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 27 日現在

研究種目：若手研究（B）
研究期間：2006 ～ 2008
課題番号：18700166
研究課題名（和文）　実環境を想定したオンラインによる音響モデルの構造化に基づく頑健な音声認識
研究課題名（英文）　On-line Speech Recognition Based on Structuring of Acoustic Models for Real Environments

研究代表者
氏　名（ローマ字）：西田 昌史（Nishida Masafumi）
所属機関・部局・職：千葉大学・大学院融合科学研究科・助教
研 究 者 番 号：80361442

研究成果の概要：本研究では、家庭内や車内といった身近な環境を想定し、話者や環境が限定できる場面における頑健な音声認識の枠組みについて検討を行った。このような環境では、閉じた複数の話者や雑音が繰り返すことが普通であることに注目し、オンラインで話者と雑音の特徴変化を音響的に同一処理で自動学習することで、複数話者・雑音の学習と認識を統一的に処理する新たな音声認識の枠組みを実現した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	500,000	0	500,000
2007 年度	900,000	0	900,000
2008 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,000,000	180,000	2,180,000

研究分野：音声情報処理，パターン認識，ヒューマンインタフェース，福祉情報工学
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：音声認識，環境適応，話者適応，強化学習，音響モデル，クラスタリング

１．研究開始当初の背景

近年、学会講演などを対象とした音声のディクテーションや音声対話などをタスクとした話者や雑音の変動に頑健な音声認識手法についてさかんに研究が行われている。

このような従来研究で想定される環境は、データごとに話者は一人であり、データごとに話者が異なっていることが明確である。また、雑音に関してもデータ中に雑音は一種類であり、データごとに雑音が異なることは明確である。

こういった環境を対象とした研究のアプローチの一つとして、話者や雑音の変動に音響モデルを適応化する手法がある。従来の適応に関する研究では、いかに少量の音声データで話者に適応するか、事前に雑音データを収集し雑音をモデル化して最適なモデルを選択するといった方法が検討されている。

このような手法では、音声認識システムを使用する環境に応じて、事前に話者や雑音のデータを収集し適応化させる必要があり、話者や雑音のモデル化と認識処理が独立に処理されているため、汎用的ではなくコストが

かかる。
さらに、音声認識システムが使用される環境は、話者や雑音も様々なものが考えられ、適切な事前対応が困難であり、音声認識を実用化する上で大きな問題点であると考えられる。そういったことから、実環境を意識した音声認識の枠組みについて検討する必要がある。

２．研究の目的

最近では、カーナビゲーションシステムや介護ロボット、情報家電のインタフェースとして音声認識のニーズが高く、より身近な環境で使用されるようになってきている。
こういったタスクでは、使用する話者としては家族、環境としては車内や家といったように、使用する環境は事例ごとには限定することができる。システムを利用する話者は一人とは限らず、話者が頻繁に変わったり同じ話者が繰り返し使用したり、テレビがついていたり消えていたり、音楽が流れていたり流れていなかったりと、環境も繰り返し変化している。
このような実環境においては、複数の話者・雑音へ対応する必要がある。しかし、事前に各話者や雑音データを収集し登録することは困難である。そこで、本研究では誰が話しているか、どんな雑音であるかを検出することなく、使用現場でオンラインにより話者と雑音の変化を自動学習しモデルを強化していく。さらに、学習したモデルを効率的に構造化することで、過去に学習したモデルを利用することができる頑健な音声認識を実現することを目的とした。

３．研究の方法

(1) 初年度は、これまで自らが検討した連続数字の認識における強化学習による環境適応手法を大語彙連続音声認識に拡張し、音響モデルの自動学習について検討を行った。
強化学習では、実行した行動に対して報酬が得られ、その報酬を最大化するような方策を学習する理論であり、行動、行動を実行したことによる状態の価値、報酬といったものを定義する必要がある。
これまで検討してきた手法では、環境への適応を MAP 推定により行い適応の割合を行動として、フレームごとに最尤な認識結果と発話全体で最尤な認識結果の数字列を比較し、その一致度を状態の価値、適応前に比べて適応後の認識結果の一致度が高くなれば大きな報酬を与えることで、適応の割合を学習した。
大語彙連続音声認識へ拡張するため、状態の価値としてフレームごとと発話全体での最尤な音素列をそれぞれ抽出し比較を行う。これにより、正しく認識ができていない場合は比較した音素列の一致度が低くなり、話者や環境が変化したと判断し適応を行うことができる。
本研究では、環境への逐次適応として MAP 推定による手法に着目し、音響モデルのパラメータに対する適応の割合を行動とした。具体的には，11 種類の MAP 推定における重み係数を行動として定義した。また、環境の変化を初期モデルと適応モデルとの尤度比により表現することで、Q-learning により行動の価値を推定した。その際、尤度比の値を 4 つの領域に離散化することで状態を定義した。これらに基づいて，環境に応じた適応の度合いを学習した。
以上の手法の有効性を確認するために、日本語話し言葉コーパスを用いた音響モデルの自動学習について評価実験を行った。

