様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　２ 年　 ６ 月　 ４ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００２２４

研究課題名（和文）誤認識原因の識別と通知に基づく音声認識のユーザビリティ改善

研究課題名（英文）Usability Improvement of Speech Recognition Based on Classification and Notification of Recognition Error Causes

研究代表者
山田　武志（Yamada, Takeshi）

筑波大学・システム情報系・准教授

研究者番号：２０３１２８２９

交付決定額（研究期間全体）：（直接経費）　　2,500,000円

研究成果の概要（和文）：実利用における音声認識の性能はユーザの話し方によって大きく変動する。しかし、一般のユーザにとって、このような性能変動を的確に把握することは極めて難しい。そこで本研究では、ユーザが発話した音声を正しく認識できるか否かを判断し、認識できないと判断した場合には、その原因を識別してユーザに分かり易く通知する手法を開発した。まず、高精度な認識成否判断と誤認識原因識別を実現するために、変調スペクトルと深層ニューラルネットワークを用いた手法を提案し、その有効性を確認した。そして、誤認識原因をユーザに通知するためのインタフェースを設計してＰＣ上に実装した。

研究成果の学術的意義や社会的意義
本研究では、ユーザが発話した音声を正しく認識できるか否かを判断し、認識できないと判断した場合には、その原因を識別してユーザに分かり易く通知する手法を開発した。このような機能は本来ユーザインタフェースの一部として備わっているべきであるが、音声認識においてはこれまで実現していなかった。本研究成果により音声認識のユーザビリティが大きく改善し、音声認識サービスのさらなる普及につながると期待できる。また、本研究を通して既存の技術では認識が難しい音声特徴が明確になり、音声認識技術のさらなる高精度化を図るための指針を得た。

研究成果の概要（英文）：The performance of speech recognition in actual use varies drastically depending on how the user speaks. However, it is difficult for general users to accurately grasp such performance fluctuations. Therefore, in this research, we have developed a method to judge whether or not a user's utterance can be correctly recognized, and if it is judged that it cannot be recognized, classify the cause and notify the user in an easy-to-understand manner. First, in order to realize the judgement and classification, we proposed a method using a modulation spectrum and a deep neural network, and confirmed its effectiveness. Then, an interface for notifying the user of the cause of recognition error was designed and implemented on a PC.

研究分野：音声情報処理学

キーワード：音声認識　ユーザビリティ　認識性能推定　誤認識原因識別　発話特徴　変調スペクトル

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

音声認識の研究分野における主要な研究目標の一つは、いかなる条件においても高精度な認識を実現することである。最近の成果としては深層学習の導入というブレークスルーがあり、認識率の劇的な改善が達成された。しかし、高騒音環境や自然発話に対する認識率は依然として低いのが現状である。

このように実利用における音声認識の性能は、周囲の音環境（雑音の特性や音量、残響の程度など）やユーザの話し方（発音、声量、発話速度など）によって大きく変動する。しかし、一般のユーザにとって、このような性能変動を的確に把握することは極めて難しい。そのため、誤認識が起こったときにその原因が分からず、正しく認識されるまで何度も繰り返し発話するといった苦労を強いられてしまう。

一方、ユーザに正しく認識される方法を適切に伝えることにより、誤認識を大幅に削減できることが知られている。例えば、早口であるときには、話すスピードを少し遅くするようにアドバイスするだけで正しく認識されるようになる。しかし、このようなユーザ主導(ユーザ側の対応)による認識率の改善を支援する研究は世界的に見てもほとんど行われていない。

２．研究の目的

本研究では、ユーザが発話した音声を正しく認識できるか否かを判断し（認識成否判断)、認識できないと判断した場合には、その原因を識別して（誤認識原因識別）ユーザに分かり易く通知する手法を確立することによって、音声認識のユーザビリティ改善を図る。このような機能は本来ユーザインタフェースの一部として備わっているべきであるが、音声認識においてはこれまで実現していなかった。本研究成果により音声認識のユーザビリティが大きく改善し、音声認識サービスのさらなる普及につながると期待できる。

３．研究の方法

(1) 認識成否判断手法の開発

利用できる情報はユーザ発話の音響信号のみであり、ユーザの発話内容を表す正解テキストは未知であるので、正解テキストと認識結果との比較によって認識の成否を判断することはできない。よって、認識成否に対応する何らかの尺度を設定する必要がある。そのような尺度の一つとして認識性能（認識率）がある。

