公開

# 研 究 終 了 報 告 書

## 「音声対話系における言語・音響モデル自動適応」

研究期間: 2018 年 10 月～2022 年 3 月
研 究 者: 武田 龍

1. 研究のねらい

本研究では，音声対話を通じた内部モデルの自動テーラーメイド（適応）の実現，を狙う．内部モデルは，特に音声認識における統計的な音響モデル（音パターン）・言語モデル（語彙や文法）を仮定し，テーラーメイドとは予めシードとして与えられた内部モデルのパラメータ（辞書・出現頻度）を各ユーザにカスタマイズすることを意味する．通常，多くのシステムは，全ユーザに対して同一のモデル（万能モデル，factory-made）を利用している．そのため，平均から外れたユーザの音声や，ユーザ特有の言い回しやシステムの知らない単語（未知語）を含む発話を，システムが正しく認識できない．

音響・言語モデルのユーザへの適応には，ユーザ毎に正解ラベル付きのデータが必要となる．例えば，音声信号（データ）とその書き起こし文（ラベル）などである．データ収集には，1) 事前に定型文をユーザに発話させる方法や，2) 事後的に実際の発話を人手で書き起こす方法がある．しかし，収集可能なデータパターンに制限がある，ユーザビリティー・ユーザへの手間・負荷や費用が大きい等の理由で，音声対話においては現実的な選択肢ではない．

本研究では，音声対話を通じてユーザからラベルを得ることで，ユーザ毎に内部モデルを適応する．つまり，システム自身がユーザ発話において「わからない点（未知パターン系列）」を検出する．次に，「わからない点」をユーザに確認する（教示を得る）ことで，ユーザの発話データと対応する「ラベル」を収集する．得られたラベルとデータに基づき内部モデルを更新する．このアプローチでは，ユーザとシステムの会話のみを要し，事前・事後的な作業が不要なユーザ個別の自動カスタマイズが実現できる．

本研究では，上記の課題に対して統計的機械学習における確率的生成モデルにより解決を図る．確率的生成モデルは，データの生成過程を確率モデルで表現したもので，未知パターンを考慮したモデル化や対話単位でのオンライン適応が容易である．一度しか現れないデータも反映できる点，ユーザの応答もデータから予測できる点も長所である．

具体的な研究項目には下記がある．

A) 言語モデルでの未知パターン対応
B) 音響モデルでの未知パターン対応
C) 音声対話システム実装と音声対話を通じたラベル獲得
D) 複数システムへの拡張

2. 研究成果

（1）概要

研究項目に対する成果の概要を述べる。

A) 言語モデルでの未知パターン対応: 査読付き国際会議 2 件
B) 音響モデルでの未知パターン対応: 査読付き国際会議 1 件

C) 音声対話システム実装と対話を通じたラベル獲得: 査読付き国際会議 2 件、査読付き論文誌 1 件, 対話ロボットコンペ（予備予選 1 位）

言語モデルでの未知パターン対応では、音素列からの未知語検出と単語分割におけるサブワード表現の有効性検証と音素列からの未知語の属性推定手法の開発を行った。これらの技術により、ユーザ発話から未知語を検出し、その属性を推定できる。前者では、自然言語処理分野で開発された教師なし単語分割技術に基づき、音素列からの未知語検出技術を実装・評価した。後者では、混合 PYSMM に基づく属性推定手法を提案し、また、識別モデルに基づく推定結果と統合することで精度向上を実現した。

音響モデルでの未知パターン対応では、深層学習と隠れマルコフモデルに基づく音響モデルにおいて、教師なしで逐次的にモデル更新する技術を開発した。これにより、ユーザ発話を聞くだけでモデルが自動的に適応され、音声認識率を改善できる。ロバストな適応フィルタ技術から着想を得た正則化と疑似エビデンスに基づく更新制御を導入することで、逐次適応の安定化を図った。また、音素認識率と疑似エビデンスの相関を確認し、ユーザの音声認識率の予測が可能なことも確認した。

