科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２０　日現在

機関番号：　１３９０１
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００２０９

研究課題名（和文）人間対人間の対話情報を事前知識とした情報検索音声対話戦略

研究課題名（英文）Dialog strategy for information retrieval usign prior knowledge obtained from human-human dialog

研究代表者
北岡　教英（Kitaoka, Norihide）

名古屋大学・情報科学研究科・准教授

研究者番号：１０３３３５０１

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：以下を実施した。
(1)対話の理解候補をグラフ構造の各ノードとして、グラフ全体でシステムの理解状態を表現する意味理解表現法とグラフ展開による理解の進展の表現、および対話において最適な展開を誘発するユーザへの応答選択戦略を提案(2)対話音声を認識して得た情報から相応しい楽曲を検索して提示する方法を提案(3)潜在意味解析を通して、理解の前提知識（文書群・音声認識結果から成る）から関連する文書を検索する、音声ドキュメント検索技術の高精度化を達成(4)楽曲検索をする対話を想定し、楽曲間の類似性に関する人間の感覚と物理量との対応を示し、検索時の事前知識とする方法を検討

研究成果の概要（英文）：We conducted below:
(1)We proposed to explain multiple understanding states using graph structure and the method to expand the nodes according to the progress of the dialog, to lead a "good" answer from the user,(2)we also proposed the method to recommend suitable music for the situation where humans were talking using the recognition result of the conversation, (3)we improved spoken document retrieval methods towards constructing a prior knowledge for understanding, (4)Considering the dialog to search music, we investigated physical features of acoustics to express the individual similarity measurement of humans, to construct prior knowledge to the search.

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：　音声対話　情報検索　潜在意味解析　楽曲間類似度

１．研究開始当初の背景

音声対話インタフェースは人間－機械インタラクションに対して有望であるとされながら、十分に実用化されているとは言い難い。その原因として、単純で決まった入力しか受け付けられない、誤認識する、誤認識からの回復が容易でない、という点がある。これに対し、人間同士の対話を参考に考えた場合、

・人間対人間の対話は自然で対機械で得られるよりも豊富な情報が含まれている
・潜在意味空間によりメディアを横断した情報の表現法が存在する

という仮説に基づいた高度化が考えられる。そして、最終的には、人間対人間対話の豊富な情報（＝事前情報・ヒューリスティクス）を基に、対機械対話をより豊かで有用にできるという考えに至った。

２．研究の目的

上記背景に基づいて、以下の２点を研究対象とし、目的の対話システムに向けた背景知識形成とそれに基づく探索の基本技術を開発する。

・音声認識における誤認識に伴う曖昧さを表現した対話理解表現
　グラフ構造上の複数ノード（＝理解結果）の保持と、ターンごとにグラフを展開して探索する理解の進行と展開におけるヒューリスティックスの導入
・潜在意味表現によるメディア横断情報表現と検索
　潜在的意味解析による言語あるいは人間同士の対話内容と音楽など異なるメディアの統一的空間における表現方法と、その空間内における検索に基づいたメディアを横断した検索手法

３．研究の方法

(1) グラフ構造の各ノードを一つの理解候補として、グラフ全体でシステムの理解状態を表現する意味理解表現法と、そのグラフを対話の入力ごとに展開することによる理解の進展の表現、および対話において最適な展開ができるためのユーザへの応答選択戦略を研究する。
(2) 対話音声を認識し、その情報からその場にふさわしい楽曲を検索して提示する方法について研究する。
(3) 潜在意味解析を通して、理解の前提となる知識（文書群・音声認識結果から成る）から関連する文書を検索する、音声ドキュメント検索技術の高精度化を研究する。
(4) 楽曲検索をする対話を想定し、音楽と、人間の感覚（特に楽曲間の類似性）とを関連付ける物理量を探索し、検索時の事前知識とする方法を研究する。

