一般社団法人 人工知能学会
Japanese Society for
Artificial Intelligence

人工知能学会資料
JSAI Technical Report
SIG-Challenge-049-8 (11/25)

アンドロイドの動作生成に向けた自然対話中のジェスチャの認識および分類に関する検討

町屋敷　大地 , 石井カルロス寿憲, 劉超然, 石黒浩
Daichi Machiyashiki, Carlos Toshinori Ishi, Chaoran Liu, Hiroshi Ishiguro
ATR(石黒特別研究所)/大阪大学, ATR(石黒特別研究所), ATR(石黒特別研究所), ATR(石黒特別研究所)
ATR HIL/Osaka University, ATR HIL, ATR HIL, ATR HIL
machiyashiki.daichi@irl.sys.es.osaka-u.ac.jp

Abstract

ロボットの動作生成を目指して, 人の対話と同時に起こるジェスチャーの分類とそれらのジェスチャー中の手の位置や, 発話との関係を調査した. ジェスチャーの分類はアノテーターによってラベルづけされ,k-means 法によってジェスチャーの手の動きのクラスタを生成した. また Wordnet からジェスチャーとともに現れる発話の上位概念を取得した. これらの取得したデータの関わりを今後も調べていき, ロボットの動作生成へとつなげていく.

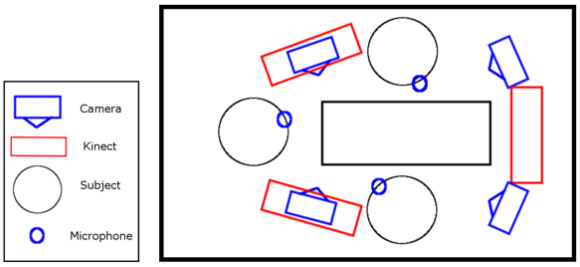

図 1: Environment setup for 3 person dialogue experiment.

1　はじめに

近い未来, アンドロイドなどのロボットがより大衆的になり, 社会の中での人に役割の一部を代替する場面が増加すると思われる. その役割の一つとして, 人と向き合って対話を行う人間の役割の代替が考えられる. その場合テキストのみによる対話システムと異なり, 対話の中に言語情報だけでなく非言語情報も含まれる. 人間とロボットが対面して自然な対話を行うためには, それらの情報からロボットが相手の伝えたいことを正しく理解したり, 自分の表現したいことが相手に伝わるように表現できることが望ましい. 当研究室では, 発話に伴う口唇・頭部・表情・腰部の動作生成に関する研究をこれまで数多く報告してきた [2][3][4][5][10][11]. 当研究では, 対話の中で出現する非言語情報のうち発話とともに現れるジェスチャーに焦点を当て, その動作生成を目的として, 人間の対面対話中に現れるジェスチャーの分析を行った.

対話の中におけるジェスチャーにはすでにたくさんの研究が行われているが, ジェスチャーをその機能ごとに分類する方法として有名なものは McNeill によるものである [13]. McNeill はジェスチャーのうち同時に話された言葉に関連する動きを映像的なジェスチャー, 隠喩的なジェスチャー, 拍子のジェスチャー, 指示を表すジェスチャーの 4 種類に分類した. 映像的なジェスチャーとは具体的な物体を表すものであり, 例えばリンゴの形を表すためにリンゴの形を手で表現するジェスチャーが当てはまる. 隠喩的なジェスチャーは抽象的な概念を表出するジェスチャーであり,「行く」などの動作や炎や心などの不定形なものを表すのに用いられる. 拍子のジェスチャーは対話の強調など目的とした手や指の振動で, 指示を表すジェスチャーは指さしジェスチャーが該当する. また複数の種類の重複が許されている. この分類のうち映像的なものを用いて CG エージェントの動作を生成する取り組みを LÜcking らが, 隠喩的なものに関しては Wordnet を用いて門野らが行っている [12] [6].