音声対話システム実装と対話を通じたラベル獲得では、1) 実システムの実装、2) 音声対話システムの開発指針、3) ユーザの年齢予測モデルの提案を行った。実システムを元に、ユーザに合わせた対話戦略の切り替えや、ラベル獲得を目的とした対話戦略の設計ができる。1 点目に関しては、基盤プログラムおよび未知語検出・音響適応技術を実時間実装した。本プログラムを対話ロボットコンペに転用し、予備予選会では 1 位を獲得した。2 点目に関しては、コンペの内容を踏まえ、その中で考慮すべき指針をまとめた。例えば、システム主導の対話、音声認識誤りへの対応などを含む。3 点目では、音声特徴量の正規化に加え、ベイズモデルに基づきラベル・声年齢の曖昧性を表す確率モデルと深層学習を統合することで、年齢予測性能が向上することを明らかにした。対話からのラベル獲得では、実システムを用いたデータ収集と、ユーザ発話の予測に基づくモデル化に取り組んだ

（2）詳細

**研究項目 A)「言語モデルでの未知パターン対応」**

【目的】

言語モデルの観点から、ユーザ発話中に含まれるシステムに未登録の単語（未知語）とパターンに対する検出技術や属性推定技術の開発に取り組む。未知語の検出ができれば, その単語についてユーザへ問い合わせることが可能となる。

【研究成果】

主な成果は以下の2つである。

1) 音素列からの未知語検出と単語分割におけるサブワード表現の有効性を検証: 査読付き国際会議発表
2) 音素列からの未知語の属性推定手法の開発: 査読付き国際会議発表

この 2 つの成果は，図 1 における step1 と step2 の部分に対応する。ここでは、ユーザ発話の表現として音素列（発音記号列）を想定する。この音素列が与えられた時に、未知語を検出し、その検出された未知語の属性を音素列から推定する。実際の対話では、この結果に応じたシステム応答が生成できる。

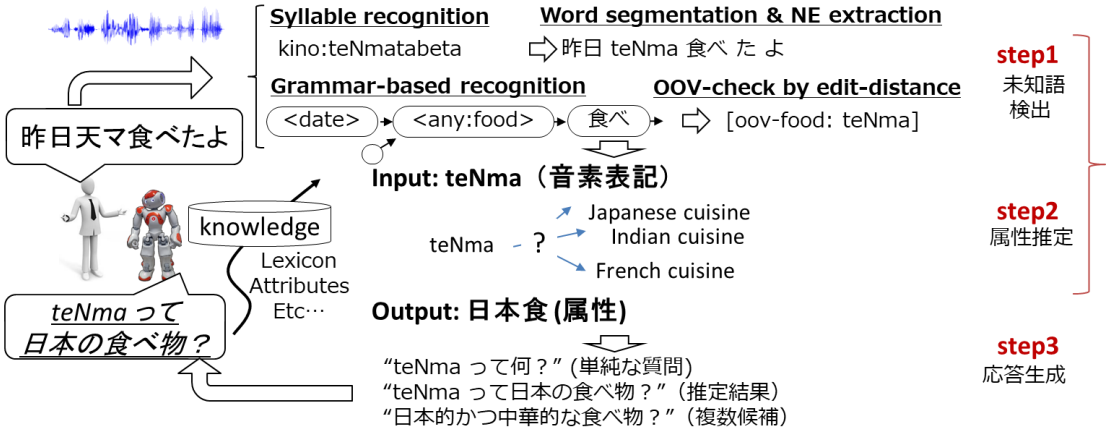


**図 1 未知語の検出および属性推定**

成果1: 音素列からの未知語検出と単語分割におけるサブワード表現の有効性検証

自然言語処理分野で開発された教師なし単語分割技術（Pitman-Yor Semi-Markov Model: PYSMM）に基づき、音素列からの未知語検出技術を実装・評価した。PYSMM は確率的生成モデルに基づいた手法であり、単語の追加・削除が容易という面から本研究課題に適している。PYSMM を音素列へ応用する場合、その単位に関して検討の余地がある。最小の単位は音素だが、音節や音素以上単語未満の単位（サブワード）を用いると未知語検出精度が改善する可能性がある。音節や最適なサブワード表現を用いた場合の未知語検出精度を検証し、音素を単位とした場合と比較し検出精度が8～18ポイント向上することを確認した。また、典型的な発話パターンに含まれる未知語を頑健に検出するために、一部に任意の音節列を受理する記述文法モデルに基づく音声認識を実装した。