(2) 二年度目は、話者や雑音の変化に応じて音響モデルをオンラインでクラスタリングすることで、効率的な音響モデルの学習について検討を行った。
オンラインでの音響モデルのクラスタリングについては、強化学習によりモデルを適応していき、上で述べたフレームごとと発話全体での認識結果の一致度が適応前後での変化量により話者や環境が変化したと判断した。この際、音響特徴から認識結果の一致度を判断基準としているため、誰がしゃべっているか、どんな環境なのかということを検出することなく、環境の変化を判断できるという利点がある。
次に、現在の環境は過去に学習したのか未知なものかを判断することで、過去に学習したものであればそのモデルを選択することで頑健な認識が可能となり、さらにモデルを強化していった。もし未知な環境であれば、モデルを新たに学習していくことができる。現在の環境が既知であるか未知であるかは、音素ごとに音響モデルである HMM の各状態の平均ベクトルを束ねて一つのベクトルとして音響空間にマッピングし、音素間のベクトルによるユークリッド距離によりモデル間の類似度を定義し、判断した。つまり、この距離が過去のすべてのモデルに対して一定値以上離れていれば、未知な環境であると判断した。
以上の手法の有効性を確認するために、雑音環境下連続数字音声データベースを用いて認識実験を行った。

(3) 三年度目は、音響モデルを構造化することで頑健な音声認識を実現するために、音響モデルの選択手法ならびにそれを踏まえた話者適応について検討を行った。
前年度において検討した手法では、音響モデルの各状態における平均ベクトルを束ねて

一つのベクトルにして音響空間にマッピングし、それらのユークリッド距離によりクラスタリングを行った。それに対して、モデル間の類似度をより効率よく算出するために、音響モデルのパラメータをダイレクトに用いるKullback-Leibler 距離によるクラスタリングについて検討を行った。

以上の手法の有効性を確認するために、より自然発話に近いデータである討論音声を用いて従来の話者クラスタリング法でよく使用されているCross Likelihood Ratio 法との比較実験を行った。

さらに、音響モデルのみならず言語モデルの構造化を目指して、音素、単語、文節といった認識単位の異なるモデルを併用した認識手法について検討を行った。

４．研究成果

(1) 初年度では、話者適応における適応割合を自動的に学習する手法として、強化学習を用いた大語彙連続音声認識を提案した。

従来の適応法では、MAP 推定における適応の割合を環境に応じて実験的に決定されていた。それに対して、本研究では図１のように、環境に応じてMAP 推定における最適な適応の割合を強化学習により自動的に学習する枠組みを実現した。

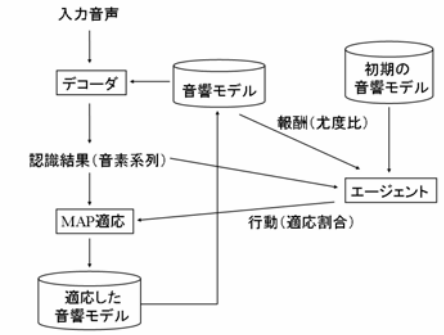

図 1 強化学習に基づく音響モデルの適応割合の推定

日本語話し言葉コーパスの 38 講演を対象に、Q-learning によるMAP 推定の適応割合の自動学習を行った。その結果、尤度比が大きく環境の変化が小さい状態では適応の割合が小さく、尤度比が小さく環境の変化が大きい状態では適応の割合が大きかった。したがって、状態に応じて適切な行動が学習されていることが明らかとなった。

また、提案手法により学習された状態価値が最大となる行動は、認識率が最大となったときのMAP 推定の重み係数とほぼ一致した。

さらに、学習データと異なる 10 講演で認識実験を行った結果、学習データで認識率が最大となったときのMAP 推定の重みによる従来の適応手法と比べて、約 65%とほぼ同等の単語認識精度が得られた。また、図２に示すように、音素認識精度では従来が 77.8%に対して、提案手法では 78.6%でありすべての評価データに対して認識精度の改善が得られ、提案手法が有効であることが明らかとなった。

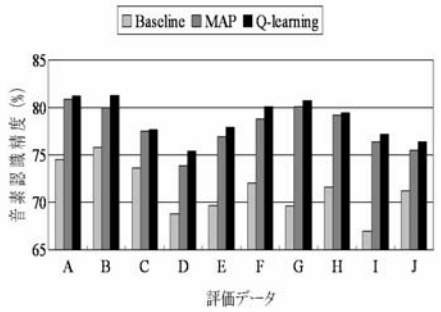

図 2 各評価データにおける音素認識精度

(2) ニ年度目では、限定された複数の話者・雑音が変動する環境下を想定して、オンラインで強化学習により音響モデルを適応し、クラスタリングする手法を提案した。本手法の処理の流れを図３に示す。

強化学習における状態の定義としては、フレーム単位での認識結果と発話全体を考慮した認識結果の一致度に着目した。この一致度をもとに TD 誤差を算出し、その値により環境の変化を判断し、それに応じてモデルの適応量を制御した。このように音響的な特徴変化を見ることで、誰が話しているか、どんな環境であるかを検出することなく、環境の変化に適応することが可能となった。

