様式 C-19

科学研究費補助金研究成果報告書

平成　21　年　4　月　1　日現在

研究種目：基盤研究(C)
研究期間：2007～2008
課題番号：19500147
研究課題名（和文）　多レベルの学習機能を有する音声対話システムの研究

研究課題名（英文）　Studies on spoken dialogue systems with multi-level learning

研究代表者
荒木　雅弘（ARAKI　MASAHIRO）
京都工芸繊維大学・工芸科学研究科・准教授
研究者番号：50252490

研究成果の概要：
音声対話システムは、音声認識・自然言語理解・対話管理・文生成・音声合成などを要素技術として構成される。本研究では、音声対話システムの家庭用ロボットなどへの応用を念頭に置き、少数の対話データから音響レベル・言語レベル・対話レベルなどの多レベルの学習が行えるアーキテクチャを考案し、それぞれのレベルにおける特定少数のユーザへの適応技術の実現と、それらの統合方式の検討を行った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	2,100,000	630,000	2,730,000
２００８年度	1,200,000	360,000	1,560,000
年度			
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：音声対話処理
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：(1) 音声対話システム　(2) 機械学習

１．研究開始当初の背景

音声対話システムは、音声認識・自然言語理解・対話管理・文生成・音声合成などを要素技術として構成される。

近年の音声認識・自然言語処理における統計的手法の発展を受け、音声対話システムにおいても機械学習を用いた様々なユーザ適応手法が研究・開発されてきた。例えば、音響モデルの学習においては MLLR 法が話者適応の基本的な手法として用いられている。言語モデルにおいては、大規模コーパスで学習したモデルを併用した少量のタスク依存コーパスからの言語モデルの学習や、active learning を利用した効率的な学習用データの選別などが国内外の研究機関から数多く報告されている。また、あまり国内での研究事例は見られないが、概念モデルの統計学習による意味解析規則の学習や、部分観測マルコフ決定過程を用いた強化学習による対話戦略の獲得などが米国・英国の研究機関などから報告されている。

このように音声対話システムにおける学習の問題は、個別要素に関してはそれぞれ適した機械学習手法の選別が進み、ある程度の学習が可能であることが報告されている。し

かし、これらをひとまとめにして、ユーザが使い込んでゆけばゆくほど様々なレベルで適切に振舞うようになるシステムは、これまで研究されてこなかった。

近い将来の実用化が見込まれている家庭用ロボットのインタフェースや、視覚障がい者用のコミュニケーション支援機器など、特定少数のユーザが、使い込んでゆくうちに性能が向上する音声対話システム技術は、必要不可欠な開発パラダイムであるといえる。

２．研究の目的

本研究では、音声対話システムの家庭用ロボットなどへの応用を念頭に置き、少数の対話データから音響レベル・言語レベル・対話レベルなどの多レベルの学習が行えるアーキテクチャを考案し、それぞれのレベルにおける適応技術の実装と、それらの統合を実現することを目的とする。

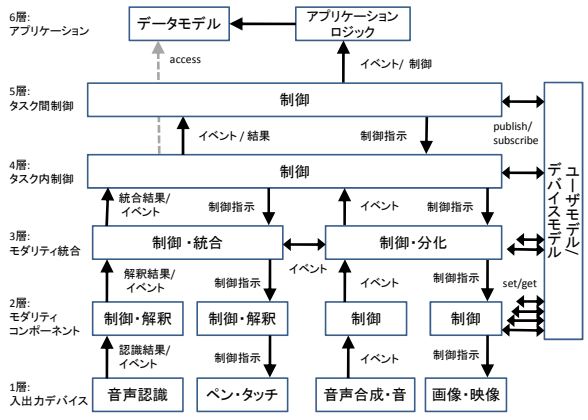
３．研究の方法

音声対話システムをさらに一般化したマルチモーダルインタラクションシステムの標準的なアーキテクチャを検討し、より応用可能性の高い機能分割を行う。そして、そのアーキテクチャのもとで、複数タスクの音声対話システムを実装し、その運用ログから各レベルの学習が行えることを確認する。さらに、あるレベルの学習結果が、他のレベルに及ぼす影響について考察を加える。

