様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 29 日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：１９７００６４１
研究課題名（和文）　画像化音声を用いた外国語発話訓練システムの開発

研究課題名（英文）　Development of the real-time speech visualization system for language acquisition

研究代表者
坂田　聡（SAKATA TADASHI）
熊本大学・大学院自然科学研究科・助教
研究者番号：80336205

研究成果の概要：パソコン用アプリケーションとしてリアルタイム動作する発話訓練システムを開発し，訓練システムにおいて言語特有の音素特徴を画像情報に変換するために必要であるニューラルネットワーク（NN）構築を行った。日本語用 NN を用い，日本語非母語話者による訓練システムの評価を行った。さらに，発話において重要なイントネーションや発話速度などの情報を付加するため，それらのリアルタイム推定について検討し，システムの発展性について検討を行った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,900,000	0	1,900,000
2008 年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,100,000	360,000	3,460,000

研究分野：総合領域
科研費の分科・細目：科学教育・教育工学　教育工学
キーワード：教育工学，可視化，音声学

１．研究開始当初の背景

近年，日本の外国語教育の分野において，「読む（reading，リーディング）」「書く（writing，ライティング）」に変わり，「きく（listening，リスニング）」「話す（speaking，スピーキング）」能力が重要視されている。しかし，外国語を学習するうえで，
・母語において，あまり（あるいは全く）使用されない音素群の存在
・母語において，正確に区別されない音素対
・イントネーションや発声速度の違い
が外国語の聞き取りや，言語の取得そのものを困難にしている。
特に「スピーキング」において最も重要なポイントは，自分の発音が正しいのか，どうすれば改善できるのかといった自分の発声に対する適切な評価のフィードバックである。
現在，教師やネイティブスピーカーの代わりに，こうした点を的確に評価・アドバイスしてくれる CALL 等のソフトウェアが数多く開発されているが，これらソフトウェアが下す評価の信頼性は，使用している音声認識システムの精度に左右されるため，発話された単 語や文脈，発話者の性別や年齢層によって認識率が大きく変化するといった問題を

内包している。このような訓練者ではなくシステムに起因する問題を解消するために，サウンドスペクトログラムのように音声の周波数分析結果を画像とし表示するシステムも見られるが，専門的知識がない者がそれ自体から音韻情報を直接的に読み取る事は困難である。

２．研究の目的

これらの問題を解決するために、先に、難聴者の聴能訓練に使用した、音声の視覚情報への変換である音声画像の使用が極めて有効だと考える。この音声画像は、音声に何ら認識的手法を用いず、その音響パラメータを融合して構成したカラー画像であり、短期間の学習により容易に音素系列の読み取りが可能である事も実証されている。

そこで、本研究では、音声の画像化情報を用いる外国語発話訓練システムの開発を行う。すなわち、外国語という未知音声を学習するために、「教示用画像化音声の読み取り」、「音声の聴き取り」及び「被験者自身が発話した画像化音声の読み取り (教示用画像との比較)」を関連づけてトレーニングに導入することで、その効果の検証と利用を図ることを目的とする。

具体的には、

(1)音声画像化システムのリアルタイム処理化

(2)画像化システムに用いる音素的特徴を抽出するニューラルネットワークの多国語対応についての検討

を行う。

３．研究の方法

(1)音声画像では、逆フィルタ制御法により音声信号から抽出した第 1〜3 フォルマント周波数の巡回比をディスプレイの R/G/B 信号に変換し、母音を視覚的に判別可能な色彩の帯で表現している。以前は，訓練者の音声を録音し、いったん保存した後で画像情報へ変換する処理を行っていたため，訓練者が調音器官（口唇や舌）を細かく調整した際に，音としてどのような変化が生じるのかを瞬時に判断することは不可能であった。しかし，逆フィルタ制御法を含む音響パラメータ抽出アルゴリズムをリアルタイム処理が可能になったことから、音声の発話 (入力) から音声画像化 (出力) までをリアルタイムで実行可能なシステムの構築が図る。

(2)音声中の子音は、LPC ケプストラム、フォルマント周波数、実効値、零交差周波数そしてラウドネスレベルをニューラルネットワークに入力し音素的特徴を抽出し、各子音に割り当てたテクスチャパターンの輝度をニューラルネット出力に応じて変化させることで音韻性を表現している。現在使用しているニューラルネットは、日本語の子音の音素的特徴 (調音位置、音源、調音様式) を抽出するように構成されているため、他言語で使用するために、日本語には存在しない、あるいは正確に区別されない調音位置や調音様式についてニューラルネットを再構築し、性能の改善を行うことで，システムの多国語対応について検討する。