４．研究成果

(1) 理解状態のグラフ探索に基づいた音声対話戦略

音声対話において誤認識は避けられない問題である。この問題を解決するために、対話ターンごとの確認を毎回行うなどの処理が必要になるが、長い対話が必要となりわずらわしいものとなる。そこで、本研究では、情報検索を、各ノードが理解の仮説候補を表現するグラフ探索問題として捉え、確認発話を減らすいくつかの新しい対話戦略を提案した。本研究では、カーナビのように地点を検索するタスク、および音楽データベースから楽曲を検索するタスクにおいてその評価を行ったが、本報告では楽曲検索のタスクにおける結果について述べることとする。

① 楽曲検索タスク

システムに対し、ユーザは音声対話によって所望の楽曲を検索するものとする。ユーザはアーティストや曲名などを発話し入力することもできるが、楽曲のジャンルや年代、歌手の性別のような情報を対話によって入力していくことで絞り込んで好みの曲を検索することもできる。例えばユーザは、「ロックの曲を探してください」「女性ボーカルの曲を聴きたいです」「１９９０年の曲をお願いします」などの入力が可能である。それに対し、システムは、次のような応答（質問）により、条件を絞る発話をユーザに促す。「ジャンルはなんですか？」「いつの曲をお探しですか」（以上のような新規情報要求）「○○○（人名）の曲でよろしいですか」「△△（ジャンル）の曲でよろしいですか」（以上のような確認）。

② グラフ探索に基づく情報検索のための音声対話理解

まず、情報検索の過程をグラフ探索であると考える。音声対話において、ユーザからの入力で得られた検索キーをスロットに埋めていく過程をグラフで表現する（図１）。グラフのノードとして（部分的な）キーワード集合を考えることで理解の探索グラフを構成する。キーが入力されるたびにノードを展開し、最終的に正しい理解に到達することで検索を実行する。

音声対話はしばしば誤りを含むが、このグラフには複数の（スコア付きの）ノードがあり、各ノードを理解、ノードの集合（木構造）を理解状態と考えることで、常時複数の理解候補を保持し、理解誤りがあっても復帰が可能となるようにする。

スコアは音声認識から得られる認識結果の信頼度に基づいて付与されている。そして、一時的に正しい理解のスコアが低下したと

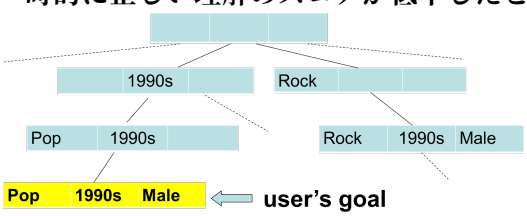

図１　情報検索対話理解のグラフ表現

しても、対話を通してトータルのスコアが上昇することによって正しい理解に到達する可能性がある。

③ システム応答選択基準

対話のゴールは、対話の最後に最も正解の可能性の高い理解仮説（ノード）を選択することである。対話においては、できるだけ早く到達できるような入力をユーザにしてもらうことにより効率が上がる。しかし、現状の理解から単純に効率的な応答を返せば、システムが誤理解をしていることをユーザに晒してしまうことになる。そこで、以下に示す「効率性」と、対話履歴に矛盾せず自然であることを示す「無矛盾性」の尺度を併用した応答の選択基準を提案した。

・効率性尺度

探索問題として考えた場合、探索空間を大きく絞り込める質問が効率的である。そこで、エントロピーに基づいて、質問 q をした後に得られる回答の集合 A_q から、相互情報量を算出する。

$$S_e(q) = I(X;q \mid n) = H(X \mid n) - H(X \mid q,n)$$

$$H(X \mid n) = -\sum_{x \in X} p(n,x) \log_2 p(x \mid n)$$

$$H(X \mid q,n) = -\sum_{x \in A_q} \sum_{x \in X} p(n,x,a) \log_2 p(x \mid n,a)$$

