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２６日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19700182
研究課題名（和文）　話者の音声認識エンジンに対する適応を促進するための基本技術に関する研究
研究課題名（英文）　A study on a fundamental technology to promote the speech recognition adaptation for speakers

研究代表者
中野　鐵兵（NAKANO, Teppei）
早稲田大学・ＩＴ研究機構・助手
研究者番号：90449348

研究成果の概要：

本研究では，人間の持つ非常に高度な適応・学習能力を積極的に活用した，人を音声認識器に適応させることで高度な認識精度の実現を可能にする手法の検討を行った．実利用環境における認識精度の劣化を招く主な原因として，想定話者の発話の特徴と実際の話者のミスマッチ（話者要因）が挙げられる．本研究では，話者要因に関して，入力音声の音響的特徴からより適切な発話様式を誘導するための手法の提案と，その効果の検証実験を行った．また，より適切な話者誘導を実現するために必要な，語彙依存な指示語の必要性について調査を行った. 音声認識技術のエキスパートによる指示例の収集とその分析を行い，より間違えやすい語に対する適切な指示語生成を可能にするエキスパートシステムを開発した．被験者実験を行い，その効果の検証を行った．

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	**1,300,000**	**390,000**	**1,690,000**
年度			
年度			
年度			
総　計	**3,200,000**	**390,000**	**3,590,000**

研究分野：総合領域
科研費の分科・細目：知覚情報処理・知能ロボティクス
キーワード：音声情報処理，ヒューマンインタフェース

１．研究開始当初の背景

音声認識システムを応用したアプリケーション（音声認識アプリケーション）は，ユビキタス環境におけるデータ入力の効率化，利便性の向上，老人や情報弱者に対する簡便なユーザインタフェースの提供などの観点から，強く普及が望まれている．その普及のためには，話者や周辺環境，利用タスクによらない高い認識精度の実現が要求される．これまでの音声認識技術の研究・開発の結果，話者やタスクを推定できればある程度の認識を実時間で行うことが可能となっている．しかしながら，今日まで開発されてきた技術だけでは実際の利用環境における要求精度の実現には至っておらず，実際に音声認識アプ

リケーションも十分に普及しているとは言えない．

実利用環境における認識精度の劣化を招く主な原因として，想定話者の発話の特徴と実際の話者のミスマッチ（話者要因）がある．従来の研究ではこれらの問題に対し，数多くのコーパスを作成する手法や，音声認識を使用する前に，あらかじめ対象とする話者の音声を入力することで，ユーザに合うような音響モデルを構築する話者適応法が採用されてきた．すなわち，システムがユーザに合わせるような手法が採られてきた．これらの研究成果により，今日では様々な実環境において音声認識機能の利用が可能になりつつある．一方，これらの技術が有効に働いているかどうかは利用者にとって不透明であり，認識精度が出ないときに利用者がどうすれば良いのかはまったくわからない．つまり利用者に対して，なぜ認識がうまくいかないのか，話し方が早すぎる／遅すぎるのか，入力デバイスに対して入力音声が大きすぎるのかなど，音声認識精度の劣化の原因を利用者に伝えるための枠組みと，そこで必要な要素技術が求められる．

２．研究の目的

本研究ではシステムがユーザに合わせるのではなく，ユーザがシステムに合わせる手法を検討する．すなわち，入力音声に対する適切なフィードバックを生成することで，ユーザの発話を認識されやすいような発話様式に誘導する．入力した音声の発話様式を分析し，あらかじめ学習したモデルを利用することによって，例えば“もう少しゆっくり喋ってください”とか，“もう少し大きく喋ってください”という助言を与える．これらの指示によってユーザの発話様式を変化させ，より音声認識が認識しやすいような発話様式に誘導する．このような手法によって，誤認識後の言い直し発話の発話様式が，音声認識性能が向上するような発話様式へ誘導されることを期待する．また，適切なフィードバックを生成するために，語彙非依存なフィードバックだけでなく，語彙依存なフィードバックの生成手法についての検討を行う．

具体的には，まず（１）語彙非依存な発話様式誘導モデルの確立を行う．語彙非依存な発話様式誘導を用いて，話者の音声入力を変化させることが可能であることを明らかにする．さらに，音響的特徴量や認識率との関係を分析し，それらのモデル化を行う．また，（２）語彙非依存話者誘導システムの検証を行う．分析によって得られたモデルを用いて話者誘導システムを開発する．開発した話者誘導システムを用いて発話様式を変化させ，音声認識率の改善可能性を検証する．次に，語彙依存な発話誘導の枠組みについての検討を行う．まず，（３）語彙依存な発話様式誘導モデルの確立を行う．語彙依存な発話誘導に関するデータ収集と分析を行い，汎用的な語彙依存な発話誘導方式のモデル化を行う．さらに（４）語彙依存話者誘導システムの検証を行う．語彙依存な話者誘導システムに対する，実際の話者の入力に対する影響を明らかにする．

