科学研究費助成事業　　研究成果報告書

平成　２６　年　　４　月　２４　日現在

機関番号：　１４３０１
研究種目：　基盤研究(A)
研究期間：　2010～2013
課題番号：　２２２４００１３

研究課題名（和文）聞き手の反応に着目した音声会話の解析と生成

研究課題名（英文）Analysis and Generation of Speech Conversation by Considering Listener's Reaction

研究代表者
河原　達也（KAWAHARA, TATSUYA）

京都大学・学術情報メディアセンター・教授

研究者番号：００２３４１０４

交付決定額（研究期間全体）：（直接経費）　　35,600,000円、（間接経費）　　10,680,000円

研究成果の概要（和文）：音声によるコミュニケーションの解析とそれに基づく音声対話システムの高度化を行った。本研究では、特に聞き手の反応に着目した新たなアプローチを提案し、そのモデル化と実現を行った。人間どうしのインタラクションにおいて、聞き手の視線・相槌や笑い声などの反応を検出し、これらの情報に基づいて興味度・理解度の推定を行った。また、聞き手の興味や焦点に応じて、話題を選択して情報提供する音声対話システムの構築も行った。

研究成果の概要（英文）：A novel approach to analysis of speech communication and design of conversational systems is investigated. It particularly focuses on listener's reaction. In human interactions, listener's reactions such as eye-gaze, backchannels and laughter are detected, and these behavior signals are combined to predict the interest and comprehension level. Moreover, a new type of spoken dialogue systems is developed, which conducts proactive information presentation based on the user's interest and focus.

研究分野：　総合領域

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：　画像、文章、音声等認識　コンテンツ・アーカイブ　エージェント　マルチモーダルインターフェース
音声会話

１．研究開始当初の背景

音声によるコミュニケーションは、太古より人間どうしの知識伝達・意見交換の根源的な手段であり、電子的媒体・Web が発達した現在においても、新たな知の創造は、主に face-to-face の議論や会話を通して行われていると考えられる。

しかしながら、講演や討論などがフォーマルな場で、明瞭かつ一方的に発話されるのに対して、日常的な議論や会話はそれほどフォーマルでなく、インタラクティブに行われるのが特徴である。このような発話の区切りや発音・言語表現が不明瞭な状況で、発話内容をすべて音声認識（テキスト化）し、言語解析して、要約や対話を行うといった従来のアプローチ（「コンテンツに基づく処理」と呼ぶ）に限界があるのは明らかである。

これに対して、システムが話の内容すべてを認識・理解しなくても、音声コミュニケーション（議論や会話など）の場にいる人間の反応を捕捉することで、要約や検索を行うための手がかりを得られるのではないか、というのが本研究の主な着想である。例えば、議論において聴衆がうなずいたり、あいづちを打っている箇所が重要と考えられる。本アプローチは、話の内容を聞いている人間の反応に基づいて音声コミュニケーションの処理（解析及び生成）を行うもので、「インタラクションに基づく処理」と呼ぶ。

２．研究の目的

本研究では、人間どうしのインタラクション、及び日常的な話題を扱う音声対話システム(会話エージェント)を対象として、上記の処理のモデル化と実現を目指す。前者のインタラクションに関する分析を行い、その知見を生かしながら、後者の会話システムを高度にしていく。

まず、会話コーパスを収集した上で、以下の課題に取り組む。

(a) 話し言葉音声認識の高度化と話者・ドメイン適応
　これまで学会講演や国会討論を対象に開発してきたものを一層洗練するとともに、話題や話者に対して、高速にかつ教師なしで適応する方法を研究する。

(b) 聞き手の反応の捕捉とインタラクションのモデル化
　聞き手の言語的・非言語的な反応、具体的には、あいづち、笑い、うなずき、視線、指差しなどに着目し、これらの現象と会話のモード、興味、重要度などとの関係を明らかにする。少なくとも音声メディアに関するものは自動抽出できるようにする。また、これらの反応に先行する発話との関係も調べる。

(c) 会話における話題と焦点のモデル化と制御
　会話における話題と焦点の推移をアノテーションした上で、発話の表層的表現との関係、ならびに上記の言語・非言語反応との関係を明らかにする。そして、会話システムにおいて話題管理と焦点制御を行う方法を研究する。

その上で、以下のシステムを作成し、評価を行う。

(d) 会話の自動アノテーション
　人間どうしのインタラクションをアーカイブ化した上で、話題区間の分割、聴衆が興味を持った箇所や理解困難だった箇所のインデキシングを行う。