４．研究成果

(1)リアルタイム音声画像化システムの開発

これまでのシステムにおいて，リアルタイム処理が実現できなかった原因の一つに，音声特徴量であるフォルマント周波数の抽出に用いる逆フィルタ制御法(IFC 法)を用いることにあった。しかし，我々は DSP を用いることで，フレーム長 20[ms](12kHz サンプリング)の音声から，IFC 法により第 1〜4 フォルマント周波数を 10[ms]間隔でリアルタイム抽出することを可能にした。

さらに, 近年の汎用 PC の高性能化に伴い、DSP の処理速度に匹敵するような高速処理が可能になった。そこで，DSP 処理に用いた

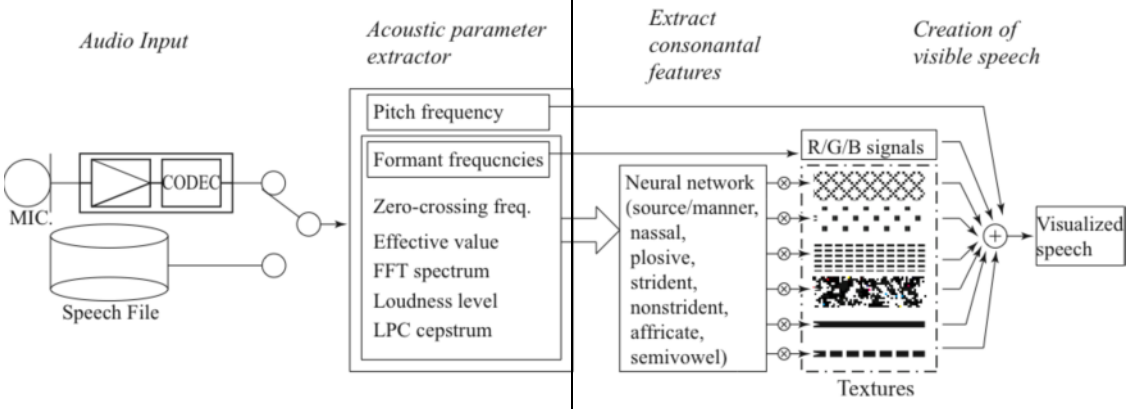

図 1．音声画像化システムのブロック図

アルゴリズムを応用することで，汎用 PC 上にアプリケーションとして音声特徴量の抽出から音声画像の生成までを実行することが可能になった。図 1 にリアルタイム画像化システムの概要を記す。

このプログラムは SDK Windows プログラムをベースに C 言語で記述されている。USB オーディオインターフェースから入力された音声は(a)A/D 変換後に 48k[Hz]から 12k[Hz]にダウンサンプリングされ，(b)音声特徴量抽出，音素特徴量抽出処理される。得られたパラメータを用いて(c)音声画像イメージが生成され，(d)電光ニュース表示のように画面に出力される。これら(a)〜(d)までの処理に要する処理時間は，Intel Celeron プロセッサ(2.66GHz)を用いた場合でも，(a)0.4[ms], (b)4.8[ms], (c)0.1[ms], (d)1.2[ms], 合計 7.5[ms]となり，フレーム長 10[ms]のに対して 75%程度の実行時間で実行可能となった。

(2)外国語発話訓練システムの画像化音声作成に用いるニューラルネットの検討

本システムでは，ニューラルネットを用いて，音声の調音様式・調音位置・声帯振動の有無に関する音響的特徴の抽出を行う。しかし，これまでに我々が開発した訓練システムは日本語ベースの発話訓練を対象としており，システムを他言語の訓練に拡張するために，ニューラルネットの再構築を行う必要がある。今回検討に使用する外国語は，英語を対象とする。

①抽出する音響的特徴と音素の対応

・音源・調音様式ネットワーク　調音様式と声帯振動の有無に関する特徴により，以下の音響的特徴を調音様式ネットワークにより抽出する。母音性（二重母音を含む），鼻音性，破裂性，　粗擦性，非粗擦性，破擦性，側音性，半母音性，有声音性，無声音性，無音性(合計 11 カテゴリ)からなる。