人間の代替を目的とする場合, 自然なふるまいを行うには, 特定の種類のジェスチャーだけでなく全ての種類のジェスチャーの生成を行う必要があり, さらに対話中の人間の動作には言語と関係のないアダプターと呼ばれる癖の動作が含まれる. この研究では, それらのジェスチャーの種類と癖がそれぞれどのようなときに現れ, 実際の手の動きとどのような関係があるのか, 個人差の影響はどの程度あるのかを分析し, ハンドジェスチャーが可能なアンドロイドのジェスチャーを生成することを最終目的とする.

2　対話データの取得

人間同士による自然な対話データの収集を目的に対話実験を行った. 様々な対話状態に対応したシステムを作るため,3 人による自由対話を記録し, そのジェスチャーを分析した. これは,2 人での対話では発話相手が決まっていて 1 人が発話するともう 1 人が必ず対応しなければならないなど状況が限定されるからである. 今回分析したのは 3 組による対話で, 被験者は互いに面識がある. 年齢層は 20 代から 40 代の男女からなる. 実験を複数回行った. 本稿ではそのうち拍子に関するもの以外のアノテーションが完了した 5 人分のデータを使用する.　実験環境は図 1 に示されるように四角いテーブルの 3 面に一人ずつ座った状態で, 約 45 分間自由対話をしてもらった. 途中, 座っている席の位置によるデータの偏りを防ぐため席順を 2 度変更した.4 台のカメラによって各被験者を撮影し,3 台のキネクトにより骨格情報および映像を, 各被験者が各々に装着したマイクロフォンと机に設置したマイクロフォンアレイから音声情報を記録した. また, 実験に使うテーブルは被験者の手が見えるように十分低くなっている.　実験後, 記録した動画に Openpose[14] を使用し, 全身の骨格 18ヶ所および, 手 21ヶ所, 顔 70ヶ所の 2 次元位置データを取得, 記録した. 座標系は水平軸を x 軸, 垂直軸を y 軸とし, データの範囲は 0 から対象となる動画の画素の各軸に対する数までとなる.

3　データ処理

3.1　対話中に現れる手の動きの分類

本研究では得られた実験データ中に現れる手に関連する動きを次の要素にしたがって定義し, 分類した.

1. ジェスチャーと癖
   対話と関連して現れる動きをジェスチャー, それ以外を癖とする. 例えば対話と関係なく自分の体を触る動きや, 姿勢の変更に関連して発生する手の動きは癖に分類される. また, 動きがない状態をホーム, その時の手の位置をホームポジションと呼ぶ.

2. ジェスチャーのフェーズ
   ジェスチャー動作には一連の流れが存在する. 本研究ではそれらを構成する要素をジェスチャーのフェーズと呼び,Kendon[8] の分類法に従って以下の 4 つのフェーズに分類する.

   - 準備
     ジェスチャーを行うために, ホームポジションから手を動かすフェーズ. 存在しない場合もある.
   - ストローク
     ジェスチャーのメインとなる意味がある動きを行うフェーズ.
   - 終わり
     準備フェーズとは反対に, 手の位置をホームポジションに戻す動き.
   - ホールド
     ストロークが起きる前後に手がホームポジションではない場所にとどまっているフェーズ.

3. ストローク内でのジェスチャーの機能
   McNeill[13] の分類のうち映像的, 隠喩的, 拍子, 指示とエンブレムの 5 つのカテゴリに分類する. エンブルムとは OK サインや V サイン, バイバイの動作など形と意味が社会的に定まっている動きである. さらに, 映像的, 隠喩的, 指示のジェスチャーに関して, 先行研究 [6] や書籍 [9] を参考に一部変更を加え, より詳細なサブカテゴリを定義した. 詳細を Figure 2 に示す. McNeill の定義では, 映像的, 隠喩的, 拍子, 指示の各カゴテリは重複可能であるが, 本研究では, 映像的, 隠喩的,, 指示それぞれと拍子との重複のみを考える. 拍子に関するラベルは以下のとおりである.

   - 動き
     - 単発
       上下運動一回のみ
     - 連続
       上下運動が連続して起こる
   - 意味
     - 強調
       強調部分で行われるビート
     - リズム
       単にテンポよく行われるビート

3.2　対話中に現れる手の動きのアノテーション

実験で得られたデータに対して 1 人のアノテーターが映像と音声によるジェスチャーの区間と機能のラベル付けを 3.1 章で定義したカテゴリに従って行った. 以降このラベルをジェスチャーの機能ラベルと呼ぶ. また, ジェスチャーと発話が同時に発生した場合, その発話を含む一文を書き起こした.