(e) 柔軟な音声対話システム（会話エージェント）
　観光名所の案内や商品の紹介、さらには時事的な話題を元に会話を行うエージェントを設計・実装する。相手の反応を見ながら、的確な話題・焦点の制御を行えるようにする。

３．研究の方法

(1) 平成２２年度

まず、人間どうしの話し言葉音声を収録し、言語的・非言語的な情報のアノテーションを行う。その上で、通常の音声認識に加えて、これらの情報を自動検出する方法を研究し、聞き手の反応との関係を分析する。次に、談話レベルのモデル化、特に話題と焦点の管理を行う方法を設計した上で、音声対話システムを多様なドメインに展開できるようにする。具体的には以下のように進める。

① 音声コーパスの収録とアノテーション
② 発話区間検出と話者決定
③ 聞き手の反応の検出
④ 音声認識の高度化
⑤ 音声対話システム（会話エージェント）の構築

(2) 平成２３～２５年度

前年度に引き続き、コーパスの収録とアノテーションを行いながら、以下の研究を進める。

① 聞き手の反応と認知状態の分析
② 音声認識の高度化
③ 会話アーカイブのアノテーション
④ 会話アーカイブのブラウザの作成
⑤ 会話エージェントの多様なドメインへの展開
⑥ 会話エージェントにおける話題と焦点の制御

４．研究成果

(1) 話し言葉の音声認識の音響モデルの高度化

話し言葉の音声認識のための音響モデルの学習には大規模なコーパスが必要であるが、話し言葉の忠実な書き起こしを用意するのは多大なコストを要する。これに対して、会議録や字幕などの忠実な書き起こしでないが、容易に入手できるテキストを活用する準教師付き学習の枠組みを研究した。提案する手法では、会議録のテキストデータに統計的話し言葉変換を適用して、会議の詳細な単位（ターン）ごとに制約の強い言語モデルを作成し、この言語モデルを用いて音声認識を行うことで、音響モデル学習のためのラベルを作成する。国会審議を対象とした音声認識評価実験により、従来手法よりも高い精度のラベルを作成できること、及びこのラベルを用いて人手のラベルを用いた場合と同等の精度のモデルを学習できることが示された。

(2) 話し言葉の音声認識の言語モデルの高度化

話し言葉の音声認識では、言語モデルがドメインに関連する表現とフィラーや口語表現などの話し言葉特有の表現の両方をカバーすることが求められる。本研究では、単語・構文などの情報に基づくルールベースの話し言葉テキスト変換と、N-gram の統計的話し言葉変換を組み合わせて、書き言葉スタイルのテキストから話し言葉スタイルの言語モデルを構築する手法を検討した。学会講演音声を対象とした評価実験において、提案手法の効果の評価を行った。

(3) 音声対話システムのための音声認識の言語モデルの高度化

音声対話システムのための音声認識における言語モデル構築のために、Web 上で収集される文から適切なものを選択する手法を研究した。従来手法では文表層のパープレキシティを用いた文選択が一般的であったが、提案手法では音声対話において利用される文書集合（＝ドメイン）との意味的な類似度を定義し、これを文選択に用いる。具体的には、ドメイン固有の述語項構造パターンに着目し、統計的な尺度を定義する。この意味的な類似度と従来のパープレキシティに基づく手法を組み合わせることも検討する。２種類の異なるドメインにおける音声認識実験によって、提案する文選択手法が有効であることが示された。

(4) 聴衆の反応に基づく会話のアノテーション

ポッドキャストやプレゼンテーション会話といった音声会話コンテンツを対象として、会話音声中の聞き手の反応に基づいて、視聴者にとって有益な箇所を抽出する手法を研究した。笑い声やあいづちを生起させる箇所（＝ホットスポット）は第三者である視聴者にとっても有益な情報を含んでいると考えられる。本研究では、笑い声とあいづちの検出を行い、検出されたそれぞれのイベントに基づいて、「おもしろスポット」と「なるほどスポット」の２種類のホットスポットを定義し、それらの抽出を行った。被験者実験によって各ホットスポットの妥当性を評価し、これらの大半が実際に被験者が興味・関心をもった箇所であることを確認した。

