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-052-7 (12/3)

アンドロイドの発話に伴うジェスチャー生成システムのオンライン化の検討

Considerations of online processing for gesture generation accompanying android's utterance

三方瑠祐, 石井カルロス寿憲, 新谷太建, 石黒浩
Ryusuke MIKATA,Carlos T ISHI, Taiken SHINTANI, Hiroshi ISHIGURO
ATR（石黒特別研究所）
ATR (HIL)
mikata.ryusuke@irl.sys.es.osaka-u.ac.jp, carlos@atr.jp,
shintani.taiken@irl.sys.es.osaka-u.ac.jp, ishiguro@irl.sys.es.osaka-u.ac.jp

Abstract

日常生活における会話では，我々人間は言葉のやりとりだけではなく，視線や声の強弱，身振り手振りといったノンバーバル情報を用いてコミュニケーションを行っている．本研究では，この身振り手振り，いわゆるジェスチャーに注目し，ロボットへの実装を目標にしている．ここで扱うジェスチャーは, ハンドジェスチャーに限定し以降は, 単にジェスチャーとのみ表記する．ジェスチャーは，表現する対象物や動きによって，数種類のカテゴリーに分けることができる．人間同士の雑談対話のマルチモーダルデータを用いて, ジェスチャーに対するラベル付与および手の動きの抽出を行った．そして, ロボットのジェスチャー生成のため, 発話内容, ジェスチャーの機能, 動きの関係性を計算したベイジアンネットワークを作成した．作成したネットワークを用いて, アンドロイドの発話に伴うジェスチャー生成を行った. この生成方法を, 人間とアンドロイドの１体１の自由対話をする状況において, オンライン化への実装を提案し, より汎用性のある手法へとつなげていく．

1 はじめに

われわれは普段, 人と会話するときには無意識の内に身体全体を動かしている．聞き手は, 話者の声からの情報だけでなく, 手や胴体の動き, 顔の表情などからも情報を読み取り, 話者は聞き手を意識しつつ, 自分の発話や身体の動きを調節する．このように, 対話を行う状況においては, 様々な非言語的コミュニケーションが必然的に起こってくる．動きには意識的なものもあれば, 無意識的なものもあり, これらが混ざり合って会話が成り立っている．本研究では, このような非言語コミュニケーションの中のジェスチャーに焦点を置いている．ここで, ジェスチャーの定義として喜多 [15] の表現を用いて「何かを伝えようという意図のもとに起こる行為の一環としてある身体の動きが発現し, それが伝えるべき内容に関連のある情報を表す」とき, 身体の動きを“ジェスチャー”と呼ぶことにする．近い未来, アンドロイドなどのロボットが社会進出により, 人間とロボットのインタラクションする機会が増加すると考えられる．このインタラクションの中で, 言語情報だけでなく, 非言語情報を表現することによって, ロボットが相手に伝えたいことを理解しやすく, 正確に伝えることが可能となり, より自然な対話の実現につながる．われわれの研究室では, 発話に伴う口唇・頭部・表情・上半身（腰部）の動作生成に関する研究を数多く報告してきた [6][7][9][11]．本研究の目的は, ジェスチャーの中でもハンドジェスチャーに注目し, アンドロイドの発話に伴うジェスチャーを生成し, より人間らしい振る舞いを実現することである．対話中のジェスチャーにはすでに数多くの研究が行われており,McNeill はジェスチャーを図像, 隠喩などの機能ごとに分類する手法を提案している. 分類の詳細については,2 で述べる．この分類において, Bergmann[2] らは，図像ジェスチャーを，対象物を IDT(Imagistic Description Trees)[14] で表現し，ベイジアンネットワークで表現されたモデル（GNetIc）を用いて，CG エージェントへの動作生成を行った．中野ら [16] は隠喩的ジェスチャーを Wordnet を用いて，単語の上位概念を分析することで，CG エージェントへの動作生成を試みている．本研究では, 分類されたすべてのジェスチャーに対して生成を行っている. 以前より進めてきた, 人間同士の収録データを分析, ラベリングすることによってジェスチャーを生成する手法 [8] を, オンライン化システムを実装し, 人間とアンドロイドの対話におけるジェスチャー生成を行った．