３．研究の方法

（１）語彙非依存な発話様式誘導モデルの確立

語彙非依存な発話様式誘導モデルを確立するために，発話様式を表す要素を軸とした発話様式空間を定義し，入力音声を発話様式空間に写像するための方式を決定した．ここでは，音声認識時のリアルタイムな写像を可能にするため，入力音声から直接計測可能な特徴量のみを用いて発話様式空間を定義した．具体的には，発話様式を表わす要素として，大きさ（発話の音圧[dB]）・速度（発話速度[モーラ数/秒]）・高さ（基本周波数[Hz]）・明瞭さ（母音のスペクトル距離）に関する４つの要素を予備実験により決定した．

次に，合計５０名（男女２５名ずつ）の被験者に対するデータ収集を行った．ここでは，被験者に一つの単語につき発話様式に関する指示を３回与える．指示としては，「大きく」と「小さく」（大きさに関する指示），「速く」と「遅く」（速度に関する指示），「高く」と「低く」（高さに関する指示），「はっきりと」と「こもって」（明瞭さに関する指示）の４軸８指示語を用いた．被験者には，１回前に発話した際の発話指示に対して，与えた指示に合った発話を行うように事前教示を行った．同一内容の発話に対し，通常発声から始め，３回の指示を与え，計４回の発話を要求した．認識単語としては，指示語の認識語彙に対する依存性を排除するため，４桁の連続数字の孤立単語認識用語彙を音素バランスと予備実験から得られた各数時の認識率を考慮した上で設計した．また，収集されたデータに対して，各々の指示語に対する特徴量の変化，認識率の変化の相関を分析した．

モデルとしては，入力音声が与えられた時に，認識率が改善するような指示語の識別モデルを構築した．認識誤りを減らすような発話様式へ誘導するための，発話様式空間上の遷移先を表す，ターゲット発話様式の決定を試みた．しかしながら，発話様式空間内におけ

る認識性能の高いサンプルから成る部分空間の識別が困難であったため，サンプルベースで指示語を選択するようにした．フィードバックの生成には，発話様式空間上での現在までの入力音声と，ターゲット発話様式との相対的な位置関係を用いる．発話様式のフィードバック方式は，ターゲット発話様式に遷移するような指示を直接ユーザに与えることで，ユーザが指示にしたがって発話することを期待する．

（２）語彙非依存話者誘導システムの検証

収集したデータと構築したモデルを用いて，指示の有効性評価実験を行った．有効な指示の選択をパターン認識の枠組みとして実現する．識別器としては support vector machine（SVM）を利用した．実験では学習データに使用する音声データの分量を２０名から４０名まで変化させ，交差検定を行った．SVM に関しては，線形カーネル，多項式カーネル，ガウスカーネルの３種類を用いた．SVM の特徴ベクトルとしては，音圧，話速，基本周波数，滑舌度（計４次元），単語事後確率（１次元），認識結果（１次元），指示の履歴（８次元），指示内容（８次元）を用いた．指示語が有効であったかどうかに関するラベルとして，指示語有効ラベルを定義した．ここでは，発話の単語認識率や単語事後確率から，与えた指示の有効であったか否かを表現した指標を，指示有効性と呼ぶ．指示有効性は，誤認識した発話に対して指示を与えることで，正しく認識させることができれば○を，また正しく認識していた発話が指示を与えた後に，誤認識を起こした場合×にする．また，認識結果が指示を与えても認識成功のまま，または認識誤りのままの場合，事後確率が増加すれば○，減少すれば×とした．

（３）語彙依存な発話様式誘導モデルの確立

語彙に依存したな発話様式誘導モデルを確立するために，まず音声認識技術のエキスパートによる指示例の収集した．ここでは，発音しにくいことばの音声要因（ＮＨＫ放送文化研究所）より，音素バランスを考慮して語彙（合計１２８単語）を設計した．データ収集システムは被験者に単語を提示し，被験者に音声入力を促す．最低２回音声入力を行い，２回連続で音声認識が成功した場合には次の単語を提示する．一度でも認識に失敗した語に対しては，１回目の発話時にどのような工夫を行ったか？，２回目以降の発話時にどのような工夫を行ったか？あなたなら素人にどのように指示するか？に関するアンケートを自由記述形式で得た．実験はビデオカメラにて撮影を行い，入力時の口もとの変化も観察した．