(5) 聴衆の興味・理解度の推定

プレゼンテーション会話における聴衆の興味・理解度の自動推定を行った。このような会話では、聴衆の視線や相槌などの振る舞いが顕著に見られる。これらの振る舞いは、興味・理解度と関係があると考えられる。また興味・理解度は、聴衆の質問や相槌などの発話行為からも推測できると考えられる。本研究ではまず、興味・理解度と発話行為の関係を分析した。次に、発話行為と聴衆の振る舞いとの関係を調べた。これに基いて、話題セグメント毎にマルチモーダルな振る舞いから、質問の生起とその種類の予測を行った。実験の結果、相槌と視線の特徴量が予測に有効であることと、それらを組み合わせることの相乗効果が確認された。この知見に基いて、会話の様子を視覚化するブラウザを作成した。

(6) 柔軟な情報推薦を行う音声対話システム

日々更新される Web ニュースなどのテキストに対して、述語項構造に着目した情報抽出を行い、それに基づいて情報検索・推薦を行う音声対話システムを構築した。まず、ドメインごとに有用な述語項構造パターンの抽出を行う指標を検討し、ナイーブ・ベイズ法に基づく抽出が有効であることを示した。また、抽出された述語項構造に完全に一致するものがない場合でも情報推薦ができるように、前述の指標に基づいて述語項の優先度を決定し、さらに、要素・用言に関して関連度を定義することによって述語項どうしの類似度を計算した。評価実験において、典型的な従来手法である Bag-Of-Words（BOW）モデルと比較して、本手法がより的確に応答生成を行えることが示された。これに加えて、ユーザからの情報要求・発話がなくなった場合に、対話履歴中の述語項との類似度を利用してプロアクティブに情報提示を行う手法を提案した。

(7) ユーザの焦点に適応的な音声対話システム

聞き手の興味に基いて会話を行うエージェントを構築した。これは、日々動的に更新される Web ニュース記事を対象として、音声

による雑談的な情報案内を行うものである。ユーザがどの情報に興味があるかという焦点に着目し、ユーザとの対話を通じて漠然とした情報要求に応えることを目標とした。本研究では、ユーザの意図推定と焦点解析をドメインにできるだけ依存しない形で機械学習により実現し、さらに部分観測マルコフ決定過程(POMDP)を用いた統計的対話制御により、ユーザの状態と焦点に最適化された情報案内モジュールの選択を行う枠組みを実現した。

５．主な発表論文等
（研究代表者、研究分担者には下線）

〔雑誌論文〕（計　14 件；すべて査読有）

[1] M.Ablimit, T.Kawahara, and A.Hamdulla. Lexicon optimization based on discriminative learning for automatic speech recognition of agglutinative language. Speech Communication, Vol.60, pp.78-87, 2014. http://dx.doi.org/10.1016/j.specom.2013.09.011

[2] 吉野幸一郎，森信介，河原達也．述語項構造を介した文の選択に基づく音声対話用言語モデルの構築．人工知能学会論文誌，Vol.29，No.1，pp.53--59，2014.

[3] 秋田祐哉，河原達也．講演に対する読点の複数アノテーションに基づく自動挿入．情報処理学会論文誌，Vol.54，No.2，pp.463--470，2013.

[4] G.Neubig, Y.Akita, S.Mori, and T.Kawahara. A monotonic statistical machine translation approach to speaking style transformation. Computer Speech and Language, Vol.26, No.5, pp.349--370, 2012. http://dx.doi.org/10.1016/j.csl.2012.02.003

[5] 三村正人，河原達也．会議音声認識における BIC に基づく高速な話者正規化と話者適応．電子情報通信学会論文誌，Vol.J95-D，No.7，pp.1467--1475，2012.

[6] G.Neubig, M.Mimura, S.Mori, and T.Kawahara. Bayesian learning of a language model from continuous speech. IEICE Trans., Vol.E95-D, No.2, pp.614--625, 2012.

[7] 吉野幸一郎，森信介，河原達也．述語項の類似度に基づく情報抽出・推薦を行う音声対話システム．情報処理学会論文誌，Vol.52，No.12，pp.3386--3397，2011.

[8] 河原達也，須見康平，緒方淳，後藤真孝．音声会話コンテンツにおける聴衆の反応に基づく 音響イベントとホットスポットの検出．情報処理学会論文誌，Vol.52，No.12，pp.3363--3373，2011.
（情報処理学会 2012 年度論文賞受賞）

[9] 三村正人，秋田祐哉，河原達也．統計的言語モデル変換を用いた音響モデルの準教師付き学習．電子情報通信学会論文誌，Vol.J94-D，No.2，pp.460--468，2011.