・鼻音ネットワーク　鼻音の調音位置に関する特徴抽出を行うネットワーク。

・破裂音ネットワーク　破裂音の調音位置に関する特徴抽出を行うネットワーク。

・粗擦音ネットワーク　粗擦音の調音位置に関する特徴抽出を行うネットワーク。

・非粗擦音ネットワーク　非粗擦音の調音位置に関する特徴抽出を行うネットワーク。

・半母音音ネットワーク　半母音音の調音位置に関する特徴抽出を行うネットワーク。

②ニューラルネットによる音響的特徴抽出}

ニューラルネットワークは，階層型ネットワークである 3 層パーセプトロンを用い，バックプロパゲーション法を用いて学習を行った。ニューラルネットの学習条件として，モーメント係数：0.8，学習率：0.2 である。

ニューラルネットの学習及びテストには，ATR が作成した「The ATR British English Speech Database」の，男女各 2 名のナレータが発話した使用頻度の高い単語（5000 語）を用いた。

この音声データに付与されている音素の開始・終了サンプル数と音韻ラベルをもとに，学習用データセットを作成した。データセットは，性別による音素毎のバランスが崩れないように考慮し，男女で音素が均等分布となるように選択し，作成したデータセットをランダムに加えながら学習を行った。その際，多様な調音現象を取り入れるため，選択された音素の全区間から任意に抜き出して使用する。

③入力フレームに関する検討

連続音声の画像化をリアルタイムで行うという目的から，時間的な遅れが生じない 1 フレームの入力に対してネットワークの出力が計算されるような構造が望ましいが，学習システムにおける画像化音声は，電光ニュース式に画像が左から右に流れていく方式であるため，子音の特徴抽出に数 10 [ms]程度の時間遅れが生じても画像の読み取りにはそれほど問題無い。また，1 フレームの入力では，破擦性の識別が困難である事から，抜き出されたフレームの前後の複数のフレームを与える事で，時間的な変化の情報も学習可能な Time-delay ニューラルネットワーク（TDNN）を採用し，入力フレームとその数について検討を行った。

選択方法は，学習の際に，選択されたフレームの前後 5 フレーム中のそれぞれのフレームを組み合わせて作成した 2 フレーム分を 1 入力ベクトルとするデータセットとしてネットワークの学習を行い，最も性能の良かったフレームを選択する。次に，残りのフレームからもう 1 フレームを追加して選択し，学習・性能評価を，ネットワークの性能の向上が見られなくなるまで繰り返す。結果として，前後のフレームより，子音から母音や子音への渡り部分に現れる特徴を学習する事で，認識率の向上が見られるが，3 フレーム以上付加に対する効果は見られなかったため，各 NN の性能が最良の状態である 2 フレーム付加の結果を図 2 に示す。

音源・調音様式ネットワークの認識結果は最大で 60[%]程度であった。この結果から，現状では不特定話者を対象とした発話訓練システムの画像化に用いるには困難であると考えられる。したがって，日本語の音声特徴量とは異なるものを用いるなど大規模な変更を行う必要があると考えらえる。

(3)音声画像フィードバックを用いた外国語学習者による日本語特殊拍の習得

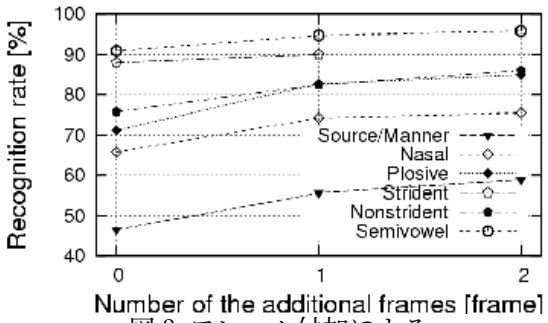

図 2.フレーム付加による
ニューラルネット識別率の変化

日本語には，特殊拍(促音，撥音，長音)と呼ばれる他の外国語にはみられない発音がある。これらを日本語特殊拍(モーラ)と呼ぶ。これらは日本語において特有の発音で，これらのモーラの正確な発声が会話内容の判断のキーとなる。

ここでは，外国人学習者が不得意とする日本語特殊拍の発話訓練を我々の開発したリアルタイム音声画像化システムにより行い，習得の程度を日本語母語話者の聴取評価(明瞭度や発話の自然性)と発話訓練の時間長測定により，リアルタイム音声画像表示機能の有効性を検討する。

実験では，事前に作成したリアルタイム画像化システムを用い，教師画像を観察しながら，訓練者が音声を入力しそのリアルタイム画像表示により，自己発声をフィードバックしながら発話訓練を行う。

①実験環境

被験者は現在熊本大学留学生で，英語を母語とする成人男性 2 名(A，B)で，次のような履歴をもつ。被験者 A は日本での留学経験 3 年目である。被験者 B は留学経験 2 年目であるが，幼少期に一年程度日本で生活した経験をもつ。