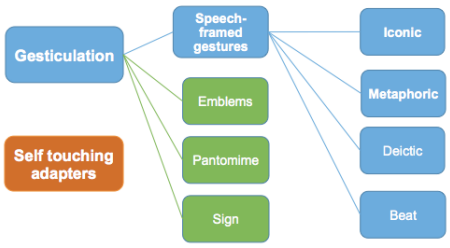

図 1: GestureCategory

2 ジェスチャーの分類

2.1 意味に関する分類

本研究では,McNeill[13], 喜多 [15] の研究より知見を得て, 発生したジェスチャーについて以下のように分類している. 本研究では, この分類をジェスチャーファンクションと呼ぶ.

1. 図像（Iconic）:
   具体的な物の形や大きさ, 状況や出来事を表現する動作

2. 隠喩（Metaphor）:
   図像とは異なり,”考え”や”記憶”などの実態がない単語のような, 抽象的な内容を空間して表現する動きである.

3. 指示（Deictic）:
   ある方向や場所を指し示すジェスチャーであり, 人差し指を伸ばした形が多く見られるが, 手のひらを使った形, 顎や目線を用いて行われる場合もある.

4. 拍子（Beat）:
   手の形や動きと意味の関係が社会的慣習ではなく, 他の要因によって固定される. 動きの特徴としては, 手を小刻みに振る動作を表し, 同期している発話が談話構造の観点から重要であることを示す.

2.2 フェーズに関する分類

また Kendon[10] は, ジェスチャーは発生してから終わるまでのフェーズを 4 種に分けることができることを提唱している. その 4 種のフェーズを以下に示す. ここでは, ジェスチャーが起きていない状態（座った状態であると仮定したとき, 手が膝の上にあるような状態を示す）をホームポジションと呼ぶ.

1. 準備（Preparation）:
   ホームポジションからジェスチャーをするために行われる移動の動き

2. ストローク（Stroke）:
   ジェスチャーの意味を表す, メインの動き

3. 終わり（Retraction）:
   ストロークが終了し, 手をホームポジションへと戻す動き

4. ホールド（Hold）:
   ストロークの前後, すなわち, 準備の後や終わりの前に見られる, 手を上げたまま動作を停止して固まっている動き

3 生成手法

この章では, 対話の収録データからどのようにジェスチャー生成を行うかの手法について説明する.

3.1 ジェスチャーファンクションのラベル付け

収録したデータは, 各被験者に対して,2.2 で述べた Kendon[10] のフェーズの分類方法に基づいた動作の区切りを行うことによりジェスチャーのラベル付けをする. ジェスチャーのフェーズの内, メインの部分となるストロークの部分に対して 2.1 で述べたジェスチャー機能に加え, 慣習的な動きを意味するエンブレム（図 1）の動きの 5 種類の分類を行った. ここで, ラベル付けを行う際, 拍子のジェスチャーは, 他の図像や隠喩に付随して見られる特徴があったので, 拍子のみ重複を許している.

3.2 動きのクラスタリング

3.1 の手法でラベル付けを行ったストロークのフェーズに対して, 動きのクラスタリングを行う.

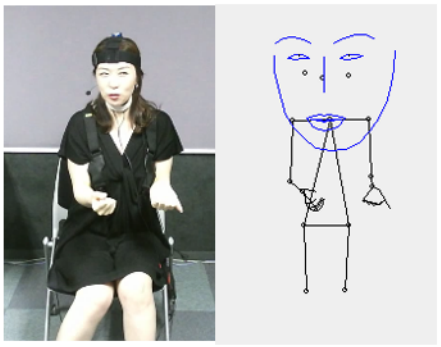

図 2: Skeleton extracted from Openpose (the face parameters are enlarged to allow better visualization)

カメラから取得した映像から,Openpose[1] を用いた骨格情報（図 2）を取得することで動作のデータを得る. また, 取得した Openpose の情報は 2 次元データなので, 人間の 2 次元のスケルトン情報を入力情報とし,3 次元のスケルトンデータを出力として得られる Martinez ら [12] が提案した手法を用いることにより 3 次元へと変換を行う.