3.3　ジェスチャーの動作の特徴抽出

実験で得られたデータから各ジェスチャーのストロークフェイズにおける手の振る舞いについて調査する. 手の動きは複雑かつ多岐にわたるため基準となるクラスタの生成を試みた. Openpose から得られるデータは各関節に対応した画素の絶対位置であるため, 話者ごとにカメラからの距離や体格差によって同じ動きが違ったデータになること

映像的	図像	物の形や大きさを形取る
	図像2	実在する人などを表すが、形をなぞっていない（丸く描く等）もの
	動作（映像的）	イメージがつきやすい、具体的な動作（行く、食べる、歩く）
暗喩的（隠喩的）	名詞（名_隠喩的）	形の定まらない名詞、単語の概念（記憶、余裕、経験）、みんな等集団を表すもの
	動作（動_隠喩的）	イメージがつきにくい、抽象的な動作（...になる、夜が明ける、経験する）
	修飾（修飾）	「優しい」、「きれい」などの修飾語
	時間（期間）	「1年間」などの一定期間
	時間（時点）	「今」、「昨日」などの時間の一時点
	考え（否定）	「否定」を表すジェスチャー
	考え（考_その他）	その他の考え
	関係（関係）	物と物の関係性（年齢が上、先輩で...）
	様子（程度）	「ちょっと」、「いっぱい」などの物事の程度
	様子（擬音）	「ぐちゃぐちゃ」、「ぶりぶり」、「スペスペ」などの擬音、擬態
	様子（様_その他）	その他の様子
指示	指示（自分）	指さし動作。指さしのストロークは区切りにくい、指示する単語の発話のタイミングを考えながら区切る方法をとる。 自分を差すときは指示（自分）とし、相手を指差す場合に相手の被験者ナンバーで差した相手を判断する。
	指示（C01～C03）	
	指示（指_その他）	
慣習的（エンブレム）	慣習的（エンブレム）	形と意味が社会的慣習として定まっている（OKサイン、バイバイ、指で数を表す）

図 2:　Gesture categoriesThe list of categories and subcategories of gesture meanings in stroke phase

を防ぐため, 平滑化によるノイズ除去を行った後, すべてのデータを首元からの相対位置に変換した. さらに, 文献 [1] を参考にして x 軸に対して右肩から左肩の距離を 1,y 軸に対してのど元から腰の長さが 1 となるようにスケーリングを行った. 本稿では, 各腕の動きとして各手首の軌跡を用いる.

アノテートされたラベルから各ジェスチャーのストロークフェイズでの軌跡を取得する. 取得した軌跡は各腕に対して x 方向, y 方向, 時間の 3 軸からなる 3 次元データとなる.　ジェスチャーの動作をその機能や同時に発言された言葉と対応づけるためには, 何らかの手法で, ジェスチャーの動きを分類しなければならない. 本研究では k-means によるクラスタリングを行う. k-means とは d 次元の n 個のベクトルの観測 $(x_1, x_2, ..., x_n) \in$ x と k 個のクラスタ $(c_1, c_2, ..., c_k) \in$ C が存在する場合

$$argmin_C \sum_{i=1}^{k} \sum_{\mathbf{x} \in \mathbf{C_i}} cost(\mathbf{x}, \boldsymbol{\mu}_\mathbf{i}) \qquad (1)$$

