で適応を施したトライフォン HMM である．

SVM は二値識別であるため，収集発話を大人と子ども

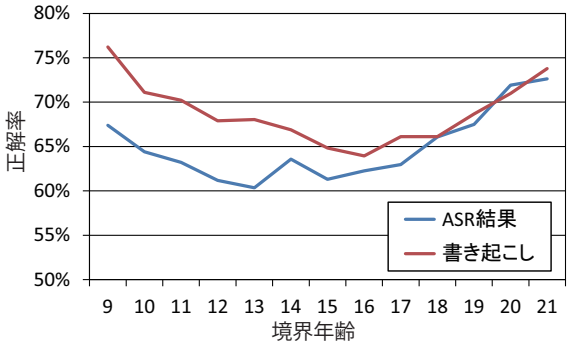

図 4: BOW を素性とした SVM 二値識別結果（正解率）

の二つの集団に分けて，2class における正解率を調査する．その際，大人と子どもの年齢の閾値となる境界年齢を設定した．例えば，境界年齢 16 歳では、16 歳以上の話者を大人、16 歳未満を子どもとみなす。評価は，収集発話の全体を 10 分割した交差検定によって行った．SVM の実装には線形カーネルの LIBSVM[7] を用いた．

図 4 に正解率を示す．横軸は境界年齢を示し，赤線は書き起こし，青線は音声認識の結果から BOW を構成したときの結果である．この結果から，BOW のみで 60%以上の正解率を得ることができることがわかった．また，音声認識の結果を用いると，低い境界年齢においての精度低下が書き起こしよりも大きいことがわかった．

4 まとめ

本稿では，発話を入力とする大人・子ども自動識別において，性能向上に寄与する言語的特徴を見つけることを目指し，単語数，品詞，新聞記事掲載単語との比較の観点でウェブ収集発話を分析した．加えて，SVM による二値識別において Bag-of-Words の利用を検討した．今後は，さらなる詳細な分析をし，識別アルゴリズムを検討する．

謝辞 本研究は，本学卒業生 仲 希望氏，宮森 翔子氏の協力によって行ったものである．両氏に深く感謝いたします．また，本研究は，(独）科学技術振興機構（JST）研究成果最適展開支援事業 A-STEP FS ステージ探索タイプの支援を受けた．

参考文献

[1] 宮森ら, 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.

[2] R.Nisimura, et al., Detecting child speaker based on auditory feature vectors for VTL estimation, Proc. APSIPA ASC, 2012.

[3] R.Nisimura, et al., Development of Speech Input Method for Interactive VoiceWeb Systems, Lecture Notes in Computer Science (Proc. HCI International), vol.5611, pp.710-719, 2009.

[4] http://mecab.sourceforge.net/

[5] 松本, 自然言語処理におけるカーネル法の利用, 第 5 回情報論的学習理論ワークショップ予稿集 IBIS2002, pp.19–24, 2002.

[6] http://julius.sourceforge.jp/

[7] http://www.csie.ntu.edu.tw/~cjlin/libsvm/