この値が小さくなる質問が効率的であるといえる。

・無矛盾性尺度

過去のユーザの発話履歴において既に入力したはずの項目の入力を再度要求すること、また誤った確認をすることは、ユーザの発話履歴と質問が矛盾しシステムの誤理解をユーザに知らしめることになる。そこで、現在保持している複数の理解候補のいずれかが正解であると仮定し、できる限り多くの理解候補と矛盾しない質問を選択することを考える。その評価値は以下のようになる。

$$S_c(q) = \sum_{n \in N} (1 - I(q,n)) \cdot P(n)$$

ここで質問 q が理解 n と矛盾する場合 $I(q,n)$=1、それ以外の場合 $I(q,n)$=0、$P(n)$ は理解仮説 n が正しい確率である。

最終的には、これらのバランスを取った値を最大とする質問をすることとした。

$$\hat{q} = \arg\max_{q} \{ w_c \cdot S_c(q) + w_e \cdot S_e(q) \}$$

④シミュレーション実験結果

ユーザと音声認識をシミュレートしてシステムと対話し、評価を行った。ユーザシミュレータは、目的とする楽曲の検索キー集合から適当に選んで入力を開始し、システムの質問に答えていく。その際に、音声認識はある設定された性能に従って誤認識をするものとした。各実験 1000 回ずつのシミュレーション対話を行った。評価の基準は、対話に要したターン数で、効率性を示す。比較対象として、毎回入力内容の確認を行う方法と、

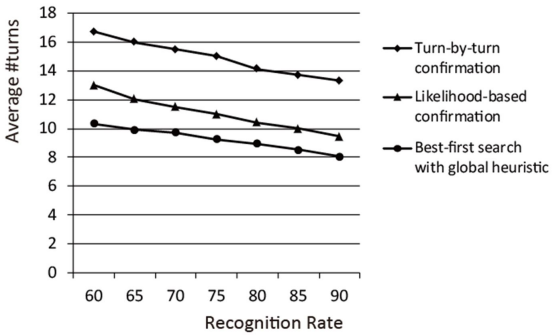

図２　対話戦略による効率性の評価結果

認識結果の信頼度が低い場合にのみ確認を行う方法を行った。結果を図２に示す。

様々な想定認識率において、提案手法の効率が良いことが分かった。また、これとは別途、システムをユーザが使用した時の主観評価を行い、これら三つのシステムの自然さを評価したところ、毎回確認するものが自然さが低く、その他二つは同程度に自然であると判断された。すなわち、提案手法は自然さを保ちながら効率的な対話が行えた。

（2）人間対人間の対話の認識に基づく楽曲連想検索

人間同士の対話には場の雰囲気や話題など多くの情報が含まれる。こうした情報は、この場にある対話システムが対話の仲間入りをする際には事前の情報として知っておくべきことであろう。本研究では、楽曲検索をターゲットとして、人間同士の対話を認識することによってその対話に合った楽曲を検索する技術を開発した。

① 音声認識結果と楽曲との関連付け手法

まず、対話を音声認識し、得られた結果に潜在意味解析を施して文書（この場合、認識結果）の類似度を表現できる文書ベクトル空間を構築する。一方、楽曲の音響信号から抽出される音響特徴量に基づいて、楽曲間類似度を表現する音響ベクトル空間を構築する。そして、これら 2 つの空間を線形変換で関連付け、相互に参照できるようにする。

文書ベクトル空間は次のように構築する。楽曲 m に関連する文書 j に含まれる単語の TF-IDF 値を要素とするベクトル $\mathbf{x}_j^{(m)}$を集め、行列 $\mathbf{X}$=($\mathbf{x}_1^{(1)}$, …, $\mathbf{x}_{J1}^{(1)}$, $\mathbf{x}_1^{(2)}$, …, $\mathbf{x}_{J2}^{(2)}$, …, $\mathbf{x}_1^{(m)}$, …, $\mathbf{x}_{JM}^{(M)}$)とする。$\mathbf{X}$ は $I \times J$ 行列であり、J は総文書数、I は文書ベクトルを作成するときに考慮する単語の総数である。そして、$\mathbf{X}$ の特異値分解 $\mathbf{X}=\mathbf{USV}^T$ により得られた直交行列 U のうち絶対値の大きな特異値に対応する上位 N 列を取り出した $I \times N$ 行列 $\mathbf{U}_N$ を用いて得られる次元削減されたベクトル $\mathbf{d}_j(m)=\mathbf{U}_N{}^T\mathbf{x}_j(m)$ を最終的な文書ベクトルとする。