[10] D.Cournapeau, S.Watanabe, A.Nakamura, and T.Kawahara. Online unsupervised classification with model comparison in the Variational Bayes framework for voice activity detection. IEEE J. Selected Topics in Signal Processing, Vol.4, No.6, pp.1071--1083, 2010. http://dx.doi.org/10.1109/JSTSP.2010.2080821

[11] 秋田祐哉，三村正人，河原達也．会議録作成支援のための国会審議の音声認識システム．電子情報通信学会論文誌，Vol.J93-D，No.9，pp.1736--1744，2010.

[12] R.Gomez and T.Kawahara. Robust speech recognition based on dereverberation parameter optimization using acoustic model likelihood. IEEE Trans. Audio, Speech \& Language Process., Vol.18, No.7, pp.1708--1716, 2010. http://dx.doi.org/10.1109/TASL.2010.2052610

[13] Y.Akita and T.Kawahara. Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Trans. Audio, Speech & Language Process., Vol.18, No.6, pp.1539--1549, 2010. http://dx.doi.org/10.1109/TASL.2009.2037400

[14] K.Ishizuka, S.Araki, and T.Kawahara. Speech activity detection for multi-party conversation analyses based on likelihood ratio test on spatial magnitude. IEEE Trans. Audio, Speech & Language Process., Vol.18, No.6, pp.1354--1365, 2010. http://dx.doi.org/10.1109/TASL.2009.2033955

〔学会発表〕（計　30 件）

[1] T.Kawahara.
Smart posterboard: Multi-modal sensing and analysis of poster conversations.
In Proc. APSIPA ASC, (plenary overview talk), 2013 年 10 月 台湾・高雄.

[2] T.Kawahara, S.Hayashi, and K.Takanashi.
Estimation of interest and comprehension level of audience through multi-modal behaviors in poster conversations.
In Proc. INTERSPEECH, pp.1882--1885, 2013 年 8 月 フランス・リヨン.

[3] K.Yoshino, S.Mori, and T.Kawahara.
Incorporating semantic information to selection of web texts for language model of spoken dialogue system.
In Proc. IEEE-ICASSP, pp.8252--8256, 2013 年 5 月 カナダ・バンクーバー.

[4] K.Yoshino, S.Mori, and T.Kawahara.
Language modeling for spoken dialogue system based on filtering using predicate-argument structures.
In Proc. COLING, pp.2993--3002, 2012 年 12 月 インド・ムンバイ.

[5] C.Lee and T.Kawahara.
Hybrid vector space model for flexible voice search.
In Proc. APSIPA ASC, 2012 年 12 月 米国・ロサンジェルス.

[6] K.Yoshino, S.Mori, and T.Kawahara.
Language modeling for spoken dialogue system based on sentence transformation and filtering using predicate-argument structures.
In Proc. APSIPA ASC, 2012 年 12 月 米国・ロサンジェルス.

[7] Y.Akita, M.Watanabe, and T.Kawahara.
Automatic transcription of lecture speech using language model based on speaking-style transformation of proceeding texts.
In Proc. INTERSPEECH, 2012 年 9 月 米国・ポートランド.

[8] R.Gomez and T.Kawahara.
Dereverberation based on wavelet packet filtering for robust automatic speech recognition.
In Proc. INTERSPEECH, 2012 年 9 月 米国・ポートランド.

[9] T.Kawahara, T.Iwatate, and K.Takanashi.
Prediction of turn-taking by combining prosodic and eye-gaze information in poster conversations.
In Proc. INTERSPEECH, 2012 年 9 月 米国・ポートランド.

[10] T.Kawahara, T.Iwatate, T.Tsuchiya, and K.Takanashi.
Can we predict who in the audience will ask what kind of questions with their feedback behaviors in poster conversation?
In Proc. Interdisciplinary Workshop on Feedback Behaviors in Dialog, pp.35--38, 2012 年 9 月 米国・ポートランド.

[11] T.Kawahara.
Transcription system using automatic speech recognition for the Japanese Parliament (Diet).
In Proc. AAAI/IAAI, pp.2224--2228, 2012 年 7 月 カナダ・トロント.

[12] T.Kawahara.
Multi-modal sensing and analysis of poster conversations toward smart posterboard.
In Proc. SIGdial Meeting Discourse \& Dialogue, pp.1--9 (keynote speech), 2012 年 7 月 韓国・ソウル.

[13] M.Ablimit, T.Kawahara, and A.Hamdulla.
Discriminative approach to lexical entry selection for automatic speech recognition of agglutinative language.
In Proc. IEEE-ICASSP, pp.5009--5012, 2012 年 3 月 京都.

[14] R.Gomez and T.Kawahara.
Optimized wavelet-based speech enhancement for speech recognition in noisy and reverberant conditions.
In Proc. APSIPA ASC, 2011 年 10 月 中国・西安.