となるクラスタを計算アルゴリズムである. ただし,$cost(\mathbf{a}, \mathbf{b})$ はベクトル $\mathbf{a}$ と $\mathbf{b}$ の距離を測る何らかのコスト関数で,$\boldsymbol{\mu}_i$ はクラスタ c_i に含まれるベクトルすべての平均値である. k-means に用いるコスト関数として, 一般的にはユーグリッド距離が使用される. ジェスチャーの軌跡を扱うにあたって, 時系列データである軌跡に一般的なユーグリッド距離を用いた方法はジェスチャーには時間的は伸縮が考えられるため有効ではない. 次に時間的な伸縮にロバストな Dynamic Time Warping(DTW) を使用することが考えられるが, ジェスチャー開始時点での手の位置や動きの大きさによる影響によりうまくいかなかった. そこで, 本研究では, ジェスチャーの軌跡を x-y 平面に射影し, その写像の類似度をコスト関数とする k-means を用いることでクラスタの生成を試る. 写像する平面の範囲は,x 軸に関して肩幅の 2 倍,y 軸に関して肩から腰の下までとし, 計算量の観点から写像した平面の画像を 64*48Pixel に縮小した, また, 縮小する際軌跡が途切れないよう縮小前の軌跡の幅を図 3 で示したものの 10 倍とした. 写像の類似度の計算には比較する点の位置のずれにある程度ロバストな SSIM を用いた. この操作により生成したクラスタをジェスチャー動作クラスタと呼ぶ.

3.4　WordNet による概念抽出

ジェスチャーの機能ラベルや動作の特徴と対話に出現する単語の関係を調べるため, 先行研究に従って WordNet[7] による単語の概念抽出を行った.WordNet は単語を synset と呼ばれる類義関係のセットでグループ化していて, 一つの synset が一つの概念に対応している.Wordnet にはこれらの一つ一つの概念に上位の概念や下位の概念といったほかの概念との関係性が記録されており, これを利用することで上位語 (Hypernym) や下位語 (Hyponym) などを取得することができる. 本研究では先行研究と同様に単語どうしの関連度を測るために国立研究開発法人情報通信研究機構が提供する WordNet(http://compling.hss.ntu.edu.sg/wnja/) を用いる. 本稿ではジェスチャーとともに現れた発話一文すべてをジェスチャーと同時に起こったとして解析を行った. アノテートされた文に形態素解析を行ったのち, 各単語を原形に変換し,WordNet を用いて概念を抽出する. 一単語には複数の概念を持つものがあるが, 統計的な問題からそのうち一つを選ばなければならない. また上位の概念も複数存在するのでどの上位概念を取得するかを考える必要がある. 今回は, その方法として単語から 2 段階上の概念を取得し, もし複数の概念が存在していたならば, 過去にすでに出現したことのあるものを優先的に選んだ.

3.5　関係性の計算

3.3 章と 3.4 章で生成したデータとジェスチャー機能ラベルの関係を計算した. 次章はその結果を示す.

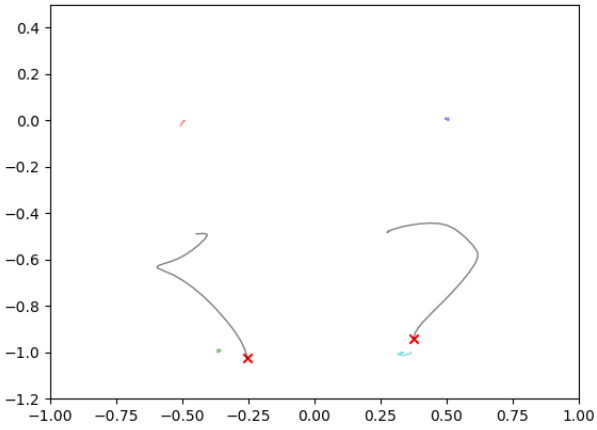

図 3: An example of projected gesture trajectories of both hands.
Cross points indicate start points of the gesture.
X-axis and y-axis are normalized to the shoulder and hip joints drawn by light dots.

4 結果

この章ではジェスチャー機能ラベル, ジェスチャーの動き, ジェスチャーと同時に出現する単語の関係に関する 3 調査結果を示す. データは実験を行ったグループのうちアノテーションが完了している 1 グループのデータを用いて計算した.