４．研究成果

(1) 音声対話システムのフレームワーク

学習可能なフレームワークに関しては、情報処理学会試行標準ワーキンググループの提案するアーキテクチャ(下図参照)に基づいて実装を行った。



このアーキテクチャにおける各階層の内容は以下の通りである。

第1層：入出力デバイス層

個別のモダリティの入出力機能を持つデバイスである。それぞれのデバイスに依存するAPIやイベントなどが用いられる。

第2層：モダリティコンポーネント層

ここでは画面表示，音声入出力、擬人化エージェント制御など、個々のモダリティの制御を行う。入出力の取扱いは分離し、モダリティ間の連携は第3層を介して実現される。SVG の rect 要素、VoiceXML の prompt 要素とSSML 、SALT の listen 要素などに対応する。

第3層：モダリティ統合層

ここでは入力の統合、出力の分化、入出力同期制御などを行う。逐次入力や同時入力の解釈、逐次出力や同時出力の同期、モダリティ拡張などが含まれる。

SMIL 2.1 （出力)、XISL （出力、統合解釈）などに対応する。

第4層：タスク内制御層

ここでは各タスク内の対話の制御と応答内容の決定を行う。フォーム処理における充足性判定や状況判断、フォーム充足のための次応答処理のモダリティ制御、バージインやシステム割り込みなどタスク内の対話遷移処理などが含まれる。VoiceXML の FIA アルゴリズムやHTML の form 要素などに対応する。

第5層：タスク間制御層

ここでは対話タスクの全般的な制御を行う。また、アプリケーション層との入出力通信を行う．SCXML 、 VoiceXMLなどに対応する。

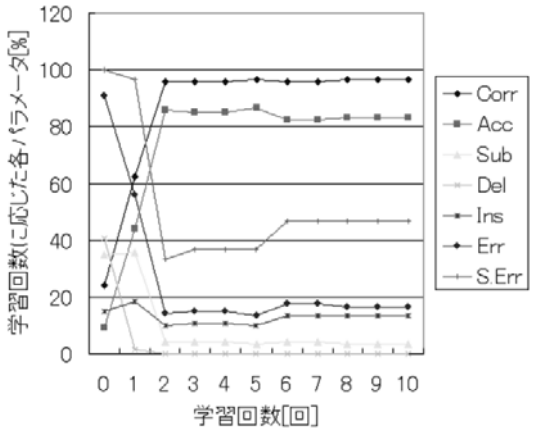
第6層：アプリケーション層

ここではデータモデルとアプリケーションロジックを実装する。

(2) 各モジュールでの学習機能の実現

① 音響モデルの学習

対話の進行に伴って収集される音声ログを適応用データとして、隠れマルコフモデル構築ツールキットHTKを用いて、MLLRによる話者適応を行い、認識率の向上を確認した。



チケット予約タスクにおける音響モデルの学習結果(1話者、30文)

② 言語モデルの学習

新聞記事モデルから生成したコーパスと、認識結果ログ(文法を用いたものとディクテーションを用いたものの両方)を重み付きで結合したものから言語モデルを学習することによって、音声認識率が向上することを確認した。

		正解率	精度
地図検索タスク	適応前	70	69
	適応後	97	93
チケット予約タスク	適応前	88	86
	適応後	89	86

複数タスクにおける言語モデルの学習結果
(1 話者、各 30 文; 単位%)

③ 他のレベルの情報の利用法の検討

意味解析にオントロジーを用いた場合に、その出現頻度に応じて言語モデルのパラメータを変化させる方式や、対話処理におけるユーザに適応した主導権の選択に応じて、出現するユーザ発話に適した意味解析規則・言語モデルのチューニングについて、その実現方法を示した。

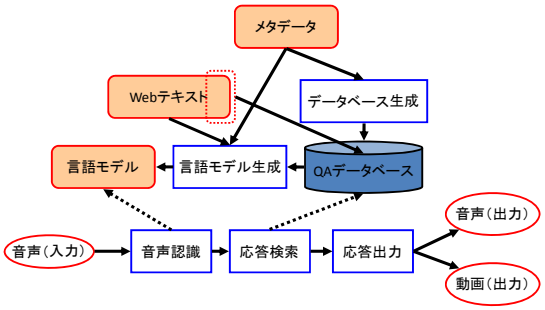
(3) ユーザモデルに対応した出力形式の決定手法

通信している端末の静的な制約および動的なユーザモデルの参照に基づき、適応的にマルチモーダル出力を制御するメカニズムを考案した。
インタラクションを通じて、ユーザの利用可能な(またはより優先的な)モダリティを決定し、そのモダリティを用いる際の端末状況に応じて、音声と画像・文字出力を制御する手法を開発した。情報の選択は、コンテンツからの静的な優先度と使用状況による動的な優先度を統合した基準に基づく。この手法を、上述のアーキテクチャに基づいて実装した。実現した出力に対して、PC, PDA, 携帯電話の各端末用のコンテンツの適切度を被験者実験で評価し、大旨良好な結果を得た。