クラスタリングに用いた情報は手首の動きの軌跡を用いる. この時系列データを分類する方法として, 本研究では k-means によるクラスタリングを行う.k-means に使用

するコスト関数として, 一般的にはユークリッド距離が使用されるが, 時系列データである軌跡に対してこれを用いた情報はジェスチャーの順序が考えられていないため有効ではない. そのため, 本研究では, データを非線形に伸縮しロバストな Dynamic Time Warping(DTW) を使用する. 図 3 にこの手法でクラスタリングを行った結果を示す. 図

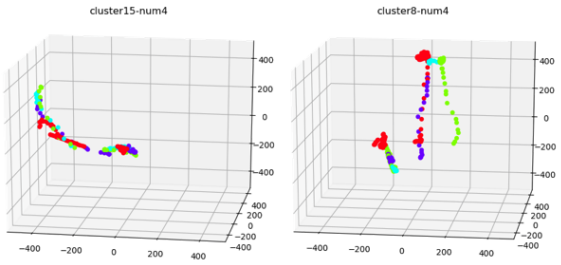

図 3: Result of wrist trajectory after clustering (Right:cluster number 15,Left:cluster number 8)

3 の cluster 15 に分類された動きは, 右手を右上へと動かしている動きと分類され,cluster 8 は左手を上へと上げる動作に分類されていることが分かる.

3.3 単語概念

ジェスチャーカテゴリや動作の特徴と収録対話での発話された単語のネットワークを作成するため, 先行研究に従って WordNet[3] を用いる.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,1 つの synset が 1 つの概念に対応している. ジェスチャーが起きた際, カテゴリーのラベル付けと同様にストロークの部分に対して, ジェスチャーの生起要因となった単語のラベル付けを人間の手によって行った. ラベル付けされた単語から WordNet を用いて, その単語に対する synset を取得することにより, ラベル付けされているジェスチャーカテゴリとの関係性を計算しネットワークを作成することができる.

3.4 全体の流れ

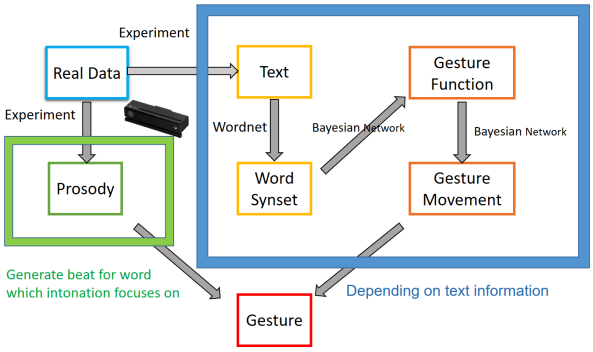

図 4: Overall processing of generation method

システムの流れの全体を図 4 示す. ロボットの発話テキスト, 発話時間を入力情報とし, 出力情報はロボットの関節軸に対応した指令値の時系列データである. 入力されたテキストに対して形態素解析を行い, 抽出した単語に対する synset を得る. 得られた synset からジェスチャー発生の有無を確率分布により計算する. ジェスチャーが発生した場合は, その synset に対応するジェスチャーカテゴリを求める. ジェスチャーカテゴリが決まれば, カテゴリに対するモーションクラスタを計算し, 動きの抽出を行う. ジェスチャーの発生の起因となった単語が発話される時間を求め, 取得した動きの 3 次元データ情報を入力し, アンドロイドの動作データを作成する.

4 アンドロイドへの実装

3.4 の処理で生成されているモーションデータは,3 次元の座標データなので, アンドロイドの各アクチュエーターの指令値への変換を行う必要がある. 本研究では, 実際行われている 3 者対話の収録データから, ある特定の人物の動きをアンドロイドにマッピングさせることでジェスチャーの再現を行っている. 収録データから, 各時間における人間の肩, 肘, 手首, 指の関節の 3 次元座標から, 各ジョイントの角度を計算している. 角度が求まれば, 可動範囲から指令値を計算することができる. 図 5 にアンドロイドのアクチュエーターの配置例を示す.

