Vol.2015-SLP-105 No.10
2015/2/28

ユーザの対話意欲自動推定を目標とした対話データの分析と音声画像特徴量の検討

千葉 祐弥[1,a)]　能勢 隆[1]　伊藤 彰則[1]

概要：対話型システムがユーザに適応して話題の提供や情報推薦を行うためには，ユーザの情報を効率的に獲得できることが望ましい．本研究では，ユーザに対して積極的に質問するインタビュー型の音声対話システムを想定する．このようなシステムとの対話では，ユーザが話したいと思う話題に関してはより詳細な情報が得られる可能性がある一方，ユーザが話したくない話題に関しては有益な情報が得られない可能性が高いと考えられるため，システムはユーザの対話意欲を考慮して質問や話題の選択を行う必要がある．本稿では，ユーザの対話意欲を自動推定するための初期検討として，人間同士のインタビュー対話の分析とその自動識別を行った．分析から，対話者自身が自分の対話意欲の高低を自覚できている場合，70〜80% 程度の精度で第三者にあたる評価者が対話意欲を判断できることが示唆された．また，評価者のアンケートに挙げられたマルチモーダル情報を利用することで，人間と同程度の精度で自動識別できることが示された．

1. はじめに

対話型システムはユーザの情報を獲得することで，それぞれのユーザに適応した親切な応答が行えるようになる．従来より，ユーザの発話から嗜好 [1] や感情 [2] などを推定する研究が多くなされてきた．一方，近年では，必ずしもタスクの達成を目指さない，所謂雑談型の音声対話システムが注目を集めている．雑談型の対話システムでは，対話の自由度がタスク遂行型のシステムに比べて高いため，タスクの設計に限定されない様々なユーザの情報を獲得できる可能性がある．本研究では，このようなユーザの情報を獲得する非タスク遂行型の対話システムとして，ユーザに対して質問することで積極的に情報を獲得するインタビュー型の対話システムを想定する．ユーザの状態を知るために質問を行うシステムは，カウンセリング対話システムなどで実現されている [3]．このようなシステムが効率よくユーザの情報を得るには，システムの質問に対してシステムが想定する以上の内容をユーザに発話してもらう必要があるが，システムが保持する質問集合からユーザの状態を顧みずに順不同で質問を提示するというような単純な方法は，ユーザの対話意欲を削いでしまうと考えられる．また，実際にはユーザが話したいと思う話題に関しては詳細な情報が得られる可能性があるが，ユーザが話したくない話題に関しては有益な情報が得られにくいと考えられるため，システムはユーザの対話意欲を考慮して質問や話題の選択を行う必要がある．

このような考察から，本研究では提示した質問や話題に対するユーザの対話意欲を推定できる対話システムの作成を目指す．しかしながら，ユーザの対話意欲が対話中のユーザの態度や発話に表出されるのか，またそれが対話相手に伝達するのかについての確証がないため，本稿ではまずはじめに人間同士のインタビュー対話の収集を行い，1) 対話相手の行動からどの程度対話意欲が推定できるか，2) どのようなユーザの行動が対話意欲の判断に有効か，を人間による対話の評価を元に分析した．また，様々なユーザの振る舞いが対話意欲の伝搬に関与すると考えられるため，本稿では評価者の内省報告を元にマルチモーダル情報の分析を行い，実際に識別実験を行った．以降では，ユーザの対話意欲の高低について，「話したい」，「話したくない」といった言葉で表現する．

2. 関連研究

本研究で扱う「対話意欲」に関しては，堂坂らの研究でも言及されており，ユーザにクイズを提示する自然言語ベースの思考喚起型対話において分析が行われている [4]．堂坂らは対話意欲を「もう一度使いたい」度合いを対話意欲が高い低いといった評価語でアノテーションを行なっている．また，対話意欲に近いものとして，「対話継続欲求」

[1] 東北大学
Aoba 1–2–3, Aramaki-Aza, Aoba-ku, Sendai 980–8579 Japan
[a)] yuya.chiba.p1@dc.tohoku.ac.jp