また、クラスタリングにおいては、現在の環境が過去に学習したものかを判断し、過去に学習していればそのモデルを選択することで高速に適応しモデルを強化して、未知な環境であればモデルを新たに学習することが可能となった。モデルの選択には、音響モデルの平均ベクトルを束ねて、それらのユークリッド距離に基づいて行った。

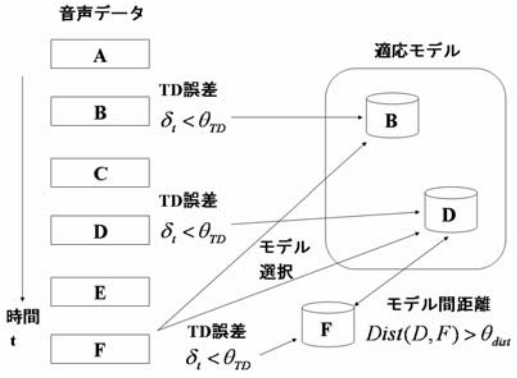

図 3 強化学習に基づく音響モデルのクラスタリング

雑音環境下連続数字認識コーパスであるAURORA-2J を用いて、複数話者・雑音環境におけるオンラインによる音響モデルのクラスタリング手法の評価実験を行った。話者 4 名とレストランや空港などの 10dB の環境雑音が繰り返し変動する発話を800 発話用意し、話者や雑音の種類を変えて 24 パターンデータを作成した。

その結果、適応を行わない場合で 40.7%、従来の MAP 適応では 53.6%、提案手法では55.9%の認識精度が得られた。また、各データセットごとに生成されたクラスタを分析したところ、平均して6 個の適応モデルが話者や雑音の変動に応じて学習されていた。

以上の結果から、提案手法により話者や環境の変動を自動学習することができた。

(3) 三年度目では、音響モデルの効率的な構造化を目指して、音響モデルのパラメータをダイレクトに用いた Kullback-Leibler 距離によるクラスタリング手法を提案した。

話者ごとにクラスタリングを行う場面を想定して、従来の話者クラスタリング法でよく使用されている Cross Likelihood Ratio 法と、新たに Kullback-Leibler 距離をクラスタ間の距離として用いて比較実験を行った。評価データには、1 回 1 時間ほどで話者が 5〜8 名ほど参加している討論音声を 10 セット用いることで、音声対話といったより自然な音声に近い形で評価を行った。

各クラスタの話者ごとに、事前にクラスタリングされたデータベースから音響的に類似したクラスタを先ほどのそれぞれの距離尺度に基づいて上位 30 個選択した。選択されたクラスタの音声データをもとに，Maximum Likelihood Linear Regression による教師無し話者適応を行った。

その結果、KL 距離では 57.3%，CLR では57.7%の単語認識精度が得られ、ほぼ同等の認識精度であった。このことから、オンライン処理では、より計算コストが小さい KL 距離によるクラスタリングが有効であることがわかった。

さらに、音響モデルのみならず言語モデルに関する構造化を目指して、音素、単語、文節単位といった異なる認識単位でモデル化した認識について、カーナビでの目的地を音声対話により検索する場面を想定して検討を行った。その結果、文節のようなより長い認識単位でモデル化したほうが高い認識精度を得ることができた。しかしながら、認識単位が長くなると部分的な信頼度を推定することが難しくなるため、音素や単語といった短い単位での認識器を併用することで、異なる認識単位の結果を統合して認識誤りを部分的に推定することができる可能性が明らかになった。

これまでの結果から、強化学習により話者や環境に依存することなく音響モデルを学習し適応化する統一的な枠組みを新たに実現することができた。

今後は、これまでの音響モデルの構造化を踏まえて、言語モデルの構造化ならびに認識単位に関する検討を行うことで、より頑健な音声認識の実現を目指していきたい。

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）

〔雑誌論文〕（計２件）

(1) 西田 昌史, 強化学習に基づく音声認識－大語彙連続音声認識への適用－, Journal of Signal Processing, vol.12, pp.117-122, 2008, 査読無.
(2) 西田 昌史, 強化学習に基づく音声認識－話者・雑音への適応とクラスタリング－, Journal of Signal Processing, vol.11, pp.353-358, 2007, 査読無.

〔学会発表〕（計３件）

(1) Masafumi Nishida, Unsupervised Training of Adaptation Rate Using Q-learning in Large Vocabulary Continuous Speech Recognition, INTERSPEECH, 28/08/07, Antwerp.
(2) 西田 昌史, 大語彙連続音声認識におけるQ-learningに基づく教師なし適応, 日本音響学会 2007 年秋季研究発表会, 2007 年 9 月 20 日, 山梨大学.
(3) 西田 昌史, 認識単位の異なる認識器を併用した認識結果の信頼度推定, 日本音響学会 2008 年秋季研究発表会, 2008 年 9 月 12 日, 九州大学.

６．研究組織

(1)研究代表者
西田 昌史（Nishida Masafumi）
千葉大学・大学院融合科学研究科・助教
研究者番号：80361442

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：