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

平成　２９　年　　５　月　３１　日現在

機関番号：　１４４０１
研究種目：　若手研究(B)
研究期間：　2015～2016
課題番号：　１５Ｋ１６０５１

研究課題名（和文）音声対話を通じた音声認識用音響・言語モデルの自動高精度化

研究課題名（英文）Automatic Improvement of Acoustic and Language Models of Automatic Speech Recognition through Spoken Dialogue

研究代表者

武田　龍（Takeda, Ryu）

大阪大学・産業科学研究所・助教

研究者番号：２０７４９５２７

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：本研究課題では，音声認識の各モデルに関してメンテナンスフリーな音声対話システムの構築を目的としている．研究課題には，ロボット上での音声対話システムの構築，音響モデルと言語モデルの基礎技術開発がある．主な研究成果では，1) Deep Neural Network (DNN) を用いたモデル構築と音信号への適応技術の開発，2) 話し言葉に対する教師なし音素列の単語分割方法の構築，3) 暗黙的確認を用いた未知語獲得戦略の構築，4) DNNに基づく音源定位技術への展開，を達成した．

研究成果の概要（英文）：Our purpose is the development of a spoken dialogue system of which models used in speech recognition is maintenance-free. The main issues are the development of spoken dialogue systems on robots, and the development of essential technologies on acoustic model and language model. The main outcomes are 1) the development of the acoustic model based on DNN and its adaptation method, 2) the development of the un-supervised segmentation of phoneme sequences for spontaneous utterances, 3) the development of the dialogue strategy for unknown word acquisition using implicit confirmation requests, and 4) the development of sound source localization method based on DNN for human-robot interaction.

研究分野：　音声対話

キーワード：　音声対話　音声認識　音響モデル　言語モデル

１．研究開始当初の背景

音声認識では，音響モデルと呼ばれる発音記号と音声特徴の対応付け，言語モデルと呼ばれる単語辞書や単語の連鎖パターンを事前にデータから統計的に機械学習している．そのため，１）平均から外れた音声特徴を持つ話者は認識精度が低下する，２）新規単語や別表現，新たな言い回しが出てきた場合に認識精度が低下する，という問題がある．これらを防ぐために，専門家による定期的なチューニングが事実上不可欠であるが，システム管理者への負担やシステムのメンテナンスコストは大きい．エンドユーザだけで音声認識に基づくシステムを運用することは困難であるため，システム自身が自動でモデルをチューニングする機能，つまり，メンテナンスフリーなシステムが強く望まれている．一方，音声対話を行うアプリケーションを想定した場合，システム自身がユーザへ応答することが可能である．そのため，未知の音声や単語によりユーザ発話を誤認識した場合でも，認識できなかったことを特定し，不明点をユーザに問い返し理解できれば，システム自身が能動的に内部モデルを更新できる．特に，Web 上の静的なテキスト情報と比較して，音声対話ではピンポイントな情報が入手可能なことが利点である．

２．研究の目的

本研究の目的は，音声認識の各モデルに関してメンテナンスフリーの音声対話システムの実現である．そのためには，人手を介さずにシステム自身がモデル更新を行う機構が必要である．本研究では 0）ロボットにおける音声対話システムの構築技術，1）音声中の誤認識箇所の推定技術，2）音声対話に基づく正解ラベルの推定技術，3)モデルパラメータの動的更新技術，を構築する．なお，人間と音声対話が可能な知能システムを運用するには，様々な人の声や変化していく言葉を常に正しく認識できる必要がある．現状の多くのシステムでは，誤認識した発話を専門家がコストをかけてログから事後的に特定し，音声に対応する正しい発音や単語表記といった正解ラベルを与え，音声認識用のモデルを更新することで実現している．

３．研究の方法

本研究は，要素技術の研究とシステム実装の 2 つを並行して進める必要がある．また，研究の進捗に応じて，推定対象の正解ラベルの種類設定，対話を行うユーザの数や特徴，といった前提条件を変更して研究を進める必要もでてくる．そのため，下記の１～4 を反復して研究内容を高度化していくスパイラルモデルに基づき研究と開発を交互に進める．1．問題設定：獲得対象の正解ラベルと音声対話の前提条件（信号，音素，単語レベル），2．対話データ収集，正解ラベル推定技術と対話戦略の確立，3．モデル更新技術の開発とシステムへの実装（高速化，省メモリ化を含む），4．改良システムの評価と実運用のための活動．この他，ロボットへの対話システム実装において必要な技術も適宜開発する．