がある．宮澤らは人間同士の対話の分析から，音声対話システムにおけるユーザの対話継続欲求を満たす対話の典型パターンについて分析を行なった [5]．分析から，ユーザの発話にフィードバックを返す，ユーザの発話行動を阻害しないことなどが挙げられている．

本研究では，対話意欲の評価語としては堂坂らの定義よりも単純化し，提示した質問や話題について「話したい」か「話したくない」かとし，ユーザの対話意欲を質問応答対毎に評価することを最終的な目標とする．対話意欲は対話の盛り上がり [6] や活性度 [7] とも関連があると考えられるが，これらの研究ではある程度の区間で盛り上がりや活性度を定義しており，質問応答レベルでの局所的な判断が難しい．部分発話毎にユーザの状態を推定する研究として，実験者が提示した話題や説明へのユーザの興味の度合い (Level of Interest; LOI) を推定する研究がある [8], [9] が，対話意欲は質問内容や質問者の態度などの影響をうけるため，必ずしも興味と同一のものではない．

また，我々の研究ではユーザが積極的に話す場面を想定しているが，この点では通常の雑談対話 [6] や，目黒らの扱う聞き役対話 [10], [11] に本質的には近い対話であると言える．特に，目黒らは聞き役対話システムの研究において，ユーザの「話したい」という欲求を満たすため「聞いてもらっている」という感覚が得られる対話システムを目指し，対話データの分析を行なっている．聞き役対話では，聞き役の自己開示の出現率は減るものの，質問の前に自己開示を行うことで話し役の自己開示を促すことができるとしている．我々の研究では，ユーザの対話意欲そのものを推定することが目的であり，マルチモーダル情報を利用した識別を行う．マルチモーダル情報の利用に関してはロボット対話 [12] やマルチモーダルインターフェース [13] などで様々な議論がなされてきた．これに関しては，我々も「考えている」,「戸惑っている」などのユーザの状態を推定する目的で，対話中のユーザが表出する音声の韻律的な情報や表情，視線などのマルチモーダル情報を用いた実験を行ってきた [14].

3. インタビュー対話データの収集

3.1 実験条件

一方の対話者が他方の対話者に対して質問するインタビュー形式の対話を収集した．本稿では，著者のうちの 1 名が質問者 (システム役)，実験参加者が回答者 (ユーザ役) を行った．最終的にはエージェントを有する対話システムの作成を目指しているが，ユーザのエージェントに対する親しみが低い状態を想定しているため，分析の容易さも踏まえ質問者と実験参加者は初対面かつ同性同士とした．実験参加者は 3 名の大学生で，全員男性であった．以下，それぞれ TK，NI，IT と記述する．質問者は，実験にあたって 1) 質問と回答のペアができるだけ明確になるよう，自分からバージインをしない，2) 対話相手の対話意欲を削がないように，適当な箇所で相槌をうつ，3) なるべく相槌や同意，相手の発話の確認などで対話を進行するようにする，といったところに注意して対話を行った．データの収録は静寂な環境で行った．質問者と実験参加者は対面して座り，両者の間には Web カメラ (LogiCool QCAM-200R) を設置し，回答者の対話中の動作を正面から撮影した．収録した動画像データは 640×360，15 fps のカラー画像 WMV ファイルとして保存した．

3.2 対話の話題

本稿では，表 1 に示す話題に関する質問を行った．対話の話題は [6] や [10] で選択されているものを参考にし，学生が対話実験に参加することを考慮した上で，できるだけ実験参加者の興味が偏らないよう様々なジャンルから 10 個の話題を選んだ．

表 1 インタビューの話題

旅行	音楽	乗り物
健康	映画	研究
料理	スポーツ	コンピュータ
ファッション		

3.3 対話データ収集の手続き

被験者には対話システムの作成のために人間同士の会話を分析するという旨を伝えた上で対話を行った．回答者はあらかじめそれぞれの話題に関する興味の度合いを 5 段階で評価する．評価値は 5:「とても興味がある」，1:「全く興味がない」であった．各話題は実験者がこれ以上新しい情報が得られないと判断したところで打ち切り，10 個の話題に関する対話が終了した時点で実験を終了した．