一方、楽曲の音響信号を様々な特徴量に基づいてベクトル表現したものを音響ベクトルと呼ぶ。音響的な特徴としては、楽曲の音

色、音量、リズムを表すものから成る。楽曲からフレーム t 毎に楽曲 m の音響特徴量ベクトル $\mathbf{y}_t^{(m)}$を得、次に、すべての楽曲から計算されるこの特徴量ベクトルの集合をベクトル量子化し、楽曲ごとの量子化コードの頻度を曲全体の特徴ベクトルとすることで音響ベクトルとする。

最終的に、文書ベクトル **d** と音響ベクトル **a** を以下のような線形変換 **W** によって関連付ける。

$$\mathbf{a} = \mathbf{W}\mathbf{d}$$

W は事情誤差が最小となるように学習する。

$$\hat{\mathbf{W}} = \arg\max_{\mathbf{W}} \sum_{m=1}^{M} \sum_{j=1}^{J_M} \left\| \mathbf{a}^{(m)} - \mathbf{W}\mathbf{d}_j^{(m)} \right\|^2$$

② 主観評価実験

雑談音声をシステムに入力した場合に上位にランクされる楽曲が、雑談にふさわしいものであるかを判断する主観評価実験を行った。被験者 4 名は、入力した雑談に対する上位 10 件の楽曲を以下の基準で評価した。

・雑談の雰囲気に沿うか
・雑談に含まれる単語から連想されるか
・音声認識の結果の単語から連想されるか

また、提案された上位 10 件の楽曲リストのうち、最もふさわしいと判断される楽曲が、雑談内容にふさわしいかを、1（ふさわしくない）～5（ふさわしい）で評価した。

使用したデータは 671 曲からなる楽曲データベースとそれを説明したレビューブログ記事、貸しである。これらを用いて文書ベクトル空間、音響ベクトル空間を構築した。それぞれ 4096 次元に圧縮されている。音声認識には Julius を用いた。雑談は A～D の 4 会話である。

評価結果を表 1 および図 3 に示す。おおよそ半分の楽曲が雰囲気に合う、連想されると判定されていることが分かる。また、特にリスト中の最もふさわしいものについては、評価値が 4～5 と高いものとなっており、ほぼふさわしいものが候補として挙がっていることも分かる。これらの結果から、検索のための事前知識として用いるには十分な絞り込みができていると考えてよい。

表 1　提案された楽曲ごとの主観評価結果

会話	A	B	C	D
雰囲気に沿う	26/40	18/40	17/40	15/40
単語から連想	21/40	15/40	14/40	14/40
認識結果から連想	20/40	15/40	16/40	13/40

(3) 音声ドキュメント検索の高精度化

テキストの検索には、従来から 3 つの主な検索手法(ベクトル空間モデル，クエリ尤度モデル，適合モデルに基づく手法)が用いられてきた。しかし、音声を認識することによってテキスト化したものを検索する音声ドキュメント処理においては、音声認識の認識誤りや未知語の影響により、多くの誤りが含

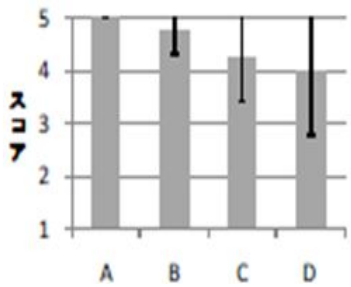

図 3　楽曲リストが雑談内容にふさわしいか否かの判定結果

まれることとなり、何らかの改良を加えて利用する必要がある。

そこで、本研究では、これらの手法に統一的な枠組みで改良を加えることで，音声ドキュメント検索における未知語や認識誤りに対処する手法を提案した．各検索手法に対し，新たな検索質問拡張手法，および音節 3-gram を単語と同様に扱う検索を単語単位の検索とスコアレベルで組み合わせる手法を開発した．

提案手法の有効性を、コンペティション型のワークショップ NTCIR-9 における SpokenDoc タスクで評価した結果，各手法で Baseline となる手法よりも検索性能が向上した．特に，確率に基づくクエリ尤度モデルに基づく手法と，適合モデルに基づく手法では検索性能が高かった．提案手法は NTCIR-9 で公表されている公式の最高精度の結果を上回る結果を得た．