Fig 4 に示す通り, ジェスチャー機能のラベルと癖の動きの分布は, 映像的ジェスチャーが 57 回, 暗喩的ジェスチャーが 221 回, 拍子ジェスチャーが 504 回, 指示的ジェスチャーが 110 回, 癖が 146 回だった. なお, 拍子ジェスチャーはほ

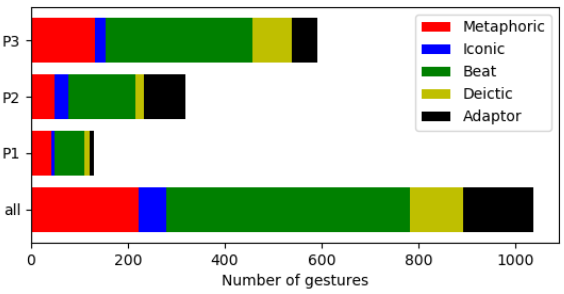

図 4: Number of occurrences of gesture meaning categories.

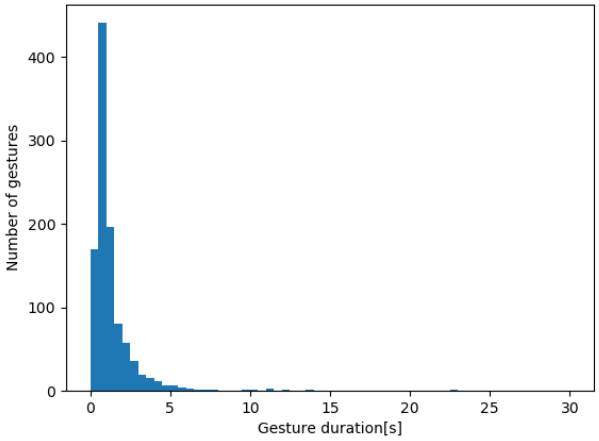

図 5: Histogram of gesture durations in stroke phases.

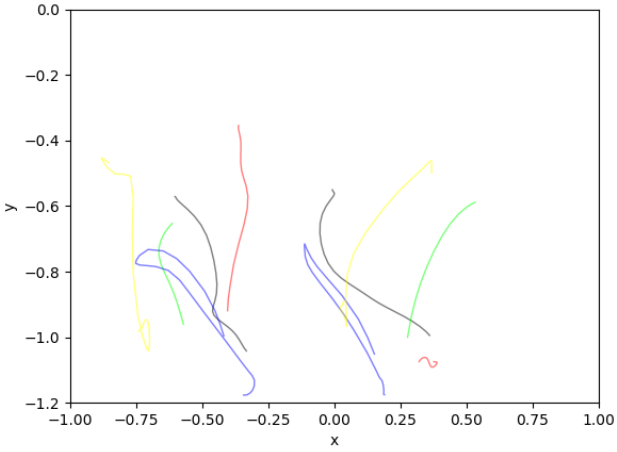

図 6: Example of a gesture cluster potentially expressing up-down movements.

かのジェスチャーと重複を許すが, 今回は, 重複していないものだけを拍子ジェスチャーとして計算した. また, 各ジェスチャーのストローク部分の継続時間は Fig 5 となり,0.5 秒から 1 秒の間が一番多かった.

4.1 ジェスチャー動作のクラスタリング

3.3 章で説明した方法を用いて,3 人分のデータを 100 クラスに分類した時の結果を示す.Fig 6,Fig 7 に示したものは 3 人分のデータを 100 クラスに分類した時のあるクラスタ内に含まれる一部のジェスチャーの軌跡の射影図である. 同じ色で示される 2 曲線が対となる左右の手の動きとなる.Fig 6 に示されるクラスタは, 腕の上下運動を表すと思われる. このクラスタは, 上または下への両手の動き 3 例と上または下への片手の動き 1 例, 両手を上にあげてから下に戻す動きを含んでいる.