従来、認識結果に対し付与された信頼度を用いて、確率的に誤りタイプ分類を行って認識性能を推定する手法が提案された。この手法では、認識結果文中の各単語を正解または 3 種類の不正解（不正解、挿入、欠落）に確率的に分類して認識性能を推定する。しかし、この手法は音声認識を実際に行う必要があるため、計算負荷が高いという問題がある。一方、入力発話から抽出した音響特徴量のみを用いて認識性能を推定する手法が提案された。これは、入力発話全体から各種特徴量の統計量を抽出し、SVR（support vector regression）を用いて認識性能を推定する。しかし、発話が短い場合に有効な統計量を算出することが難しいので、推定精度が低下するという問題点がある。

そこで本研究では、時間フレームを単位とする音響特徴量を用いて認識誤り区間を推定するというアプローチを検討する。認識性能は、発話全体のフレーム数と認識誤り区間のフレーム数の比に基づいて算出することができる。また、認識性能のみではなく、認識誤り区間が分かるので、後段の誤認識原因識別の適用の際に有用である。本研究では、これを実現するために、音響特徴量として変調スペクトル、推定器として BLSTM（bidirectional long short-term memory）を用いた認識誤り区間推定法を提案し、その有効性を検証する。ここで、変調スペクトルは各種特徴量の時間軌跡のスペクトル表現として定義される。認識誤りの原因の多くは発話速度とその変動に関係していることから、認識誤り区間の推定に適していると考えられる。さらに、推定モデルとして CRNN（convolutional recurrent neural network）を用いることを検討する。CRNN は、畳み込みニューラルネットワークと再帰型ニューラルネットワークを統合したものであり、変調スペクトルのような 2 次元特徴マップからなる時系列データの分析に適している。

(2) 誤認識原因識別手法の開発

音声認識における誤認識は様々な原因によって起こるが、本研究では特にユーザ側で対処できる話し方（発話特徴）に注目する。ここで、誤認識の原因となる発話特徴としては、発音、発話速度、フィラー、言い淀みなどが挙げられる。これらの発話特徴は、ユーザにフィードバックし易く、同時にユーザにとって比較的容易に改善できると考えられる。

そこで本研究では、早口、遅口、フィラー、言い淀みという発話特徴を識別するために、(1)の手法と同様に、音響特徴量として変調スペクトル、識別器として BLSTM を用いた発話特徴識別手法を提案し、その有効性を検証する。さらに、誤認識は発話全体ではなく発話の一部に局所的に存在することに着目し、識別器である BLSTM にアテンションという注視機構を追加することを検討する。

(3) 誤認識原因をユーザに分かり易く通知する手法の開発

① 雑音に起因する誤認識に対する手法

誤認識を回避することができる適切な発話音量をリアルタイムに推定し、ユーザに通知する

手法を提案し、実携帯端末を用いたユーザ評価を含む有効性評価を行う。

② 話し方に起因する誤認識に対する手法
誤認識原因となる発話特徴をユーザに通知するインタフェースを設計してＰＣ上に実装する。

４．研究成果

(1) 認識成否判断手法の開発

提案手法の有効性を検証するために、認識誤り区間を推定する実験を行った。本実験では、宇都宮大学パラ言語情報研究向け音声対話データベース、重点領域研究「音声対話」対話音声コーパス、音声対話データベース 96 年版から、男性と女性の各 25 名の計 5,415 個の音声データを用いた。ここで、各音声データから認識誤り区間と認識正解区間のフレーム数の比率が 1:1 になるように不連続フレームが生じないように切り出しを行った。本実験では、単語信頼度を用いた手法（音声認識エンジン Julius が出力した単語信頼度と閾値を比較して単語単位で推定する）、BLSTM を用いた手法（提案手法）、CRNN を用いた手法（提案手法）、CNN を用いた手法（CRNN から RNN 部を取り除いてフレーム単位で推定する）を比較した。

図 1 に各手法の F 値を示す。

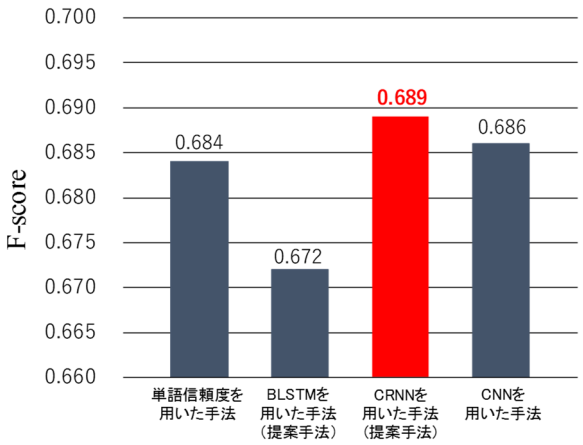

図 1：各推定手法の F 値

図 1 のように CRNN を用いた手法の F 値が最も高いことが分かった。一方、CNN を用いた手法との差は僅かであることから、CRNN を用いた手法においては RNN 部よりも CNN 部の方が性能改善に寄与していると考えられる。また、CRNN を用いた手法は音響特徴量のみから推定しているにもかかわらず、音声認識の結果を利用するという点で有利である単語信頼度を用いた手法を僅かながら上回っていることが分かった。

