様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ４ 日現在

機関番号：１４７０１
研究種目：若手研究(B)
研究期間：2013～2014
課題番号：２５８７０４３８

研究課題名（和文）ウェブインタフェースに向けた聴覚機構に基づく大人・子ども話者識別技術の研究

研究課題名（英文）A method for identifying adult and child speaker based on auditory model for web system interface

研究代表者
西村　竜一（Nisimura, Ryuichi）

和歌山大学・システム工学部・助教

研究者番号：００３７９６１１

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究は、発話を入力としたコンピュータによる大人・子ども識別法の研究開発を行った。(1) 聴覚特徴の抽出と識別アルゴリズムへの組み込み：深層学習のニューラルネットワーク（DNN）の導入を試みた。(2) 特徴的な言語情報を引き出すための対話戦略決定法の検討：収集した発話の言語情報であるBag-of-Words及び新聞記事に含まれる語彙の共起情報を特徴量として検討した。(3) ウェブサービスとしてのシステム改良：これまでPC端末向けに開発してきたが、Android端末でも提案システムが動作するように改良を加えた。また、音声入力ユーザインタフェースのデザインについて実験を通じて検討を加えた。

研究成果の概要（英文）：In this study, we have developed web-based speech interface system to identify adult and child speakers as follows: (1) Auditory features were integrated into the identification algorithms to improve accuracies. We have introduced deep learning neural networks (DNNs). (2) Interactive strategy method of determining language information were investigated. Co-occurrence information of vocabulary included in the newspaper articles and Bag-of-Words features were tested as linguistic features. (3) We have improved the proposed system as web service. The system is able to run on the Android terminal in addition to the PC terminal. We also studied about the design of the voice-enabled web user interface on the basis of the experimental results.

研究分野：音情報処理

キーワード：音声情報処理　ウェブシステム　こども　ディープニューラルネットワーク　言語情報　インタフェースデザイン

１．研究開始当初の背景

本研究は、生体情報の一つとして、発話を入力とする、大人・子ども自動識別技術を、ICT（情報通信技術）の基盤であるウェブ上に展開する。つまり、人間の肉声を用いて、ウェブの利用者が大人か子どもかを判断する技術を、さまざまなアプリケーションに組み込み可能なウェブサービスとして実装する。現在、「安心・安全な国民生活の実現」に向けた ICT の活用が求められている。インフラとしての、インターネットの拡大に伴い、未熟な子どもが ICT 技術に接することが多くなっている。同時に、インターネットの「匿名掲示板」や「アダルトサイト」が社会問題化しており、親や保護者は、これら危険な情報から、子どもが守られることを求めている。このような社会情勢のなかで、大人・子どものオンライン識別技術は、欠くことのできない基盤技術である。

２．研究の目的

提案法は、発話を入力に用いた自然な対話を繰り返し、利用者に負担を与えることなく、精度向上を得ることができることに利点がある。ただし、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。そこで、本研究では、聴覚理論及び対話研究を工学的に発展させ、10 代後半の若年者を対象に、技術移転が可能なレベルとみなす 80%の精度を得ることを目標とする。

３．研究の方法

(1) 聴覚特徴の抽出と識別アルゴリズムへの組み込み

識別アルゴリズムに深層学習のニューラルネットワーク（DNN）の導入を試みた。DNN は、従来型である隠れマルコフモデル（HMM）と異なり、そのままでは入力信号の時間伸縮を適切に扱うことが難しい。これに対し、識別単位をフレームとした場合でも高い能力を示すことを実験から確認した。

実験にあたり、特徴量（GCMC，ガンマチャープ変調係数）は、動的圧縮型ガンマチャープ聴覚フィルタバンク（GCFB）から抽出した成分に変調スペクトルを付与することで構成した。GCFB は、聴覚末梢系の周波数分析機能を模擬したフィルタバンクである。我々の先行研究では、音声認識に一般に使用される特徴量（MFCC，メルケプストラム係数）の内部フィルタバングであるメルフィルタバンクと比較して、GCFB を用いた声道長比推定の有効性を確認している。声道長は、人の体型の一つのファクタであるため、GCFB は話者年齢の推定に有効な特徴を抽出する手段と考える。本研究では、GCFB の出力に対し、MFCC の導出と同様に、対数及び離散コサイン変換