実験後，あらためて回答者にはアンケートを実施した．アンケート項目は

Q1: それぞれの話題についてどの程度話したいと感じましたか
Q2: 各話題についてどの程度話せたと思いますか
Q3: 対話全体を通しての満足度はどの程度ですか
Q4: 対話はどの程度楽しかったと感じますか

であった．評定は全て 5 段階で行なってもらい，Q1 に関しては 5:「話したいと感じた」，1:「話したいと感じなかった」，Q2 に関しては 5:「十分に話した」，1:「話し足りなかった」，Q3 に関しては 5:「満足」，1:「不満」，Q4 に関しては 5:「楽しかった」，1:「楽しくなかった」がそれぞれ対応している．以下では Q1 の評価値を対話意欲の本人評価と読み替える．また，Q3 と Q4 の質問に関しては [10] で指摘されている通り，システム役が質問し続けると尋問調の対話になってしまうことが想定されるため，どの程度対話が快適に行われていたかを調べる目的で行った．実験に

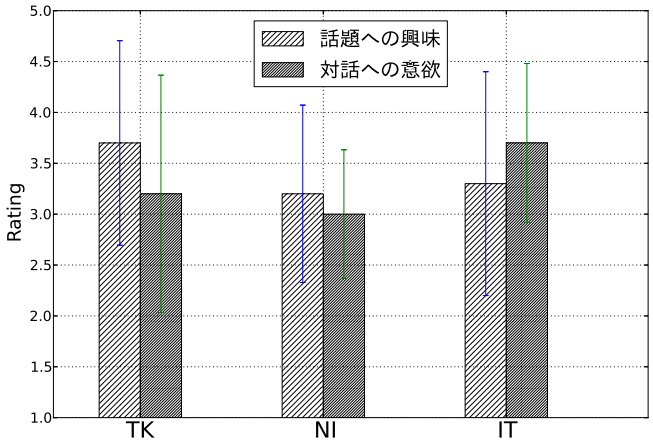

図 1　実験参加者の自己評価

より収集されたデータの総時間は 52 分 42 秒であった．

3.4　対話収集実験の結果

収集した対話データの例を表 2 に示す．本稿では 3 名の実験参加者に対して 10 個の話題に関する対話を行ったため，$3 \times 10 = 30$ の対話データが収集された．しかしながら，実験参加者 NI の「料理」に関する対話の収録に不備があったため，以降では残りの 29 対話のデータを扱う．

3.5　実験参加者のアンケート集計結果

対話実験参加者の各話題への興味及び対話意欲の自己評価の平均値及び標準偏差を図 1 に示す．誤差棒は標準偏差である．図より，各話題への興味と対話意欲には関連が見られるが，必ずしも一致していないことがわかる．これは，質問者の質問の振り方や対話態度などが影響していると考えられる．本稿では，ユーザの表層に現れる情報から対話意欲を推定することが目的であるので，どうすればユーザの対話意欲を高められるかについては検討を行わない．しかしながら，対話全体の満足度に関する評価の平均値及び標準偏差は 4.00 ± 0.82，楽しさに関する評価は 4.67 ± 0.47 と，いずれの実験参加者も 3 点以上の評定をつけており，対話そのものは快適に行われていたと言える．どの程度話せたかに対する評価も概ね高い値であった．

4.　対話データの評価実験

4.1　評価実験資料の作成

分析を簡単にするため，本人の対話意欲の評定が高かった対話データと低かった対話データのみを評価に用いる．本稿では，それぞれの実験参加者に対して評定が中央値より高い点数がついているものを「話したい」対話データ，低い点数がついているものを「話したくない」対話データとして扱った．ここで，それぞれの実験参加者の評定の中央値は TK:3，NI:3，IT:4 であった．中央値の評定がついた対話データを除いて最終的に残ったのは 19 対話であった．評価実験に利用するデータの内訳を表 3 にまとめる．