(4) 楽曲間の類似判定における個人の類似性に対する許容度の導入

楽曲を検索する対話を想定すると、検索対象となる楽曲間の類似性を数値的に表現する必要が生じる。楽曲間類似度に関しては多くの音響特徴量が提案されてきており、それである程度表現できることは分かっている。しかし、各個人がもつ個性によって判定が異なることも多く、これを表現しない限り個人に適した楽曲の検索はできない。そこで、本研究では、各個人が楽曲のペアを類似判定する際にどの程度音響的に似ていれば類似していると判定するかを「許容度」と呼び、それを用いた類似性判定モデルを構築した。

楽曲のペアが類似するか否かの判定には、音響的な類似性と聴取者の許容度が影響すると考え、ある聴取者 i が楽曲ペア j を「似ている」と判定する確率を、ロジスティック関数を用いて以下のようにモデル化する。

$$p(e_{ij} \mid s_i, p_j) = \frac{1}{1 + e^{-(s_i + p_j)}}$$

ここで、eij は聴取者 i の楽曲 j に対する類似判定結果であり、eij=1 ならば「似ている」と判定したものである。Si, pj はそれぞれ聴取者 i の許容度と楽曲ペア j の類似度に相当するパラメータである。Si が大きいほど被験

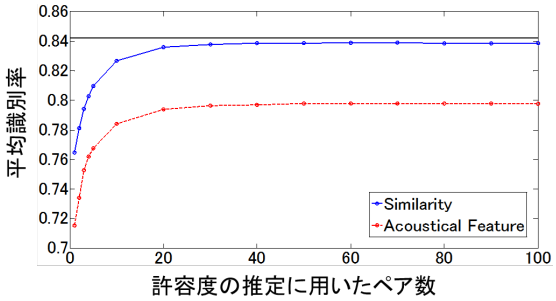

図４　許容度を用いた類似性判定推定結果

者 i は「似ている」と判定する可能性が高く、pj が大きいほど楽曲ペア j は似ていると判定される可能性が高い。被験者実験により集めた 27 名による 200 ペアの判定結果によりこの式の値の総積が大きくなるように si と pj を推定した。

この方法だと、未知の楽曲ペアの類似度が推定できない。そこで、音響特徴から本推定と同等の類似度を推定することも試みた。本研究で音響特徴量として用いたのは、短時間特徴量（メル周波数ケプストラム係数（MFCC）、インテンシティ、スペクトルセントロイド、スペクトルフラックス、スペクトルロールオフ、高周波数エネルギー）をベクトル量子化して楽曲ごとのクラスタのヒストグラムの対数である。そして、ヒストグラム間のユークリッド距離を楽曲ペア j 間の距離とし、

$$\hat{p}_j = -0.190 d_j + 4.543$$

によって類似度を推定した。

また、どの程度の楽曲ペアを用いれば許容度の推定が可能かを調べるために、楽曲ペア数を変えた推定実験も行った。

こうして推定された許容度と楽曲ペア類似度を用いて類似度判定結果を推定する実験を行った。その結果を図４に示す。青線が、式により推定された許容度と類似度で判定結果を推定（識別）した時の識別率、赤線が楽曲ペア間距離から楽曲類似度を推定した値を用いた場合の識別率である。横軸は推定に用いた学習ペア数であり、おおよそ 30 程度用いると飽和していることが分かる。また、200 ペアすべて用いた場合が黒線で書かれており、本実験の上限値と言ってよい。このように、本手法で楽曲ペアが個人の判定傾向も反映して、約 80%以上の精度で類似性判定の結果が推定できることが分かった。

こうした個人性を反映した距離尺度により音響空間を構築することによって、個人により適した楽曲の検索を可能にすることができると考えられる。

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４件）

① 北岡教英, 矢野浩利, 杉本夏樹, 山本一公, 中川聖一, 複数理解候補の保持と効率性・自然性を考慮した応答生成による誤認識に頑健な音声対話戦略とその評価, 電子情報通信学会論文誌（D）, 査読有, Vol. J85-D, 2012, 982-994