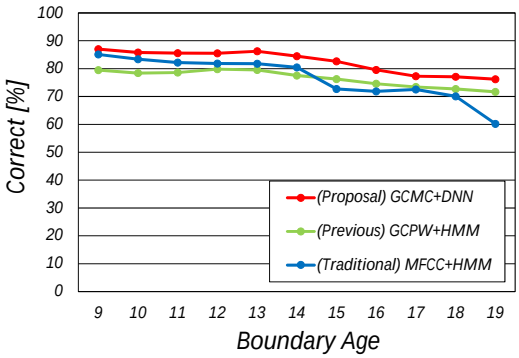

図１「大人」「子ども」2 クラス識別正解率

を適用し、パワー項である 0 次項目を取り除いた$\{b_1,\cdots, b_{12}\}$（12 次元）を特徴量とする。加えて、MFCC でのΔ項に代わる動的な特徴量として、0 次項を含む$\{b_0,\cdots, b_{12}\}$の時間方向における帯域制限変調スペクトル（13 次元）を付与し、GCMC（合計 25 次元）とする。本研究での使用する発話サンプルは、音声ウェブシステムを通して録音された発話であり、録音環境はコントロールされていない。言語情報に関わるとされる 2～16Hz の帯域を抽出することで録音状態の差異の抑制を意図している。

後述の実験で比較するために、MFCC（12 次元）にΔMFCC、ΔPower を加えた計 25 次元の特徴量を用意した。加えて、GCFB の 25 チャンネルのフィルタ出力にΔを加えた特徴量（GCPW）とも比較する。

DNN の実装には、Python のライブラリである Theano を利用した。識別処理はフレーム単位で行い、出力は、「大人女性」「大人男性」「子ども」「無音区間」の 4 つである。中間層のニューロン数を 512 とする 5 層の DNN を作成した。入力層のニューロン数は、25 次元の GCMC に前後 5 フレームを加えて 275 とした。特徴量は、平均分散により正規化して用いた。フレーム単位の識別であるため、発話セグメント単位の出力の決定には投票方式を用いた。フレーム単位で最も出力回数の多かったクラスをセグメント単位の出力とする。ただし、「無音区間」の出力はカウントから除外した。

比較する HMM は「大人女性」「大人男性」「子ども」の３クラスを持つ GMM-HMM（状態数 3）である。学習には HTK、デコーダには Julius を用いた。

実験結果として、入力発話を「大人」と「子ども」の 2 クラスに識別したときの正解率を図 1 に示す（出力結果の「大人女性」「大人男性」は同一の「大人」クラスとみなす）。

実験に用いたのは、クラウドソーシングで集めた 2,360 個の発話である。話者は、2 歳から 59 歳まで分布している。この実験では、モデルの学習から評価用のサンプルを除外する 10 分割交差検証によって実環境を再現した（話者オープン）。

図中の縦軸は、正解率であり、横軸は、大人と子どもの境界とする発話者の年齢であ

る。例えば、境界が 16 歳の場合、16 歳以上の話者を大人、16 歳未満の話者を子どもとみなす。青線が MFCC と HMM を利用する一般法、赤線が提案法（特徴量 GCMC、識別手法 DNN）である。また、比較のために、緑線に GCPW と HMM を用いた先行法の結果を示す。

図 1 から確認できるように、総じて提案法（赤）が高い性能を示している。例えば、従来法（緑）に対し提案法（赤）は、境界年齢 16 歳のとき 5.0 ポイント、18 歳のとき 4.3 ポイントの向上、平均すると 5.7 ポイントの向上である。境界年齢 16 歳のときは、正解率が 79.6%であった。また、従来からの問題である境界年齢上昇にともなった正解率低下を抑えることに成功している。

しかしながら、聴覚特徴の抽出処理におけるアルゴリズムの高度化やデータ処理量増加に伴うシステム全体の処理時間増加が問題となった。現状では、GCMC の抽出に膨大な計算を必要とし、2 秒の発話に対する結果提示には 1 分程度の時間を要する。

この問題に対し、メルフィルタバンクなどの簡易な特徴量との比較を追加するなど多角的な分析を加えた。また、聴覚特徴の抽出処理自体にも、計算スピードの高速化に向けた改良の検討をはじめた。