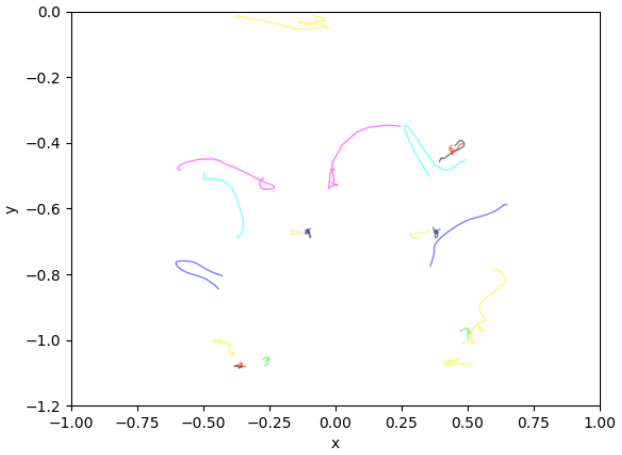

図 7: Example of an inefficient gesture cluster containing different types of trajectories.

表 1: Word concepts mostly appeared along with gestures.

出現回数	
概念	回数
交流	131
個人	124
鑑定	113
動静	102
原因物	77
考える	63
性状	57
属性	57
事	53
オーガナイゼーション	46

4.2 ジェスチャー機能ラベルと出現する単語の関係

Table 1 はジェスチャーとともに出現する概念数を,Table 3 に示したものは各ジェスチャー機能ラベルとそれと同時に出現する単語の概念の関係を表したものである. 左から出現した概念名, 出現回数, 出現した回数のうち, 各ジェスチャー機能ラベルとともに現れる確率を示す. 例えば, 指示カテゴリの中の聴覚コミュニケーションの概念は, 全会話中に 72 回出現し, その中の 36 回は指示的なジェスチャーとともに現れたということを表している. また,Table 2 で示されるのは, ジェスチャー機能のカテゴリごとの名詞の出現度を示す.

5 考察と今後の展望

Fig 4 より, ジェスチャーの機能の割合やジェスチャーが起きる頻度には, 同じ対話に参加していても個人差があることが分かる. ジェスチャー動作のクラスタリングについて,Fig 6 のようにある程度クラスタリングができているクラスタもあったが, できていないものも散見された. その原因として以下のことが考えられる. 一つ目はある点にとどまっているジェスチャーと動いている点では動いている点の影響が大きいことがあげられる. これは射影平面上でジェスチャーの軌跡が通る距離が変化するためでありその結果 Fig 6 内でみられるような, 片一方の手のジェスチャーは類似しているがもう片一方の手の動きは類似していないものが同一のクラスタになる現象が発生すると思われる. 次に, 往復動作と往のみの動作の区別がほとんどできないことも原因として考えられる. さらに, 今回の実験データでは指先の情報を使用しておらず, ジェスチャーが発生しているはずの区間でも軌跡では動いてないように見えるものが多数存在した. これらの改善のために, クラスタリングの手法の見直しや, 指の情報を追加するなど対策を今後行っていく.

表 2: Nouns mostly appeared in different gesture meaning categories.

映像隠喩		指示		エンブレム	
単語	回数	単語	回数	単語	回数
人	34	自分	13	人	9
何	23	私	9	一	6
感じ	9	人	8	二	5
みたい	8	相手	5	年間	3
一	7	一	5	三	3
異性	7	ここ	4	片手	2
友達	6	タイプ	4	感じ	1
それ	6	何	4	対	1
二	5	こと	3	五	1
の	5	みたい	3	日	1

次に, ジェスチャーの機能と単語の対応について,Table 2 では, 名詞に限定して単語の調査をし,Table 1 や Table 3 では,Wordnet を用いて概念を検索したものである. 前者では, 映像隠喩に感じやみたいといった比喩表現, 指示に人物, エンブレムに数字が多い系傾向にあることが分かるのに対し後者のほうは傾向がつかみにくい. これは Wordnet で上位概念に上りすぎたため一つの概念がカバーする範囲が広くなりすぎている可能性が考えられる. 以降は Wordnet をどこまでたどるか, どの品詞をたどるべきかなどを検討していく. また,Wordnet には一つの単語に複数の上位概念が候補として出てくることがあり, 前後の文脈などで正しいものを選択す機構も必要と思われる. これらの課題を解決しながら, 自然な動作が行えるよう動作生成へのモデルの検討も進めていく.

6 謝辞

この研究は JST, ERATO（グラント番号：JPMJER1401）の一環として行われたものです. ジェスチャーの分類やラベリングに協力いただいた三方瑠祐氏, 村瀬妙子氏, 奥野

表 3: Word concepts mostly appeared in different gesture meaning categories.