② 川渕将太, 宮島千代美, 北岡教英, 武田一哉, 楽曲間の類似判断における個人性データの収集とその分析, 情報処理学会論文誌, 査読有, Vol. 54, 2013, 1350-1361

③ Norihide Kitaoka, Yuji Kinoshita, Sunao Hara, Chiyomi Miyajima, Kazuya Takeda, A graph-based spoken dialog strategy utilizing multiple understanding hypotheses, Information and Media Technologies, 査読有, Vol. 9, 2014, 111-120

④ 陳 伯翰, 北岡教英, 武田一哉, 発話セグメントクラスタの評価とそれに基づく改良ボトムアップクラスタリングによる話者ダイアライゼーションの高精度化, 電子情報通信学会論文誌（D）, 査読有, Vol. J97-D, 2014

〔学会発表〕（計１２件）

① 大橋宏正, 柘植 覚, 北岡教英, 武田一哉, 北 研二, クエリ拡張と音節認識の統合による音声ドキュメント検索, 日本音響学会春季研究発表会, 2013.

② 川渕将太, 宮島千代美, 北岡教英, 武田一哉, 楽曲間主観的類似度データの収集実験, 日本音響学会春季研究発表会, 2012

③ 大橋宏正, 柘植 覚, 北岡教英, 武田一哉, 北 研二, 音声ドキュメント検索におけるクエリ拡張と音節認識の併用の効果, 電子情報通信学会音声研究会, 2012

④ Shota Kawabuchi, Chiyomi Miyajima, Norihide Kitaoka, Kazuya Takeda, Subjective similarity of music: Data collection for individuality analysis, APSIPA ASC 2012, 2012

⑤ Satoru Tsuge, Hiromasa Ohashi, Norihide Kitaoka, Kazuya Takeda, Kenji Kita, Spoken document retrieval using combinational use of distances of multiple vector spaces and query expansion with optimized weight parameters, NCSP'13, 2013

⑥ 市川 賢, 北岡教英, 柘植 覚, 武田一哉, 北研二, 単語空間と音節空間を併用した音声ドキュメント検索手法への潜在的意味解析の適用, 音声ドキュメント処理ワークショップ, 2013

⑦ 川渕将太, 宮島千代美, 北岡教英, 武田一哉, 楽曲間の類似判定における許容度の推定, 情報処理学会 SIGMUS 研究会, 2013

⑧ Shota Kawabuchi, Chiyomi Miyajima, Norihide Kitaoka, Kazuya Takeda, Modeling subjective evaluation of music similarity using tolerance, EUSIPCO, 2013

⑨ Ken Ichikawa, Satoru Tsuge, Norihide Kitaoka, Kazuya Takeda, Kenji Kita, Spoken document retrieval using both word-based and syllable-based document spaces with latent semantic indexing, APSIPA ASC 2013, 2013
⑩ **川渕将太, 宮島千代美, 北岡教英, 武田一哉, 北 研二, 楽曲間主観的類似判定における個人性分析手法の検討, 日本音響学会春季研究発表会**, 2014
⑪ **市川 賢, 柘植 覚, 北岡教英, 武田一哉, 北 研二, 音声ドキュメント検索手法における拡張クエリの超平面によるモデル化と潜在意味解析の適用, 日本音響学会春季研究発表会**, 2014
⑫ **北岡教英, 市川 賢, 柘植 覚, 武田一哉, 北 研二, 種々のテキスト検索モデルの頑健性向上による音声ドキュメント検索の高精度化, 音声ドキュメント処理ワークショップ**, 2014

６．研究組織

(1)**研究代表者**

北岡　教英（KITAOKA, Norihide）
名古屋大学・大学院情報科学研究科・准教授
研究者番号：１０３３３５０１

(2)**研究分担者**

武田　一哉（TAKEDA, Kazuya）
名古屋大学・大学院情報科学研究科・教授
研究者番号：　２０２７３２９５

宮島　千代美（MIYAJIMA, Chiyomi）
名古屋大学・大学院情報科学研究科・助教
研究者番号：　９０３３５０９２