表 3　評価用対話データの内訳

	対話意欲高	対話意欲低	Total
TK	5	3	8
NI	2	2	4
IT	2	5	7
Total	9	10	19

4.2　評価実験の手順

対話者の対話意欲はどの程度対話相手に伝わるのか，及び対話を質問応答対に分割しても評価することができるのかを調べるため，以下の 2 つの評価実験を行った．評価実験には 3 名の評価者（男性 2 名，女性 1 名）が参加した．以下，E1，E2，E3 と呼ぶ．評価者には本研究で目標としている対話システムの説明を行い，よく理解してもらった上で実験を行った．それぞれの評価者はいずれの回答者とも面識がなかった．

評価者に 19 個の対話データを提示し，評価者はそれぞれの対話データについて回答者の対話意欲を「話したい」，「話したくない」のどちらかから判定した．質問は「ユーザはこの話題についてどう感じていると思いますか」であった．実験後，評価者の評価と回答者の本人評価との一致率を計算し，第三者が回答者の対話意欲をどの程度判断することができるのかを分析する．

本稿では回答者間の個人差は議論しないため，評価者は同一の回答者のデータを連続して評価し，各対話データは何度視聴しても良いとした．同時にシークバーによる再生箇所の選択も許可した．ただし，話題の提示順に関しては順不同であった．また，活発に対話を行う回答者は，あまり話せない話題に関する質問については積極的に話題を変えようとする傾向が観察されたため，アプリケーションの下部には現在の質問が何の話題に関するものであるかを表示した．

4.3　評価実験の結果

評価実験 1 によって得られた評定の評価者間の一致率と，対話実験参加者の本人評価との一致率を表 4 にまとめる．表の各要素の上段は評定の純一致率を示し，下段は Cohen の κ 係数を示す．ここで，Cohen の κ 係数は評定の一致率を評価する尺度である．表より，それぞれの評価者の評定と本人評価の純一致率は平均すると 7 割程度であり，対話者が対話意欲の高低を自覚している対話（すなわち，本人評価が高かった対話と低かった対話）に関しては，ある程度対話観察者に対話意欲が伝わると考えられる．特に，3 名の評価者の多数決をとった評定の場合は $\kappa = 0.582$ と最も高い一致係数が得られた．しかしながら，それぞれの評価者の評定と本人評価を比較すると，E1 の評定と本人評

表 2　収集された対話データの書き起こし例 (NI)

話題	質問応答対	話者	発話
乗り物	T1	I1	乗り物とかにはあまり興味はないの？
		S1	乗り物はそうですね．本当乗れれば何でもいいんで．
	T2	I2	移動ができれば？
		S2	移動ができて，まあ，寄ったりとか，変にがたがたしなければ．
	T3	I3	原付とかは？
		⋮	
	T7	S7	まあ自家用車で行くか，それか電車とかで行った方が良いかなって感じです．
映画	T1	I1	最近見た映画で面白かったものは？
		S1	最近見た映画ですか．そうですね，何個かあるんですけど，やっぱりメジャーなところだと ···
		⋮	
	T17	I17	そっか，じゃあ最近のはいまいちなんだね．
		S17	あー，どうですかね．最近のもたまーに当たりがあるんですよ．〈えー、おもしろいな．〉だからちょっと見ちゃうんですけど．

表 4　対話データの評価の一致率（下段は κ 統計量）

	E1	E2	E3	多数決
本人評価	0.789	0.737	0.684	0.789
	(0.573)	(0.481)	(0.380)	(0.582)
E1	—	0.632	0.684	—
		(0.311)	(0.424)	
E2		—	0.737	—
			(0.417)	
E3			—	—

表 5　分析ラベル（多数決）

	対話意欲低	対話意欲高	Total
TK	2	6	8
NI	3	1	4
IT	3	4	7
Total	8	11	19

価との一致率が $\kappa = 0.573$ で中程度の一致率を示す一方，E3 は $\kappa = 0.380$ と低い一致率を示しており，対話意欲の評定には個人差が認められる．