[15] M.Mimura and T.Kawahara.
Fast speaker normalization and adaptation based on BIC for meeting speech recognition.
In Proc. APSIPA ASC, 2011 年 10 月 中国・西安.

[16] M.Ablimit, T.Kawahara, and A.Hamdulla.
Lexicon optimization for automatic speech recognition based on discriminative learning.
In Proc. APSIPA ASC, 2011 年 10 月 中国・西安.

[17] T.Hirayama, Y.Sumi, T.Kawahara, and T.Matsuyama.
Info-concierge: Proactive multi-modal interaction through mind probing.
In Proc. APSIPA ASC, 2011 年 10 月 中国・西安.

[18] C.Lee, T.Kawahara, and A.Rudnicky.
Combining slot-based vector space model for voice book search.
In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS), pp. 27--35,

2011年9月 スペイン・グラナダ.
[19] Y.Akita and T.Kawahara.
Automatic comma insertion of lecture transcripts based on multiple annotations.
In Proc. INTERSPEECH, pp.2889--2892, 2011年8月 イタリア・フィレンチェ.
[20] R.Gomez and T.Kawahara.
Denoising using optimized wavelet filtering for automatic speech recognition.
In Proc. INTERSPEECH, pp.1673--1676, 2011年8月 イタリア・フィレンチェ.
[21] K.Yoshino, S.Mori, and T.Kawahara.
Spoken dialogue system based on information extraction using similarity of predicate argument structures.
In Proc. SIGdial Meeting Discourse \& Dialogue, pp.59--66, 2011年6月 米国・ポートランド.
[22] R.Gomez and T.Kawahara.
Optimizing wavelet parameters for dereverberation in automatic speech recognition.
In Proc. APSIPA ASC, pp.446--449, 2010年12月 シンガポール.
[23] T.Kawahara.
Automatic transcription of parliamentary meetings and classroom lectures -- a sustainable approach and real system evaluations --.
In Proc. Int'l Sympo. Chinese Spoken Language Processing (ISCSLP), pp.1--6 (keynote speech), 2010年12月 台湾・台南.
[24] K.Yoshino and T.Kawahara.
Spoken dialogue system based on information extraction from web text.
In Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS) (LNAI 6392), Vol.Demo. Paper, pp.196--197, 2010年9月 御殿場.
[25] T.Kawahara, K.Sumi, Z.Q.Chang, and K.Takanashi.
Detection of hot spots in poster conversations based on reactive tokens of audience.
In Proc. INTERSPEECH, pp.3042--3045, 2010年9月 幕張.
[26] G.Neubig, M.Mimura, S.Mori, and T.Kawahara.
Learning a language model from continuous speech.
In Proc. INTERSPEECH, pp.1053--1056, 2010年9月 幕張.
[27] T.Kawahara, N.Katsumaru, Y.Akita, and S.Mori.
Classroom note-taking system for hearing impaired students using automatic speech recognition adapted to lectures.
In Proc. INTERSPEECH, pp.626--629, 2010年9月 幕張.
[28] R.Gomez and T.Kawahara.
An improved wavelet-based dereverberation for robust automatic speech recognition.
In Proc. INTERSPEECH, pp.578--581, 2010年9月 幕張.
[29] Y.Akita, M.Mimura, G.Neubig, and T.Kawahara.
Semi-automated update of automatic transcription system for the Japanese national congress.
In Proc. INTERSPEECH, pp.338--341, 2010年9月 幕張.
[30] T.Kawahara, Z.Q.Chang, and K.Takanashi.
Analysis on prosodic features of Japanese reactive tokens in poster conversations.
In Proc. Int'l Conf. Speech Prosody, 2010年5月 米国・シカゴ.

〔図書〕（計 0件）

〔産業財産権〕
○出願状況（計 0件）
○取得状況（計 0件）

〔その他〕
なし

６．研究組織

(1)研究代表者
河原 達也 (KAWAHARA TATSUYA)
京都大学・学術情報メディアセンター・教授
研究者番号：00234104

(2)研究分担者
角 康之 (SUMI YASUYUKI)
公立はこだて未来大学・システム情報科学部・教授
研究者番号：30362578

秋田 祐哉 (AKITA YUYA)
京都大学・学術情報メディアセンター・助教
研究者番号：90402742

森 信介 (MORI SHINSUKE)
京都大学・学術情報メディアセンター・准教授
研究者番号：90456773

(3)連携研究者
なし