次に，アンケートの結果を分析し，語彙に依存した指示語を抽出した．ここでは，アンケートの結果を指示語の種類や指示対象となる音素の位置を基準に手動で分類し，指示語の軸となる語を得た．分析によって得られた指示語をベクトル表記し，各々の認識単語に対して指示ベクトルを関連付けた．認識単語と指示ベクトルを関連付けるために，単語を子音＋モーラ単位で分割し，１つの単語を複数のユニットで表現できるようにデータを構築した．各々の単語ユニットに対して，指示対象の子音・母音の位置情報を含む指示ベクトルを割り当てた．さらに語彙と指示ベクトルとの関係を分析し，語彙を入力とした時の指示語に関する決定木を作成した．すなわち，単語を入力としたときの，指示語の有無と指示語がある場合のその具体例の生成を可能にした．実験はビデオカメラにて撮影を行い，遠隔操作が可能な環境で行われた．

（４）語彙依存話者誘導システムの検証

構築した決定木を用いて，指示語を与えた時の効果の検証実験を行った．ここでは，語が入力として与えられるとその語に対する指示語を出力する指示語生成システムを開発した．実験では，被験者は音声認識実験を３セッション行う．第一セッションでは４０語，第二セッションでは４０語，第三セッションでは６０語，それぞれ一語あたり３回の音声入力を行う．そのうち，第二セッションのみ，生成された指示語を提示する．認識語彙としては，システムによって指示語が出力される語彙のうち，音素表記の重複がない６０語を選択した．それらを，全てのセッションにて使用される２０語（グループＡ），第一・第三セッションのみに使用される２０語（グループＢ），第二・第三セッションのみに使用される２０語（グループＣ）に分類した．実験では，それぞれのグループ・セッションによって認識率がどのように変化するかを検証した．

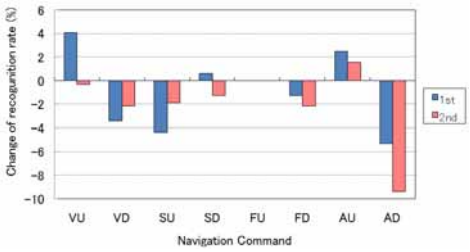
４．研究成果

＜研究の主な成果＞

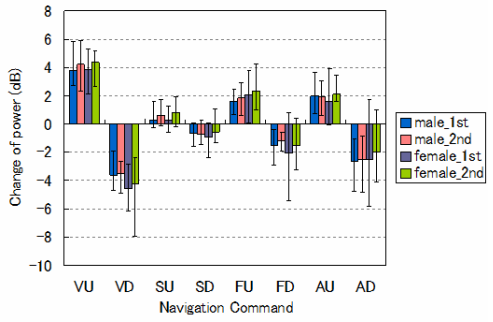
（１）語彙非依存な発話様式誘導モデルの確立

指示語を与えた時の音声入力の分析により，指示語と入力音声の音響的特徴量の相関関係を明らかにした．まず，指示に従って発話様式が変化することで，単語認識率が変化す

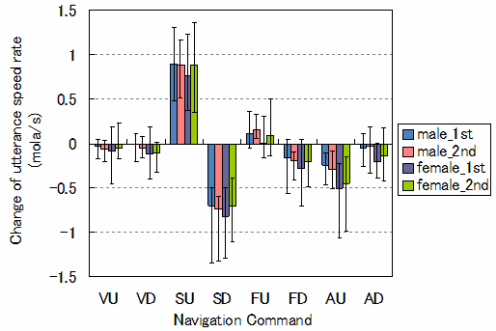
ることが確認された．大きく（VU）やはっきり（AU）は，認識率の向上が見られるが，小さく（VD）やゆっくり（SD），こもって（AD）等は認識率が減少した．また，これらの変化は，認識が失敗→成功となった語と，成功→失敗となった語が相互に存在し，その数の大小にて起こっていたことが確認された．