(2) 誤認識原因識別手法の開発

提案手法の有効性を検証するために、誤認識原因を識別する実験を行った。本実験では、日本語の対話音声データベースである重点領域研究「音声対話」対話音声コーパスと宇都宮大学パラ言語情報研究向け音声対話データベースを用いた。話者は男性 10 名、女性 10 名であり、これらの話者の音声データの中から音声認識エンジン Julius によって誤認識された 1,400 個の音声データを用いた。ここで、音声データの長さは 1～10 秒である。誤認識された音声データには発話特徴のラベルを付けた。ここで、フィラーと言い淀みについては、各音声データベースに付属の書き起こしテキストに従ってラベルを付けた。一方、早口と遅口については実際に聴取し、人手でラベルを付けた。

表 1 に各識別手法の F 値を示す。

	MFB+BLSTM	MS+BLSTM	MS+BLSTM+ATTENTION
	音響特徴量：メルフィルタバンク出力 識別器：BLSTM	音響特徴量：変調スペクトル 識別器：BLSTM	音響特徴量：変調スペクトル 識別器：BLSTM with Attention
早口	0.572	0.596	0.610
遅口	0.576	0.595	0.705
フィラー	0.601	0.675	0.686
言い淀み	0.601	0.666	0.691

表 1：各識別手法の F 値

まず、MS+BLSTM は MFB+BLSTM よりも高い F 値が得られることが分かった。これは、音響特徴量として静的なスペクトル特徴量であるメルフィルタバンク出力よりも、その時間変動をスペクトル表現した変調スペクトルの方が適していることを意味している。次に、MS+BLSTM+ATTENTION は MS+BLSTM よりもさらに高い F 値が得られることが分かった。これは、アテンションという注

視機構が有効に機能していることを意味している。図 2 に示すように、MS+BLSTM+ATTENTION により言い淀みの区間を適切に捉えていることを確認できた。

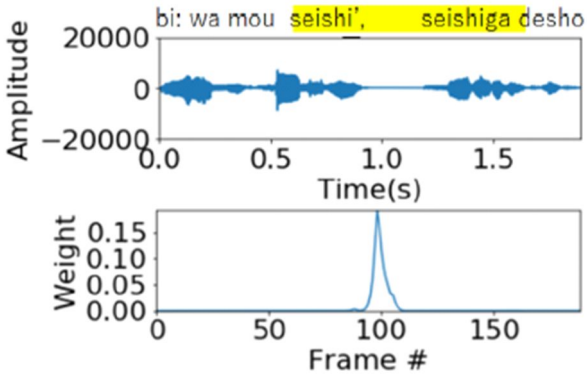

図 2：提案手法による言い淀み区間の検出結果の例

(3) 誤認識原因をユーザに分かり易く通知する手法の開発

① 雑音に起因する誤認識に対する手法

携帯型の端末を様々な雑音環境に持ち運んで音声認識を利用することを想定し、提案手法をタブレット端末（Google 社の Nexus7）上に実装した。そして、被験者評価を含む有効性評価を行った。提案手法はあくまで音声入力の補助機能であるため、提案手法の有効性を確認するために、音声認識を行って結果を表示するというシンプルなアプリケーションを作成した。被験者は、雑音環境において机の上に置いたタブレット端末から少し離れた位置で単語リストにある 10 個の単語をそれぞれ発話する。その際、全ての単語が正しく認識されるまで発話を繰り返し行い、発話が終了したときの総発話回数と認識率を算出する。提案手法を用いない場合と用いる場合の総発話回数と認識率を表 2 に示す.