暗喩的			映像的			振動			指示		
概念	回数	確率	概念	回数	確率	概念	回数	確率	概念	回数	確率
行ない	21	0.428571	個人	124	0.169355	オーガナイゼーション	46	1	ピリオド	19	0.333333
遷移	40	0.425	動く	27	0.115385	抽象的実体	42	0.804878	聴覚コミュニケーション	45	0.272727
考え	12	0.416667	ピリオド	19	0.111111	時	25	0.708333	原因物	77	0.263158
通信	19	0.368421	原因物	77	0.105263	動静	102	0.67	考える	63	0.193548
属性	57	0.333333	交流	131	0.1	機器	12	0.666667	属性	57	0.157895
性質	24	0.333333	容態	32	0.096774	交流	131	0.653846	性状	57	0.142857
機器	12	0.333333	性状	57	0.089286	容態	32	0.645161	個人	124	0.137097
時	25	0.291667	事	53	0.078431	もの	36	0.638889	交流	131	0.130769
結付き	21	0.285714	抽象的実体	42	0.073171	事	53	0.627451	鑑定	113	0.125
内容	29	0.275862	属性	57	0.070175	動く	27	0.615385	事	53	0.117647

美紀氏に感謝する.

参考文献

[1] Hwang, S. J., et al. "Ada-Boost based Gesture Recognition using Time Interval Window." 2015

[2] C. Ishi, C. Liu, H. Ishiguro, N. Hagita. "Evaluation of formant-based lip motion generation in tele-operated humanoid robots," Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012), pp. 2377-2382, October, 2012.

[3] C.T. Ishi, C. Liu, H. Ishiguro, and N. Hagita."Head motion during dialogue speech and nod timing control in humanoid robots," Proc. of 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2010), pp. 293-300, 2010.

[4] C. Ishi, T. Funayama, T. Minato, and H. Ishiguro (2016). "Motion generation in android robots during laughing speech," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2016), pp. 3327-3332, Oct., 2016.

[5] C.T. Ishi, T. Minato, H. Ishiguro. (2017). "Motion analysis in vocalized surprise expressions and motion generation in android robots," IEEE Robotics and Automation Letters, Vol.2, No.3, 1748 - 1754, July 2017.

[6] 門野友城, 高瀬裕, and 中野有紀子. "隠喩的ジェスチャの分析とジェスチャ自動付与に向けた検討." 人工知能学会全国大会論文集 29 , 2015: 1-3.

[7] Kyoko Kanzaki, Francis Bond, Noriko Tomuro and Hitoshi Isahara, "Extraction of Attribute Concepts from Japanese Adjectives",.LREC-2008, Marrakech, 2008

[8] Kendon Adam, "Gesticulation and speech: two aspects of the process of utterance" ,The Relationship of Verbal and Nonverbal Communication,pp. 207-227, 1980

[9] 喜多　壮太郎, 身体とシステム　ジェスチャー　考えるからだ, 金子書房, 2002 .

[10] S. Kurima, C. Ishi, T. Minato, and H. Ishiguro. Online Speech-Driven Head Motion Generating System and Evaluation on a Tele-Operated Robot, IEEE International Symposium on Robot and Human Interactive Communication (ROMAN 2015), pp. 529-534, 2015.

[11] C. Liu, C. Ishi, H. Ishiguro, and N. Hagita. Generation of nodding, head tilting and gazing for human-robot speech interaction. International Journal of Humanoid Robotics (IJHR), vol. 10, no. 1, January 2013.

[12] Lücking, Andy, et al. "The Bielefeld speech and gesture alignment corpus (SaGA)." LREC 2010 workshop: Multimodal corpora?advances in capturing, coding and analyzing multimodality. 2010.

[13] McNeill, David. Hand and mind: What gestures reveal about thought. University of Chicago press, 1992.

[14] Shih-En Wei and Varun Ramakrishna and Takeo Kanade and Yaser Sheikh ,"Convolutional pose machines", CVPR, 2016