４．研究成果

（１）Deep Neural Network（DNN）音響モデルの省メモリ・高速化

本研究では，省メモリ・高速計算が可能な Neural Network（NN）音響モデル構築のために，いくつかのパラメータを離散化した Discrete DNN の実現が目的である．この技術は，ロボットでのシステム実装に不可欠である．本目的を達成するには，本質的な 3 つの要求条件，1）パラメータ離散化における誤差の削減，2）高速計算の実装，3）DNN ノードサイズの削減，を行う必要がある．1）に対しては重みパラメータのモデルとその学習アルゴリズム，2）に対しては一般的な CPU で利用可能なテーブル化を用いた実装方法，3)に対しては層毎に偏りのあるノード削除方法を提案した．提案法１)では，NN の各ノードに適切なパラメータ境界を設定することで量子化の際の誤差を削減する．提案法 2）は，NN のパラメータを数ビットにエンコードすることで CPU キャッシュ内に収まるよう NN のメモリ量を削減し，処理速度の向上を実現する．提案法 3）では，各層においてそれぞれのノードの活性化度を実データから計算し，層依存の活性化スコアを用いて量子化した DNN のノード削減を行う．2-bit 量子化 NN を用いた実験では本手法を適用することにより 8-bit 量子化 NN と同程度の単語正解率を維持した．また，メモリ使用量は 95%削減し，NN のフォワード計算における処理速度も 74%高速化した．

（２）話し言葉を対象とした言語モデル

人は，フィラーや言い淀みといった，予測できない単語を発話中に含めてしまう．通常の N-gram 言語モデルを用いた場合，これらが原因で単語予測精度の低下を招く．本研究では，一種の文脈中の単語選択が可能な，予測に用いる文脈の混合に基づく言語モデルを提案した．可能な部分文脈の重み付き混合として条件付き確率を計算することで，予測できない単語によって引き起こされる負の効果を抑圧する．部分文脈のパターンは組み合わせ的に増加するため，この混合重みのチューニングは重要である．我々は，生成過程が stick-breaking process と可変長 Pitman-Yor 言語モデルで表現されるベイズモデルを用いてこの問題を解決する．評価実験では，予想が難しい雑音を含むテキストに対するパープレキシティーで，提案した言語モデルが従来の N-gram 言語モデルよりも性能を上回ることを明らかにした．

（３）複素活性化関数に基づく DNN 音源定位

音源定位は，ロボットでの音声対話において，ユーザや音イベントの位置検知に必須である．本研究では， DNN を用いた識別的な音源定位手法を提案した．DNN への入力は周波数領域の複素特徴量（位相差・強度差情報）であり，出力は離散化された音源位置（ロボットから見た方向を表すラベル）を用いる．単純に全帯域の複素特徴量を 1 つの実数ベクトルとみなして，全結合ネットワークの DNN への入力する方法は，特に信号対雑音比が低いデータでは学習に失敗する．本研究では，複素構造（強度・位相）やチャネル構造の欠落が誤差伝播学習を阻害する原因であると考え，それらを陽に扱う方向依存活性化関数を導入する．評価実験により，既存の定位手法よりも方向ラベル推定精度が向上することを確認した．

（4）DNN を用いた複数音源定位
本研究では DNN に基づく複数音源定位の学習方法を提案する．DNN は，音源位置のラベルに関する事後確率を識別的に推定するため，高い定位精度を達成できる．従来の音源定位用 DNN は 1 音源を想定しているため，実環境で用いるためには複数音源への拡張必要がある．しかし，単純な拡張は，位置ラベルや学習用データのパターン増加を招き，異なる音源数にわたるラベルの一貫性に欠けるという問題がある．音源定位では，例えば， 1 音源や 2 音源以上の場合，といった音源数が動的に変化する．本研究ではこれらの問題を次のように解決する．前者に対しては，独立な音源位置モデルを，後者に関してはブロック単位での一貫性を持った順序付き位置ラベルの付与で対応する．評価実験では，提案法によって学習した定位用 DNN は，従来手法よりもブロック単位の定位精度で最大 18 ポイントの改善を示した．

（5）DNN 音源定位の適応技術
本研究では，音源定位用 DNN の教師なし適応手法を取り扱った．DNN に基づく音源定位は学習用データと似た音データに対しては，高い定位性能を達成できている．一方，もし音源が異なる残響環境や未学習の音源位置にある場合，定位精度は大きく低下する．この問題は DNN に基づく音響モデルでも同様に問題となるため，知見や解決方法は音響モデルへも転移可能である．我々はこの問題を，観測した音信号に DNN パラメータを教師なし適応することで解決を試みた．エントロピー関数を目的関数として用い，勾配法に基づいてパラメータを最適化する．過学習が起こるが，線形変換ネットワークのような適応用ネットワークの利用やパラメータ更新の早期終了技術によって回避を行う．実験によって，未学習の音源位置や残響環境データに対して，最大 20 ポイント定位性能が改善することがわかった．

（6）クラス推定に基づく対話戦略の構築
対話システムにおいて，自らの知識にない単語（未知語）への対応が課題である．質問により未知語を獲得する手法は提案されているが，雑談対話において質問を逐一行うと，ユーザにとっては煩わしい．本研究では，雑談対話中に現れた未知語のクラスを対話中に獲得するために暗黙的確認を用いることを提案した．まず，未知語の表記からその所属クラスを推定する．推定を最下位クラスと中間クラスとの 2 つのレベルで行い，その結果から暗黙的な確認要求を生成することで，対話を継続させつつ知識を獲得することを狙う．この際，推定結果の正誤の判定は，推定時に得られる確信度に対するしきい値処理により行った．