成果2: 音素列からの未知語の属性推定手法の開発

混合 PYSMM に基づく属性推定手法を提案し、また、識別モデルに基づく推定結果と統合することで精度向上を行った。前者は単語の全体における音素の出現パターンを、後者では接尾辞などの局所的な音節パターンから、単語の属性を予測する。例えば、未知語「トスダータ」がどこの国の料理か推定する場合、言語特有の発音列が手掛かりとなる。料理名に対する国名および地名に対する観光属性を推定するタスクで評価を行い、統合手法が個々の手法と比べて、最大 4 ポイント属性推定精度が向上することを確認した。

【研究目的の達成状況】

未知語検出に対する基本的な処理の開発や対話システムからの利用を想定したサーバ化は達成した。また、実用性を考慮し、CSJ や BCCWJ などの言語コーパスを用いたモデル学習も実施した。実際のユーザ発話を用いた精度評価も小規模ではあるが行った。一方で、実際の

タスクにおける効果の大規模な評価は達せられておらず，また、性能に関しては研究の余地が残されている。後者に関しては、音響モデルと連動した処理が該当する。例えば、音声認識誤りの考慮、音響特徴量を考慮した単語分割モデルやサブワードを前提として音響モデル構築などがある。

**研究項目 B)「音響モデルでの未知パターン対応」**

【目的】

音響モデルの観点から、ユーザ発話中に含まれるシステムにとって未知の音響特徴を学習する技術や検出する技術の開発に取り組む。ここで、未知の音響特徴は音声認識誤りを引き起こすため、音声認識精度の向上技術および音声認識誤り箇所の検出技術と捉えている。音声認識誤り箇所を検出できれば、ユーザへ内容確認することも可能となる。

【研究成果】

主な研究成果は以下の1つである。

・音声認識中における音響モデルの教師なし適応技術の開発: 査読付き国際会議発表

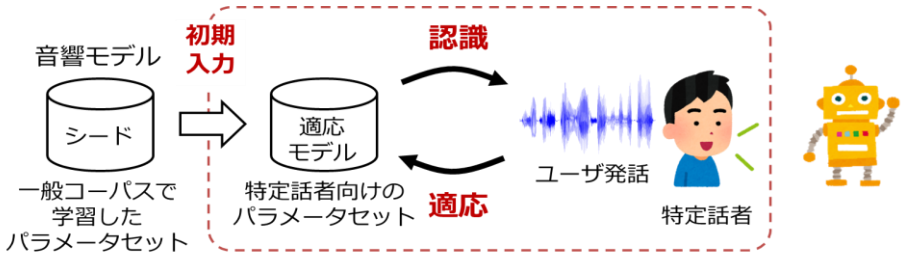


**図 2 音響モデルの教師なし適応**

本成果のモデル適応技術を図 2 に示す。教師なし適応とは教師ラベルを用いずにモデルパラメータをデータに適応することを意味する。ここでは、一般的なコーパスで学習した音響モデルを初期モデルとして用いる。ユーザ発話の書き起こしなしに、音声認識と音響モデル適応を逐次的かつ交互に行うことで、リアルタイムにモデルを更新する。

深層学習と隠れマルコフモデル（DNN-HMM）に基づく音響モデルにおいて、教師なしで逐次的にモデル更新する技術を開発した。DNN-HMM は確率的生成モデルに深層学習を部分的に用いたハイブリッドモデルであり、基本的な音響モデルのひとつであるため取り上げた。周辺化尤度をコスト関数とし，その最大化に基づいて DNN パラメータを更新する。しかし、過剰適応により、主に無音・ポーズラベルの認識結果に陥る問題がある。本研究ではロバストな適応フィルタ技術との関連に着目し、特定音素の識別に関する正則化と、疑似エビデンスに基づく更新制御を提案した。提案法により、教師なし適応を安定化させ、音素認識率を1ポイント程度向上することを確認した。また、疑似エビデンスの値が音素認識率と相関があることがわかった。この値を用いることで、音声認識が難しい話者や音声認識誤りの箇所を推定できる可能性があることが示唆された。