5. マルチモーダル情報の分析

5.1 評価者の内省報告

対話意欲の評価基準に関する評価者アンケートでは，言語情報や音声の韻律的情報，交代潜時，視線，表情，ジェスチャに関するコメントが得られた．特に言語情報は，高い対話意欲の対話に関して，聞かれた内容だけでなくそれ以上の内容を話している，特定の固有名詞が発話されているといったことが挙げられた．また，表情が笑顔であるか，ジェスチャの大きさが大きいかなどの評価指標や，音声に関しても発話の抑揚の大きさが対話意欲の推定に影響するという指摘があった．その他には，交代潜時の長さは，長い場合に対話意欲が低く，バージイン気味に発話するなど，特に短い場合には対話意欲が高く感じられるといった報告があり，視線に関しては質問者の方を向いていると対話に興味がありそうに感じられるといった報告があった．以上のアンケート結果より，それぞれの特徴量について分析を行った．ここでは，センサーデータの都合上，視線と交代潜時以外の特徴量に着目する．ラベルに関しては被験者の多数決の結果を利用する．表 5 に評価者のアノテーションの多数決の結果を示す．

5.2 言語情報及び音声情報の抽出

収録された動画像情報から，質問者及び回答者の発話を書き起こし，形態素解析を行うことで発話に含まれる品詞情報を抽出した．形態素解析エンジンには MeCab を用い，IPA 辞書を辞書データとして用いた．固有名詞に関しては，解析が容易でないと考えられる 24 エントリをユーザ辞書に追加した．

また，発話の抑揚の変化について分析を行うため，音声から 10 ms ごとに $F0$ を抽出した．$F0$ の取得には Snack を用いた．

5.3 画像情報の抽出

画像情報については，表情と身振りの変動の分析を行った．ある時刻における表情変動量は，Constrained Local Model (CLM)[15] によって検出された当該フレームの顔領域に対してオプティカルフローを計算し，その大きさの総和で定量化した．例を図 2 に示す．図では，2 ピクセルおきにオプティカルフローの方向と大きさが描画されており，瞬きによる右眼領域の縦方向の変動と発話による口唇領域の縦方向の変動が観測できる．

また，本稿では差分画像として計算された変動を身振りの変動として扱った．CLM によって得られた顔領域をマスクした上で，フレーム全体の差分画像を計算し，その変動量の大きさを当該時刻の身振り変動量とする．図 3 に得られた差分画像を示す．当該の時刻では対話者が両手を下

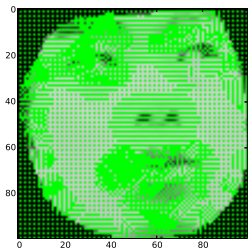

図 2　顔向きの変動

図 3　身振りの変動画像

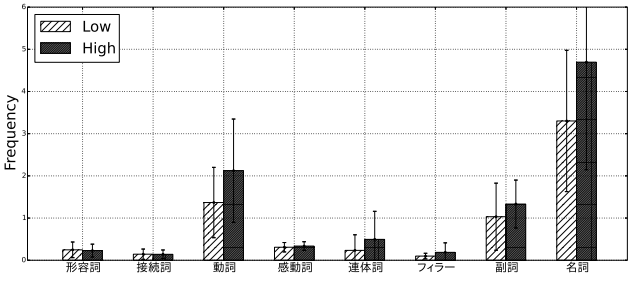

図 5　各品詞の頻度の分布

げているため，その変動の様子が差分画像によって観測できる．

5.4　マルチモーダル情報の分布

図 4 として各マルチモーダル情報の分布を示す．図より，形態素数，顔変動量，身振り変動量は，「話したい」と評価された対話データの平均値に関して「話したくない」と評価された対話データを上回っており，「話したい」と評価された対話データは回答者が身振りや表情の変動を交え，より活発に話している傾向にあることがわかる．一方で，$F0$ に関しては対話意欲が高いデータと低いデータで大きな違いは見られず，「話したい」対話データにおいて抑揚が大きいという事実は確認できなかった．それぞれの特徴量に対して被験者とラベルを要因とした繰り返しのある二元配置分散分析を行ったところ，顔変動量に関しては有意差が得られ（$F = 7.731, p = 0.01560$），身振りの変動量に関しては有意傾向であった（$F = 3.509, p = 0.0837$）．このことから，特に顔変動量，身振り変動量が対話意欲の推定に有効であると言える．また，$F0$ の平均値に関してもラベル要因に関して有意傾向（$F = 3.761, p = 0.0546$）が得られたが，これは各発話の平均値をサンプルとしており，他の特徴量に比べてサンプル数が多いことが原因であると考えられる．