（7）音素列の教師なし形態素解析
音素列の教師なし分割は，ユーザとの音声対話を通じた未知語獲得において本質的なプロセスである．この分割は入力音素列を単語に相当する分割された音素列に変換することを意味する．未知語の音素列は，信号と単語の中間表現として不可欠である．なぜなら，音情報だけでは，直接その単語のスペルを得ることはできないからである．Pitman-Yor semi-Markov model (PYSMM) は，教師なしで音素列を分割できる有望な手法である．本研究では，音素列の文脈区別と次の音素列の分割位置を予測するために，音素長の文脈モデルを導入した．これは分割ラベルの生成確率を近似しているため，効率的な音素列の分割が期待できる．会話文に対して提案モデルの適用を行い，差ベルの生成確率が学習に作用することを確認した．

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 Ryu Takeda, Kazuhiro Nakadai and Kazunori Komatani: "Acoustic Model Training based on Node-wise Weight Boundary Model for Fast and Small-footprint Deep Neural Networks," 査読有, Computer Speech & Language, 2017. 10.1016/j.csl.2017.02.002

② Ryu Takeda and Kazunori Komatani: "Noise-robust MUSIC-based Sound Source Localization using Steering Vector Transformation for Small Humanoids," 査読有, Journal of Robotics and Mechatronics, Vol.29, No.1, pp.26-36, 2017. 10.20965/jrm.2017.p0026

〔学会発表〕（計 11 件）

① Ryu Takeda and Kazunori Komatani: "Unsupervised Adaptation of Deep Neural Networks for Sound Source Localization using Entropy Minimization," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.2217-2221, Mar. 7, 2017. [査読有 採択率 48.5% (1220/2518)]

② Ryu Takeda and Kazunori Komatani: "Bayesian Language Model based on Mixture of Segmental Contexts for Spontaneous Utterances with Unexpected Words," Proceedings of International Conference on Computational Linguistics (COLING), pp.161-170, Dec. 13, 2016. [査読有 採択率 32.4% (337/1039)]

③ Ryu Takeda and Kazunori Komatani: "Discriminative Multiple Sound Source Localization based on Deep Neural Networks using Independent Location Model," Proceedings of IEEE Workshop on Spoken Language Technology (SLT), pp.603-609, Dec. 16, 2016. [査読有 採択率 60.9% (89/148) regular paper]

④ Kohei Ono, Ryu Takeda, Eric Nichols, Mikio Nakano and Kazunori Komatani: "Toward Lexical Acquisition during Dialogues through Implicit Confirmation for Closed-Domain Chatbots," Proceedings of Second Workshop on Chatbots and Conversational Agent Technologies (WOCHAT), 2016.

⑤ Ryu Takeda and Kazunori Komatani: "Sound Source Localization based on Deep Neural Networks with Directional Activate Function Exploiting Phase Information," Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.405-409, Mar. 23, 2016. [査読有 採択率 47.1% (1265/2682)]

⑥ Ryu Takeda, Kazuhiro Nakadai and Kazunori Komatani: "Acoustic Model Training based on Node-wise Weight Boundary Model Increasing Speed of Discrete Neural Networks," Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp.52-58, Dec. 14, 2015. [査読有 採択率 47.8% (107/224)]

⑦ Ryu Takeda and Kazunori Komatani: "Performance comparison of MUSIC-based sound localization methods on small humanoid under low SNR conditions," Proceedings of IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), pp.859--865, Nov. 4, 2015. [査読有]

⑧ 武田 龍, 中臺一博, 駒谷和範:"量子化 Deep Neural Network のための有界重みモデルに基づく音響モデル学習", 第 46 回 AI チャレンジ研究会, Nov. 2016.

⑨ 武田 龍, 駒谷和範:"方向依存活性化関数を用いた Deep Neural Network に基づく識別的音源定位", 第 112 回音声言語情報処理研究会, July 2016.

⑩ 大野 航平, 武田 龍, エリック ニコルズ, 中野 幹生, 駒谷 和範, "対話を通じた未知語獲得に向けた暗黙的確認の提案", 第 111 回音声言語情報処理研究会, Mar. 2016.

⑪ 大野 航平, 武田 龍, エリック ニコルズ, 中野 幹生, 駒谷 和範, "雑談対話における未知語や属性の獲得のための質問生成", 情報処理学会第 78 回全国大会, Mar. 2016.

〔その他〕
ホームページ等
http://www.ei.sanken.osaka-u.ac.jp/members/rtakeda/

６．研究組織
(1)研究代表者
武田　龍（TAKEDA, Ryu）
大阪大学・産業科学研究所・助教
研究者番号：２０７４９５２７