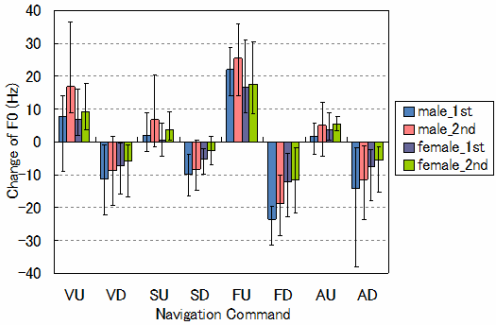
音圧に関する変化に関しては，大きく（VU）及び小さく（VD）に対する音圧の変化量が，他の指示と比較して大きいことが確認された．すなわち，声の音圧を変動させる指示を与えることで，指示通りに声の音圧を変化させることが容易であることがわかった．また，音高や滑舌度に関する指示を与えたときでも，音圧の変化が起こることが確認された．



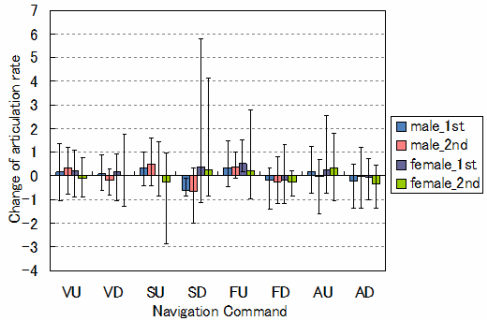
速度に関する変化に関しては，早く（SU）及びゆっくり（SD）に対する発話速度の変化量が，他の指示と比較して大きいことが確認された．すなわち，発話速度を変化させるような指示を与えることで，指示通りに発話速度を変化させることが容易であることがわかった．また，音圧や音高に関する指示では，発話速度の変化はほとんど見られないが，“はっきり”と言う指示を与えたときには，発話速度が遅くなることが確認できた．



基本周波数（F0）の変化に関しては，高く（FU）及び低く（FD）に対する音高の変化量が，他の指示と比較して大きいことが確認された．すなわち，声の音高を変動させるような指示を与えることで，指示通りに声の音高を変化させることが，容易であることがわかった．特に男性と女性を比較すると，ほとんどの指示に関して，女性の変化量が大きいことも確認できた．

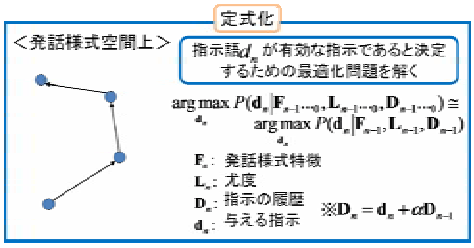


滑舌度の変化に関しては，指示語と音響的特徴量の明示的な相関が得られなかった．“はっきり”という指示に対して滑舌度の上昇が得られるケースも存在するが，“こもって”という指示に関しても，被験者によっては上昇するケースが見られた．この原因として，“こもって”という指示が，発話様式を変化させるという点では直感的でなく，非常に曖昧である点が挙げられる．実際に，被験者が“こもって”という指示を与えられたときに，発話様式の変え方が分からず，戸惑っている場面が見られた．そのために，“はっきり”と“こもって”の指示に対する発話様式の変え方について記述式のアンケートを実施した．アンケートの結果によれば，これらの違いは，口の開け方や動かし方に関連していることが分かった．これらの特徴をより正確に得るための特徴量の抽出方法が別途必要であることが示唆される．



これらの結果から，音圧，話速，基本周波数，滑舌度は間接的に音声認識の認識率に影響を与える変数と考えられ，下記の定式化を行

った．このとき，発話様式を表わす特徴量として，音圧，話速，基本周波数，滑舌度を使用する．



（２）語彙非依存話者誘導システムの検証

収集したデータを用いて学習した識別器による，指示の有効性評価実験を行った．実験では，線形カーネル，多項式カーネル，ガウシアンカーネルを使用した場合で比較を行ったが，ガウシアンカーネルを使用した，２０名を学習データとした時の交差検定における結果が最も良く，識別率は６５％であった．学習に使用する人数の増加による効果は得られず，また，使用する特徴量を変更することによる効果も得られなかった．指示の履歴情報を用いることの効果は，各カーネルに関して少量であるが指示有効性の上昇として表れた（最大２．３％の指示有効率の向上）．しかしながら，今回の評価方法は○か×の二者択一制であり，実験の結果からは提案手法の有効性を確認できたとは言えなかった．

本実験結果から得られた仮説としては，認識率に支配的な影響を与える発話様式の変化は語彙非依存に決定されるのではなく，語彙に依存した方式で決定されるという点である．すなわち，発話速度や高さを変化させることによって発生する認識率の変化は，発話する語に含まれる子音の種類や位置，母音の数等に大きく依存するという仮説である．これは，認識が容易な語（１００％であった語）では，どのような指示を与えても認識率が変化しなかったのに対し，誤認識が発生する語の場合は，発話様式を変化させることで認識率が変化しやすいという点が，予備実験を含む本研究にて行われた数多くの実験においても多く見られたことから仮説として立てられたものである．