訓練に用いる単語は，「墓」「発火」や「角」「カード」など，日本語特殊拍(長音，促音，撥音)の有無で意味が異なるミニマルペア 13 組である。検証には大学生 9 名(日本語母語話者)による評価を行う。

②実験手順

０　実験の手順として，各被験者に対して，単語リストを提示し，防音室にてリストを見ながら単語を読み上げて録音を行う。その後，発話訓練システムを用いて音声画像を見ながら日本語特殊拍の発話訓練を 1 時間半程度実施した。その際に，訓練者の発話方法に対する特別な指示はせず，訓練者は自己の発話と教師音声の画像を比較しながら試行錯誤的に発話を矯正していく。訓練後に再度防音室にてリストを見ながら単語を読み上げて録音を行う。この工程を週に 1 回 4 週行い，音声画像に慣れた 3,4 週目の音声を聴取し，評価実験に用いた。

③日本人による評価実験

訓練の前後で録音した単語を日本語母語話者に対して録音した単語リストを提示して単語を聴取し，ペアのうちどちらの単語に聞こえるかを回答し正答率を算出することで，訓練よる効果を検証する。

表 1 に評定者 9 名が聴取したときの平均正答率を示す。これより，2 名の被験者とも訓練後の正答率がよくなっていることがわかる．これにより，システムによる訓練の向上がみられた。被験者 A については，日本語学習歴が長いため，あまり効果がみられなかったと考えられる。しかし，被験者からは好意的な感想が得られ，日本語学習経験の少ない話者には有効ではないかとの意見が得られた。

表 1．大学生による聴取判定平均正答率

	被験者 A	被験者 B
訓練前	93%	88%
訓練後	94%	94%

また,図 3 は特殊拍ごとに時間長を測定し，時間長の範囲と分散，平均値を表したグラフである。横軸は左から被験者 A(訓練前後)，被験者 B(訓練前後)，日本人の標準的な平均時間長である。時間長測定は，波形の目視によって長音は母音区間，促音は無音区間，撥音は鼻音区間を用いた。グラフから，訓練前後で特殊拍の有無によって時間長を変化させて発声できていることがわかる。また，同じ評定者による聴取により自然性の評価も行い，被験者 A は 60%，B は 40%となった。B の自然性が低かった理由として，図 3 より日本人の平均と比べて B の時間長が長いため，聴取しても自然な日本語に聞こえなかったことが挙げられる。

この結果より，開発した音声画像化システムにより外国人学習者に日本語特殊拍の訓練を行い，習得の程度の評価を行った。発話の向上がみられ，システムの有効性を示すことができた

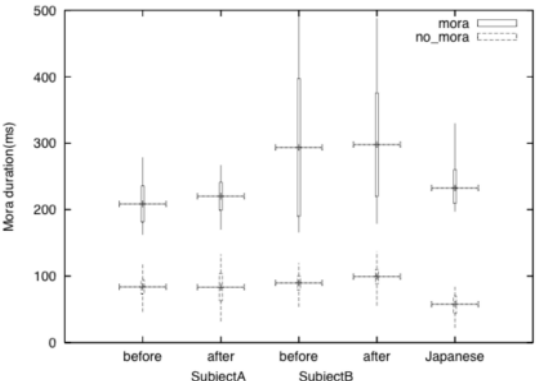

図 3. 特殊拍持続時間長の分布

(4)発話学習のためのプロソディ特徴量のリアルタイム推定

①プロソディ特徴量のリアルタイム推定

現行の発話訓練システムでは，ピッチ情報や音素持続時間の視覚化が可能であるため，文節単位・文単位での発話速度を定量化しフィードバックを行うことができる。この発話速度やイントネーションといったプロソディ特徴量は，外国語学習者だけでなく，聴覚障害者や発話障害者にとっても有益な情報となり得る。

例えば，日本語は，音節型リズムに類似したモーラ型リズムを持つ言語で，ひとつひとつの音節がほぼ同じ強さ・長さで発音される傾向（モーラ等時性）がある。英語のようなストレスタイミングの原則を優先させる言語を母語とする話者にとって，モーラ型リズムの習得は難しいことから，学習の補助的情報として，自身の発話の構音速度をフィードバックするといった利用が可能だと思われる。また，文章全体での発話速度（モーラ/秒）ではなく，分節単位での構音速度が求められれば，運動性発話障害者やパーキンソン病患者にみられる発話速度の異常を調整するリハビリテーションへの応用が可能であると思われる。