また，品詞タグの出現頻度を図 5 に示す．グラフはそれぞれの話題における発話あたりの品詞の出現頻度の平均値と標準偏差を示している．図より，全体としては「話したい」と評価された対話データにおいて「動詞」や「名詞」に分類される単語が増加していることがわかる．これは「話したくない」と評価された対話データでは，質問者の質問に対して「はい」や「そうですね」といった簡単な発話で応対が終了するのに対して，「話したい」と評価された対話では回答者の具体的な意見が伴う発話が多いからである．

表 6　識別に用いたマルチモーダル特徴量

モダリティ	特徴量	備考
言語	「動詞」頻度	発話数で正規化
	「名詞」頻度	発話数で正規化
	「副詞」頻度	発話数で正規化
	形態素数	–
音声	$F0$	セッション平均値の平均
画像	表情変動	時間平均
	身振り変動	時間平均

表 7　識別精度（識別結果の一致率）

	純一致率	κ 統計量
本人評価	0.784	0.568
E1	0.895	0.759
E2	0.526	0.159
E3	0.579	0.283

6.　SVM を用いた識別実験

最後に，分析に用いた特徴量を用いて Support Vector Machine（SVM）を用いた識別実験を行った．本稿では，表 6 に示す 7 次元の特徴量を採用した．

それぞれの特徴量は個人差の影響を除くため各対話者の平均を減算したものを用い，本人評価を正解ラベルとして学習を行った．本研究では，対話システムとして対話中に対話相手に適応する枠組みを想定しているため，妥当な処理であると言える．SVM のカーネルには RBF カーネルを用い，グリッドサーチによってパラメータを決定した．実験サンプルが少ないため，実験は Leave-one-out 法によって行った．識別結果と各評価者の評価結果との一致率を表 7 として示す．

結果を表 4 と比較すると，SVM による識別結果は本人評価との一致率が 78.4%，κ 統計量が 0.568 と，人間による評定とほぼ同程度の精度で一致していることがわかる．一方，それぞれの評価者との一致率を見ると，評価者 E1