【研究目的の達成状況】

音響モデルの自動適応という観点では、教師なし適応とそのリアルタイム実装により実現した。コーパスのデータセットによる評価ではなく、実際のユーザ発話を用いた場合にも効果があることを確認した。教師なし学習に基づいた適応手法を開発したので、教師あり学習への拡張は原理的に容易である。一方で、音声認識誤り箇所の推定や、誤り箇所をユーザに問い合わせたモデル適応に関しては取り組みを試みたが、成果という形には至っていない。実際のタスクや長時間の対話における適応能力の評価も必要である。

**研究項目 C)「音声対話システム実装と対話を通じたラベル獲得」**

【目的】

実際の音声対話において，未知語のラベルや属性を獲得するためのシステム実装と対話戦略の構築を行う。未知語や認識誤りの確認に対するユーザ応答の理解（ユーザ応答理解）やユーザ特性の推定を行うことで、自然なやりとりで知識の理解が可能となる。また、音声対話システムに各種技術を実装し、実際の対話中でのラベル獲得を実現する。

【研究成果】

主な研究成果は以下の 2 つである.

1）音声信号からのユーザ年齢推定: 査読付き国際会議発表
2）音声対話システムの基盤構築とシステム開発指針の提案: 査読付き国際会議発表，査読付き論文誌

成果1: 音声信号からのユーザ年齢推定

年齢分布や話者数の異なる複数のデータセットからの年齢予測モデルを提案した。各データセットでの音響特性や年齢分布の異なりが原因で、実ユーザの音声信号に対する年齢予測に失敗する。特徴量正規化に加え、ベイズモデルに基づきラベル・声年齢の曖昧性を表す確率モデルと深層学習を統合することで、予測性能が向上することを明らかにした。予測されたユーザ年齢は、システム発話や説明の表現をユーザへ適応することに用いる。

成果2:音声対話システムの基盤構築とシステム開発指針の提案

音声対話システムの基盤構築では、再利用可能な対話システムモジュールの設計・実装を行った。1 回の対話が短いタスクを想定し、その中で考慮すべき指針をまとめた。例えば、システム主導の対話、音声認識誤りへの対応、システムの状態設計などを含む。新学術領域「対話知能学」主催の対話ロボットコンペに参加し、指針に基づき設計したシステムは予備予選会で 1 位を獲得した。

【研究目的の達成状況】

音声対話システム基盤の実装、ユーザ応答理解や対話戦略を確立するための実データ収集と分析は達成した。簡易なユーザ応答理解を音声対話システムへ実装し、実ユーザとの短い対話における未知語検出・属性推定およびラベル獲得性能の評価に取り組んだ。特に音響モデル適応に関しては、実ユーザでも効果があることを確認した。一方で、ユーザ応答理解部の正確なモデルや対話戦略の確立、実タスクにラベル獲得おける評価は未達成である。

**研究項目 D）「複数システムへの拡張」**

研究の主たるねらいである、研究項目 A）～C）を優先して取り組むよう、研究計画を変更した。複数システムへの拡張に関しては未達成となる。

3. 今後の展開

本研究では、音声認識における言語モデル・音響モデルを音声対話中に適応する技術の土台を確立した。各要素技術の精度向上は継続的な課題であるが、人間でも聞き誤りや未知語があるため、対話を通じて上手く学習する技術がより重要となる。今回開発した技術やシステムを土台にして、対話戦略を追求することが今後の展開の一つである。

実際の音声対話システムでは、音声認識のモデルだけでなく、音声対話システムで用いられるモデルをユーザや環境へ自動適応する必要もある。例えば，システムの持つ知識，ユーザ特性やユーザが持っている知識の逐次学習は、メンテナンスフリーな対話システムには必要不可欠である。また、具体的なタスクを設定した音声対話では想定外の事態が起こりがちであり、システムの不審な挙動や発話はユーザへ不安を与える。例えば、ロボットによる家事手伝いタスクでは、子供やペットなどが急に介入し、タスク遂行を阻害する状況もありうる。そのような事態に対しても、システム自身が想定外の事態を認識し、人との対話を通じてその事態を学習できるのであれば、ユーザは安心してシステムを継続利用できる。