音声のプロソディ情報をリアルタイムに推定するために，さきに提案された音声特徴ベクトル推定エンジンを用いる。この推定エンジンは，フォルマント周波数や基本周波数などの音声特徴量から音素特徴・音素距離までの階層構造を持つ特徴ベクトルを分析フレーム周期（10ms）でリアルタイムに出力可能である。出力された音声特徴ベクトル要素である音声特徴量(母音性, 有声性, 無声性, 無音声）と音素距離（母音距離ベクトル）の時系列パタンに，判定アルゴリズムを用いて連続音声中のモーラ区間の検出を行う。さらに，モーラ区間から各モーラ長を算出することで，発話中の無音区間を含まない構音速度を推定する。

②長文音声を用いた構音速度推定

音声特徴ベクトルの要素から推定されたプロソディ特徴量を用いた構音速度の検出実験を行った。

使用した音声は，標準ディサースリア検査の発話速度判定に用いられる「北風と太陽」（223 モーラ）を，正常発声の成人男性 1 名が通常速度発話（Normal）と故意に発話速度を変えた変動速度発話（Variable）の 2 種類を用いる。

図 4 に，発話全体の視察によりモーラ区間を推定し得られた構音速度（Manual）と自動推定（Auto）の結果を示す。また，構音速度を算出するタイミングを，より瞬間的なモーラ単位と比較的長い文節単位の 2 種類を想定し，100[ms]と 1.0[s]の 2 種類で推定を行っている。自動推定による構音速度算出のタイミングは，構音速度が変動する場合でも，多少のずれはあるものの視察に対して自動分析が追従できることがわかる。しかし，詳細に確認すると，構音速度を検出できていない区間が見られる。これは，現段階のアルゴリズムが日本語特有の特殊モーラを考慮していないため，長音や促音，撥音の検出を行えないためであり，これらを別途検出するアルゴリズムを用意する必要がある。

③結論

リアルタイム推定される音声特徴ベクトルを用いた発話学習のためのプロソディ特徴量の推定を行った。構音速度を視覚的にリアルタイムフィードバックすることが可能であるため，ディサースリアに伴う構音速度の正常化や日本語発話学習における日本語リズムの習得に効果的であると期待される。

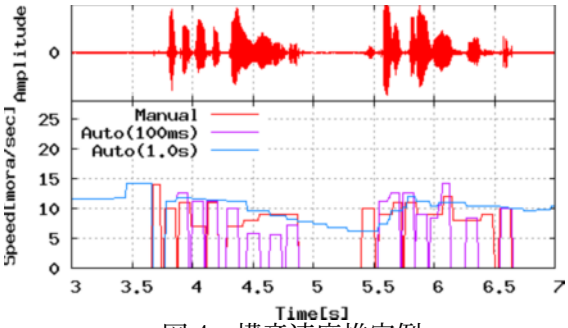

図 4．構音速度推定例

(5)まとめ

以上の結果より，本研究において開発されたリアルタイム発話訓練システムは，現状では多言語対応にできなかった。原因としては，システムの枠組みを大きく変化させずにニューラルネットの係数入れ替えのみで多数言語へ対応させるという当初計画では，多様な言語特徴を適切に抽出することができなかったためである。

しかし，当初計画からはずれはしたものの，日本語非母語話者(外国人)による日本語学習においては，主観的な評価は高く，定量的にもその効果が確認されている。また，プロソディ情報のリアルタイム推定・提示を実装することで，言語学習に限らず，難聴者や発話障害者の訓練補助ツールとして利用することも可能である。

したがって，画像化音声を用いた本訓練システムは，総合的な発話学習・訓練をリアルタイムで行う基本システムとしては評価できる結果であり，音声特徴量抽出プログラムを追加することで当初の目的であった多国語学習への応用も十分期待できると考える。

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　3 件）
① 坂田聡，外国語発話訓練システムの画像化音声生成に用いるニューラルネットの検討，日本音響学会 2008 年春季研究発表会，H20.3.19，千葉工業大学
② 米倉達郎，音声画像フィードバックを用いた外国人学習者による日本語特殊拍の習得，日本音響学会 2008 年秋季研究発表会，H20.9.10，九州大学
③ 坂田聡，発話学習のためのプロソディ特徴量リアルタイム推定法，日本音響学会 2009 年春季研究発表会，H21.3.19，東京工業大学

６．研究組織
(1)研究代表者
坂田　聡（SAKATA TADASHI）
熊本大学・大学院自然科学研究科・助教
研究者番号：80336205
(2)研究分担者
(3)連携研究者
上田　裕市（UEDA YUICHI）
熊本大学・大学院自然科学研究科・教授
研究者番号：00141961