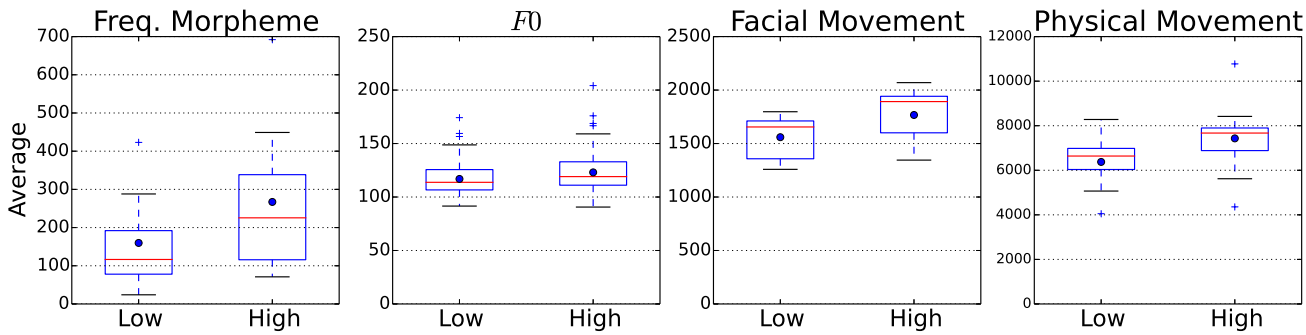

図 4　各マルチモーダル情報の分布

との一致率が高いが，評価者 E2，E3 との一致率は低く，必ずしも人間のような評価を行っているわけではないと言える．

7. まとめと今後の予定

本稿では，積極的にユーザの情報を取得する対話システムとして，ユーザに対してインタビューを行う対話システムを想定し，対話データを収集した．対話システムがユーザの話したいという態度を自動で推定するための初期検討として，人間同士の対話データを分析した．人間による評価実験では，対話者自身が対話意欲の高低を自覚できている話題の会話に関して 70〜80% 程度の精度で第三者が回答者の対話意欲を判断できることが示された．続いて，これらのマルチモーダル情報に関して実際に特徴量の抽出を行い，それぞれがどのような分布になっているのかを示した．分析により，特に顔変動量に関してラベル要因による有意差が得られ，$F0$ 特徴量および身振りの変動量に関して有意傾向であった．このことから，回答者が身振りを交えた活発な対話を行っていると，第三者は回答者の対話意欲を「話したい」と評価する傾向にあると言える．最後に，これらの特徴量を用いて SVM による識別実験を行ったところ，本人評価との一致率は 78.4% であった．これは，人間による評価とほぼ同程度の一致率である．

今後は実際にマルチモーダル情報を扱う対話システムを構築し，より多くのユーザとの対話データを収集・分析する予定である．

謝辞

本研究は日本学術振興会特別研究員奨励費 263989 の助成を受けた．

参考文献

[1] A. N. Pargellis, H. K. J. Kuo, and C. H. Lee, An automatic dialogue generation platform for personalized dialogue applications, Speech Communication, 42:329–351, 2004
[2] A. Metallinou et. al., Context-sensitive learning for enhanced audiovisual emotion classification, IEEE Trans. on Affective Computing, 3(2):184–198, 2012
[3] T. W. Bickmore, D. Schulman, and C. L. Sidner, A reusable framework for health counseling dialogue systems based on a behavioral medicine ontology, Journal of biomedical informatics, 44(2):183–197, 2011
[4] 堂坂浩二，奥梓，東中竜一郎，南康浩，前田英作，思考喚起型対話におけるユーザ対話意欲の分析，人工知能学会全国大会，2011
[5] 宮澤幸希，常世徹，桝井祐介，松尾智信，菊池英明，音声対話システムにおける継続欲求の高いインタラクションの要因，電子情報通信学会論文誌 A，95(1):27–36, 2012
[6] 徳久良子，寺嶌立太，雑談における発話のやりとりと盛り上がりの関連，人工知能学会論文誌，21，pp.133–142，2006
[7] 守屋悠里英, 田中貴紘, 宮島俊光, 藤田欣也, ボイスチャット中の音声情報に基づく会話活性度推定方法の検討，ヒューマンインタフェース学会論文誌，14(1):283–292，2012
[8] B. Schuller et. al., Audiovisual recognition of spontaneous interest within conversations, In Proceedings of the 9th international conference on Multimodal interfaces, pp.30–37, 2007
[9] W. Y. Wang and J. Hirschberg, Detecting levels of interest from spoken dialog with multistream prediction feedback and similarity based hierarchical fusion learning, In Proc. SIGDIAL, pp.151–161, 2011
[10] 目黒豊美，東中竜一郎，堂坂浩二，南泰浩，聞き役対話の分析および分析に基づいた対話制御部の構築，情報処理学会論文誌，53(12):2787–2801，2012
[11] T. Meguro, R. Higashinaka, K. Dohsaka, Y. Minami, and H. Isozaki, Analysis of listening-oriented dialogue for building listening agents, In Proc. SIGDIAL, pp.124–127, 2009
[12] P. McGuire et. al., Multi-modal human-machine communication for instructing robot grasping tasks In Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems, pp.1082–1088, 2002
[13] M. Pntic and JM Leon Toward an affect-sensitive multimodal human-computer interaction, In Proc. IEEE, 91(9):1370–1390, 2003
[14] Y. Chiba, T. Nose, A. Ito, and M. Ito, User Modeling by Using Bag-of-Behaviors for Building a Dialog System Sensitive to the Interlocutor’s Internal State, In Proc. SIGDIAL, pp.74–78, 2014
[15] J. M. Saragih, S. Lucey and J. F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Computer Vision, 91, pp.200–215, 2011
[16] G. Farnebäck, Two-frame motion estimation based on polynomial expansion, Image Analysis, pp.363–370, 2003