また、話者の身体情報を示す特徴の一つとして声道長に着目し、日本語母音データベースを用いた任意発声の相対的声道長の推定手法について検討を加えた。

(2) 特徴的な言語情報を引き出すための対話戦略決定法の検討

過去に我々が収集した発話データの語彙情報を統計的に分析して、対話システムに組み込むことができる形態で、その特徴を抽出した。具体的には、Bag-of-Words 及び新聞記事データベースに含まれる語彙の共起情報を検討した。

本研究では、クラウドソーシングで収集した発話データのうち、「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った。これは、この質問が、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、言語情報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。

形態素解析ツール Mecab 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 2 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

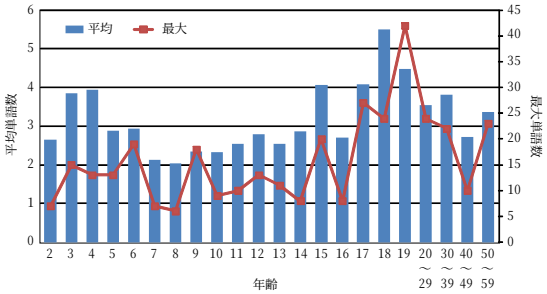

図 2 収集発話に含まれる単語数の平均・最大

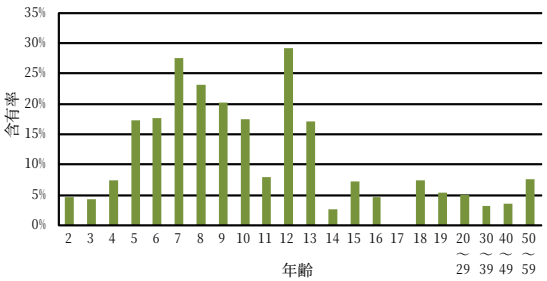

図 3 収集発話に含まれる感動詞の含有率

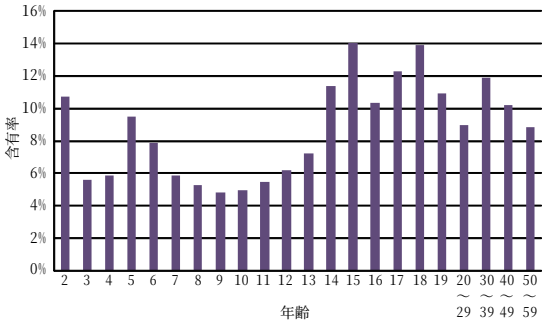

図 4 収集発話に含まれる助詞の含有率

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。

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 3 に感動詞、図 4 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。

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

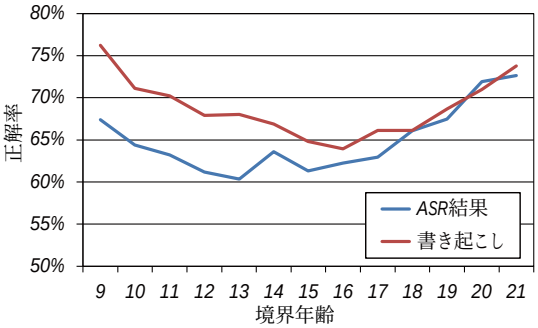

図 5 BOW を素性とした SVM 二値識別結果（正解率）

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 単語を抽出できた。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。

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。

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 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の二つの集団に分けて、2class における正解率を調査した。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 を用いた。

図 5 に正解率を示す。横軸は境界年齢を示

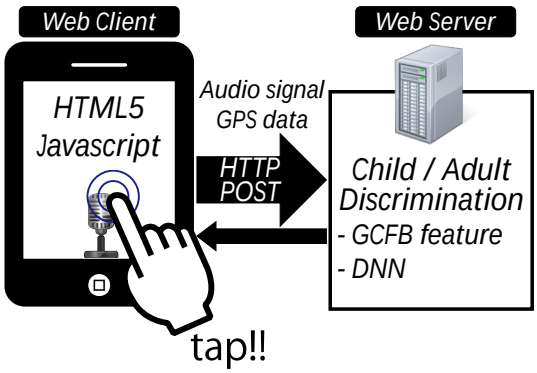

図 6 開発システムの構成

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．

(3) 公開試験に向けたウェブサービスとしてのシステム改良

提案システムの本格的な公開に向けて、システムの改良を行った。図 6 にシステム構成を示す。これまでに行った提案法をウェブサーバ側プログラムに適用することで、最新の研究成果を広く公開できるようにした。同時に、これまでのデモシステムは、主に PC 端末向けに開発してきたが、より多くの利用者を獲得するために、Android 端末でも提案システムが動作するように改良を加えた。また、音声入力のユーザインタフェースを設計する際のデザインについて実験を通じて検討を加えた。これは、音声入力ユーザインタフェースに慣れていない利用者が適切に本システムを利用できるようにするための改良である。ユーザビリティを向上し、多くの人に利用してもらうのに必要な検討であり、有意義な知見を得ることができた。