本研究成果を社会実装に繋げるには、他のモデルでの適応機能の実現と継続的な実証実験を行えるパートナーとの連携を進める必要がある。実用的なタスクでは、音声認識のモデル以外にも知識やシステム状態、ユーザの振る舞いを表すモデルも使われる。適応に伴うユーザへの適切な確認方法もモデルの性質に異なる可能性があり、具体的なタスクを取り上げて研究を進める必要がある。また、このような実タスクで研究を進めるためには、企業との連携を図ることが一番の近道である。一方で、根本にある音声認識という観点では、実用に耐えうる性能やモジュールの利便性も要求される。これらは社会実装に不可欠であるため、研究・開発を進めつつ、アウトリーチ活動を通じて企業との連携を進めたい。

4. 自己評価

【研究目的の達成状況】

目標とする音声対話システムを構築するために必要な最低限の技術開発は研究項目ごとに達成できた。一方で、精度向上や実タスクにおける長期的な利用における性能評価など積み残した課題もある。各要素技術においても研究要素がまだあるため、ひとつひとつをもう少し深堀して、今後の大きな成果へとつなげていきたい。

【研究の進め方（研究実施体制および研究費執行状況）】

毎年学生協力（1～2 名）を得る予定が最終年度のみの協力に留まったが、研究代表者がシステム実装を含め精力的に一人で進めた。研究費はおおよそ予定通りに執行できたが、新型コロナの影響や研究計画の変更に伴い、研究期間の後半では執行額の調整を行った。

【研究成果の科学記述および社会・経済への波及効果】

本研究成果では、音声認識モデルの学習・適応を扱い、実時間での処理が可能であることを示した。この先にある、未学習事象一般に対しても対話的にモデルを適応する、という自動テーラーメイド化の実現可能性に一歩近づいた。一般的な教師あり学習とは異なる本研究の方向性を突き詰めれば、あらゆるシステムについて、種となるモデルさえ作れば、ユーザや現場をシステム自身が対話を通じて学習できるようになる。これが実現できれば、音声インタフェースの使い勝手が一段向上し、各種情報機器やロボットの在り方・使い方を変えるのは間違いない。例えば、家電製品は家事を自動化したが、近いうちにロボットが各種タスクを自動実行するようになると考える。この時、想定外のトラブルが起きても、音声で簡単に指示を理解してくれるシステムであれば、情報格差や老若男女問わず簡単に自動化の恩恵を受けられる。このような社会への潜在的な波及効果は大きいと考える。

5. 主な研究成果リスト

（1）代表的な論文（原著論文）発表

研究期間累積件数：6件

| |
|---|
| 1. Word Segmentation from Phoneme Sequences based on Pitman-Yor Semi-Markov Model Exploiting Subword Information. Proceedings of SLT. 2018. 763-770. |
| サブワード情報を援用した Pitman-Yor Semi-Markov Model に基づく音素列からの単語分割手法を提案した。単語分割に用いる単位に関して、音節とサブワードを用いた場合の上限性能を評価した。また、自動推定したサブワードを用いることで、音素に基づく単語分割よりも未知語検出性能が向上することを確認した。 |
| 2. Frame-wise Online Unsupervised Adaptation of DNN-HMM Acoustic Model from Perspective of Robust Adaptive Filtering. Proceedings of Interspeech. 2020. 1291-1295. |
| DNN-HMM に基づく音響モデルのフレーム単位での教師なし適応手法を提案した。教師なし学習で用いられる周辺化尤度をコスト関数にし、認識と DNN パラメータの適応を交互に行う。その際問題となるのが、過剰適応による音声認識の失敗である。本研究では、ロバストな適応フィルタ技術から着想を得た正則化とパラメータの更新制御により安定した適応を実現した。 |
| 3. Age Estimation with Speech-age Model for Heterogeneous Speech Datasets. Proceedings of Interspeech. 2021. 4164-4168. |
| 年齢分布や話者数の異なる複数のデータセットからの年齢予測モデルを提案した。各データセットでの音響特性や年齢分布の異なりが原因で、実ユーザの音声信号に対する年齢予測に失敗する。特徴量正規化に加え、ベイズモデルに基づきラベル・声年齢の曖昧性を表す確率モデルと深層学習を統合することで、予測性能が向上することを明らかにした。 |

（2）特許出願

研究期間全出願件数：0 件（特許公開前のものも含む）

（3）その他の成果（主要な学会発表、受賞、著作物、プレスリリース等）

特になし