	提案手法を用いない場合	提案手法を用いる場合
駅ホーム（約0dB）	64回/78%	61回/82%
駅ホーム（約-10dB）	80回/63%	70回/71%
百貨店ホール（約0dB）	80回/63%	70回/71%
百貨店ホール（約-10dB）	135回/37%	78回/64%

表 2：提案手法を用いない場合と用いる場合の総発話回数と認識率

提案手法を用いることによって総発話回数が総じて少なくなっており、潜在的な誤認識を削減できることを確認した。

② 話し方に起因する誤認識に対する手法

誤認識原因となる発話特徴をユーザに通知するインタフェースを設計してＰＣ上に実装した。実装したインタフェースのスクリーンショットを図 3 に示す。

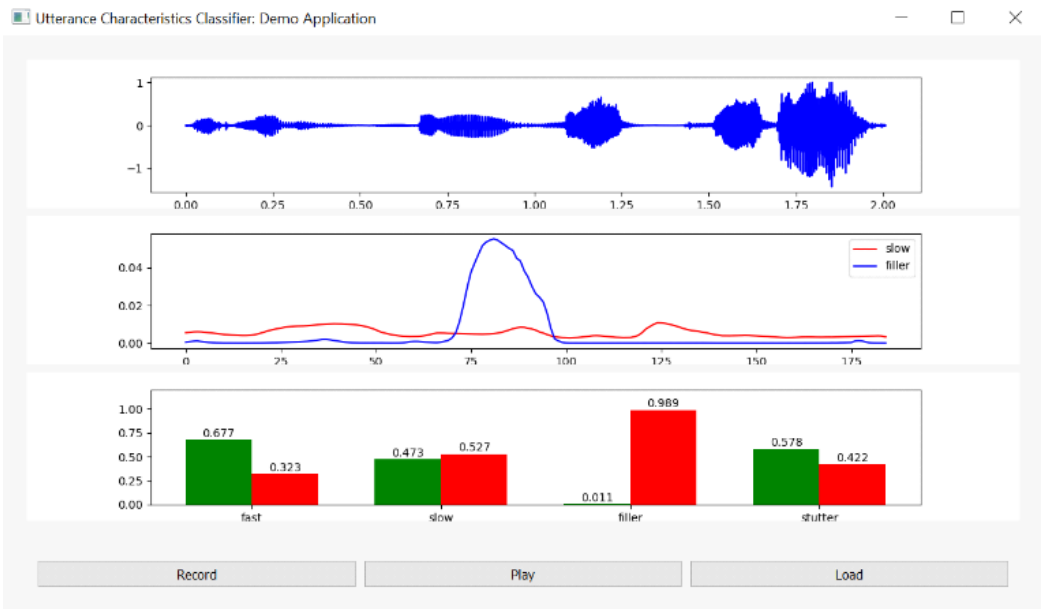

図 3：実装したインタフェースのスクリーンショット

本インタフェースでは、最下段において各発話特徴の非存在確率（緑）と存在確率（赤）、及び中段において存在確率が非存在確率を上回っている発話特徴の時間区間をユーザにグラフィカルに提示する。これにより、ユーザは自身の発話を速やか、かつ容易に改善することが可能となる。

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計7件（うち招待講演　0件／うち国際学会　3件）

１．発表者名 Jennifer Santoso, Takeshi Yamada, Shoji Makino
２．発表標題 Classification of causes of speech recognition errors using attention-based bidirectional long short-term memory and modulation spectrum
３．学会等名 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference 2019（国際学会）
４．発表年 2019年

１．発表者名 サントソ ジェニファー，山田武志，牧野昭二
２．発表標題 BLSTMと変調スペクトルを用いた発話特徴識別の検討
３．学会等名 日本音響学会2019年秋季研究発表会
４．発表年 2019年

１．発表者名 舒禹清，山田武志，牧野昭二
２．発表標題 BLSTMを用いた音声認識誤り区間推定の検討
３．学会等名 日本音響学会2019年秋季研究発表会
４．発表年 2019年

１．発表者名 舒禹清，山田武志，牧野昭二
２．発表標題 発話の時間変動に着目した音声認識誤り区間推定の検討
３．学会等名 日本音響学会2020年春季研究発表会
４．発表年 2020年

１．発表者名
Jennifer Santoso, Takeshi Yamada, Shoji Makino

２．発表標題
Categorizing error causes related to utterance characteristics in speech recognition

３．学会等名
RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2019 (NCSP'19)（国際学会）

４．発表年
2019年

１．発表者名
Takahiro Goto, Takeshi Yamada, Shoji Makino

２．発表標題
Novel speech recognition interface based on notification of utterance volume required in changing noisy environment

３．学会等名
RISP International Workshop on Nonlinear Circuits, Communications and Signal Processing 2018 (NCSP'18)（国際学会）

４．発表年
2018年

１．発表者名
後藤孝宏，山田武志，牧野昭二

２．発表標題
音声認識における誤認識原因通知のための印象評定値推定の検討

３．学会等名
日本音響学会2018年春季研究発表会

４．発表年
2018年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考