4．研究成果

現在のところ、大人と子ども発話識別における正解率の最良値は、79.6%である。これは、当初、目標とした 80%にほぼ匹敵する。また、言語情報の統合などに関しては、改善の余地が残っており、今後の研究に繋がる成果を残すことができた。

しかしながら、処理速度に関しては、実験によって、その問題点が表面化したが、問題を解決するまでには至らなかった。処理速度の問題は、計算機の発展とともに、ある程度は自然に改善するものであるが、今後も継続的に検討を加える必要がある。

音声ウェブシステムのユーザインタフェースデザインに関する検討は、これまでに無い新しい観点での研究であり、一定の成果を残すことができた。応用システムの本格的な普及に向けて、重要になると思われるので、

本研究の結果をスタートアップとして、新しい研究課題を提案していきたいと考える。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計　9 件）

① 田藤千弘, 西村竜一, 入野俊夫, 河原英紀, ウェブアプリケーションにおける音声入力 UI の設計と評価について, 日本音響学会 2015 年春季研究発表会, 2015 年 3 月 16 日, 中央大学後楽園キャンパス（東京都） 査読無し
② 田藤千弘, 西村竜一, HTML5 による音声入力ウェブアプリケーションの開発キット, 日本音響学会 2014 年秋季研究発表会, 2014 年 9 月 4 日, 北海学園大学豊平キャンパス（北海道・札幌市） 査読無し
③ Minori Matsuyama, Ryuichi Nisimura, Hideki Kawahara, Junnosuke Yamada, Toshio Irino, Development of a mobile application for crowdsourcing the data collection of environmental sounds, HCI International 2014, 2014 年 6 月 25 日, Heraklion, Crete, Greece, 査読有
④ 松山みのり, 津田貴彦, 西村竜一, 河原英紀, 山田順之介, ROCKON: スマホを用いた環境音の収集と認識システム, 情報処理学会音学シンポジウム 2014, 2014 年 5 月 24 日, 日本大学文理学部(東京都) 査読無し
⑤ 松山みのり, 津田貴彦, 西村竜一, 山田順之介, 入野俊夫, 河原英紀, クラウドソーシングによる環境音収集に向けたスマホアプリの開発, 電子情報通信学会 2014 年総合大会, 2014 年 3 月 19 日, 新潟大学五十嵐キャンパス（新潟県・新潟市） 査読無し, 電子情報学会学術奨励賞受賞
⑥ 西村竜一, 田藤千弘, 音声から大人と子どもを識別するウェブシステムにおける DNN の適用, 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 11 日, 日本大学駿河台キャンパス（東京都） 査読無し
⑦ 小林真優子, 坂口諒, 西村竜一, 入野俊夫, 河原英紀, 日本語母音データベースを用いた声道長推定法の校正について, 日本音響学会 2014 年春季研究発表会, 2014 年 3 月 11 日, 日本大学駿河台キャンパス（東京都） 査読無し
⑧ 松山みのり, 津田貴彦, 西村竜一, 河原英紀, 入野俊夫, 環境音収集アプリのための UI 設計 ～クラウドソーシング型データ集積サービスの提案～, 日本音響学会関西支部第 16 回関西支部若手研究者交流研究発表会, 2013 年 12 月 8 日, 産業技術総合研究所関西支部（大阪府・池田市） 査読無し
⑨ 小林真優子, 坂口諒, 西村竜一, 入野俊夫, 河原英紀, 日本語母音データベースを用いた任意発声の相対的声道長の推定について, 日本音響学会 2013 年秋季研究発表会, 2013 年 9 月 27 日, 豊橋科学技術大学（愛知県・豊橋市） 査読無し

〔その他〕
ホームページ等
http://w3voice.jp/

６．研究組織
(1)研究代表者
西村　竜一（NISIMURA, Ryuichi）
和歌山大学・システム工学部・助教
研究者番号：00379611