（３）語彙依存な発話様式誘導モデルの確立

収集した４名のエキスパートの指示例の分析を行った．実験では，より多くの指示例を得やすくするために，音声認識器の設定を誤認識が多くなるように行った．認識率はそれぞれ，６２．５％，４５．３％，６８．８％，７２．７％であった．抽出されたルールとしては，出現頻度の多い順に，強度（弱く，やや弱く，やや強く，強く），速度（ゆっくり，ややゆっくり，やや速く，速く），明瞭性（ぼかす，ややぼかす，ややはっきり，はっきり，意識する，口を大きく），長さ（短く，やや短く，やや長く，長く），連続性（一瞬区切る，つながりを丁寧に，連続するように，一息で），高さ（低く，やや低く，やや高く，高く），テンション維持（やや音量を保つ，音量を保つ），強調（やや強調，強調，とても強調），丁寧さ（丁寧に，抜けないように），特に（特に），という指示が得られた．これらの指示は主に先頭の子音や，濁音・促音・拗音等に対して行われており，語全体に対して指示語が与えられる例は少なかった．

データ収集に利用した認識単語を単語ユニットに分割し，それぞれ位置情報を併せ持つ指示ベクトルとの関連づけを行い，決定木分析を行った．分析のアルゴリズムとしてＣ４．５アルゴリズムを用いた．分析対象のデータとしては，指示ベクトルを持たないユニットの数と指示ベクトルを持つユニットの数が均一でない（１０：１程度）ため，決定木を階層的に作成した．結果としては，clarity（はっきり）と intensity（強く、大きく）と speed（速さ）に関する木が生成された．

（４）語彙依存話者誘導システムの検証

構築した決定木を用いて，指示語を与えた時の効果の検証実験を行った．被験者は１０名，全て２０代男性である．セッション毎の認識率の平均とその標準偏差はそれぞれ，{（０．８９５，０．０５１），（０．８８７，０．０６９），（０．８９７，０．０８９）}となり，分散分析の結果より統計的有意差は得られなかった（p=０．８２２２）．グループ毎の語彙に関しても，グループＡでは，{（０．８８３，０．０５８），（０．８９０，０．０７５），（０．８８０，０．１１６）}，グループＢでは，{（０．９０７，０．０５６），（０．９１７，０．０９４）}，グループＣでは，{（０．８８５，０．０７２），（０．８９３，０．０７２）}となり，いずれも統計的優位は得られなかった．これらの結果から，提案手法に基づく発話様式誘導システムでは，話者の発話様式を認識率に対して影響を与えるという観点からは効果がないことが示された．すなわち，音声認識技術のエキスパートによって得られた，提案手法に基づく語彙依存な発話様式に関する指示語作成のエキスパートシステムでは，人を音声認識器に適応させるために有効な指示を出すことが出来ないことが明らかとなった．

＜まとめと今後の展望＞

本研究では，人を音声認識器に適応させる手法として，１．入力された音声の音響的特徴のみから，語彙に依存しない形式でフィードバックを与える手法と，２．語彙に依存した発話様式に関する指示を出すことで，より適切な発話様式を人が学習するための手法について，それぞれモデル化とシステムの実装，検証実験を行った．しかしながら，いずれの手法も効果的な手法としては実証できなかった．すなわち，人を音声認識器に適応させる手法としては本手法の抜本的な修正が必要である．本研究の成果により明らかになったように，語彙依存・非依存のいずれの指示に対しても，話者はおおよそ期待通りに発話様式を変化させることが可能であり，それに伴う音響的な特徴量の変化も発生する．しかしながら，それらの特徴量によって認識率の変化を説明することは困難であり，現在まで明らかになっている結果からは，発話様式の指示に関するシステムを構築することはできない．今後の展望としては，認識率を直接・もしくは間接的に説明可能な支配的な変数の発見が求められる．例えば，音声データだけでなく，画像入力も併用した新たな変数などである．これらの変数が見つかれば，本研究で行ったアプローチにて発話様式誘導システムの構築が可能になることが期待される．

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）

６．研究組織
(1)研究代表者
中野　鐵兵（NAKANO　TEPPEI）
早稲田大学・ＩＴ研究機構・助手
研究者番号：90449348