携帯端末に対する出力例

(4) ユーザの利用状況に合わせた言語モデルの適応

ビデオコンテンツ鑑賞時の質問応答機能をタスクとして選択し、これまでの質問内容に基づいて Web を検索することで関連ページを取得し、そこから言語モデルの適応を行う手法を開発した(下図参照)。



上述のアーキテクチャに基づいて、複数のコンテンツ(伝統技能鑑賞ビデオおよび料理手順の説明ビデオ)に対するインタラクティブプレゼンテーションシステムを作成し、ユーザ発話の収集によるシステムの精度向上可能性を評価した。
コンテンツとしては、京織物の「絣」の解説ビデオ（再生時間約 11 分、シーン数 8）お

よび「焼きそば」の作り方のビデオ（再生時間約 9 分、シーン数 4）を選び、それぞれ MPEG-7 でメタデータを付与し、そこから半自動的にプレゼンテーションシナリオを生成した。

このプレゼンテーションシナリオについて、書き起こし文を 10 名程度の協力者に提示して合計 93 文の質問文を得て、それらを 3 人の被験者が読み上げたものをデータとして用いた。

結果として、複数のコンテンツに対して、ベースラインからの単語認識率の向上および質問応答精度の向上を確認した。

		成功(%)	表記不一致(%)	失敗(%)
絣	Web60k	49.9	6.6	43.4
	提案手法	67.3	11.6	21.1
焼きそば	Web60k	65.0	4.1	30.9
	提案手法	77.8	4.1	18.2

単語認識率の向上

		応答精度			シーン同定精度	
		応答成功(%)	関連有り(%)	応答失敗(%)	一致(%)	不一致(%)
絣	書き起こし質問文	9.7	30.6	59.7	48.4	51.6
	実験者平均	2.7	23.7	73.7	36.6	63.4
焼きそば	書き起こし質問文	9.7	35.5	54.8	51.6	48.4
	実験者平均	8.6	17.2	74.2	51.6	48.4

質問応答精度の向上

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 M. Araki: Filling the gap between a large-scale database and multimodal interactions, In Proc. Third International Conference on Large-Scale Knowledge Resources, LKR 2008, LNCS Vol. 4938, pp.179-185, Springer, 2008.（査読有）
② M. Araki: Proposal of a Markup Language for Multimodal Semantic Interaction, In Proc. Workshop on Multimodal Interfaces in Semantic Interaction, pp.58-62, 2007.（査読有）

〔学会発表〕（計 5 件）

① 服部 貴志, 荒木 雅弘: コンテンツのメタデータを利用した質問応答手法の提案, 人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A803-05, 2009 年 3 月 13 日, 早稲田大学 大久保キャンパス.
② 中川 祐一, 荒木 雅弘: 様々なデバイスに適応するマルチモーダル出力分化手法, 人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A803-04, 2009 年 3 月 13 日, 早稲田大学 大久保キャンパス.
③ 守時 理裕, 荒木 雅弘: 音声対話システムにおけるユーザ適応技術の統合手法の提案, 人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A702-3, 2007 年 11 月 12 日, 関西学院大学 大阪梅田キャンパス.
④ 寺村 真加寿, 荒木 雅弘: オントロジーを利用した音声入力質問文の解析, 人工知能学会 言語・音声理解と対話処理研究会, SIG-SLUD-A702-2, 2007 年 11 月 12 日, 関西学院大学 大阪梅田キャンパス.
⑤ 新田 恒雄, 桂田 浩一, 荒木 雅弘, 西本 卓也, 甘粕 哲郎, 川本 真一: マルチモーダル対話システムのための階層的アーキテクチャの提案, 情報処理学会音声言語情報処理研究会, 2007-SLP-68-2, 2007 年 10 月 19 日, 早稲田大学 大久保キャンパス.

〔図書〕（計 1 件）

① 荒木雅弘：フリーソフトでつくる音声認識システム - パターン認識・機械学習の初歩から対話システムまで -, 総ページ数 232 ページ、森北出版, 2007.

６．研究組織
(1)研究代表者

荒木 雅弘（ARAKI MASAHIRO）
京都工芸繊維大学・工芸科学研究科・准教授
研究者番号：50252490

(2)研究分担者

(3)連携研究者