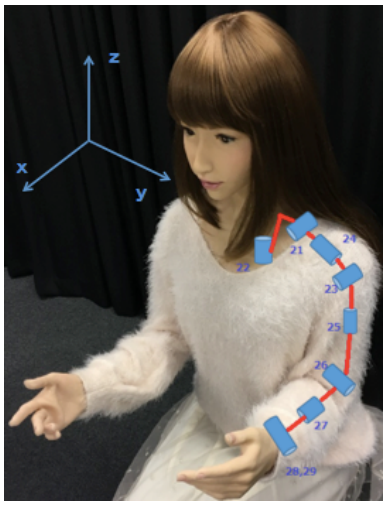

図 5: Arrangement and number of the left arm joint

4.1 指令値の計算

3 次元の座標は, 収録した人間の肩から腕の長さでスケーリングされている. よって, 腕の長さが異なる場合においても, 3 次元の値の情報は腕の伸ばし具合を示すことになるので, ジェスチャーの再現度も保たれる. 角度計算におけるアンドロイドの座標系は, アンドロイドの股間を原点とし, 真っ正面に手を伸ばす方向を x 軸の正の向き, 腰から頭へ向かう方向を z 軸の正の向きと設定する. この座

標系を基準とし，DH 法 [4][5] に基づいて各ジョイントに固有の座標系を設定する．隣接する各ジョイントの座標系間の変換行列を求めることで，各ジョイントの回転角度から 3 次元座標を，基準点（今回は胸の中心）に近いジョイントから順に求めていくことができる．このような，角度を与えてジョイントの座標を推定することは，順運動学と呼ばれている．一方，ジョイントの座標を与えて角度を求めることは逆運動学として知られている．本研究では，逆運動学と内積，外積を用いることで，3 点（肩，肘，手首）の 3 次元座標から角度計算し, 指令値を求めた．

ジェスチャーの再現において, 手先の動きが重要であると考え, 手先のずれを最小限に抑えるため, 肘の位置を推定することにより, 角度計算を行った．入力情報の手首の位置を中心とし，アンドロイドの前腕の長さを半径とする球と入力情報の肩の位置を中心とし，アンドロイドの上腕の長さを半径とする球の交円を計算する．この交円の点が，推定する肘の位置の候補となる．候補の点を 1 つにしぼるために，入力情報として与えられる肩，肘，手首の 3 点から生成される平面を用いた．この平面と交円の交点を求めることで，肘の位置が 2 つに絞られる．絞られた 2 つの点の内，入力情報の肘の点に近い点を肘の推定点とした．

4.2　ストロークの判別

また, ストロークの動きのみを扱っているので,3.4 の処理の段階では, 準備と終わりの動きが生成されていない．加えて, ジェスチャーが発生していても, 片方の手はストロークだが, もう片方の手はホームポジションの場合がある．動きのない側の手は, ホームポジションにあるはずだが, アンドロイドのホームポジションと人間のホームポジションには, ずれがあるため, 手がストロークであるのか判別を行い, ジェスチャーが起きていない場合は, ホームポジションに手を移動させる必要がある．この判別を行うために, 手首の高さに対してある閾値を与え, どちらの手が動いているのかの判別を行っている．閾値を超えた時点をストロークの開始と判断し, ホームポジションからその時点における指令値の値を用いて準備の動きを, 逆に閾値を超えていたが下回った場合は, ストロークの終了と判断し終わりの動きを作成する．収録データの解析した結果に基づいて, 準備の時間は 0.8 秒, 終わりの時間は 1.2 秒の固定値に設定している．

5　オンライン化

本研究では,3 で述べた生成法を用いて, アンドロイドと人間の 1 対 1 の対面対話の状況を考え, アンドロイドの発話に対してオンラインでのジェスチャー生成を行う．アンドロイドと人間の対話は自由に行うものとし, 発話衝突が起きた際は, 相手の発話が終わるまで, アンドロイド側が相手の発話終了を待つように設計している．このシステムでは, このようなターンテイキングを行うため, 単純にモーションを再生するのみではなく, 発話衝突が発生し, ジェスチャーをしていた場合は手を下ろすように動作生成をするなど, モーション再生途中で動作の変更をすることが必要になる．また, 発話が終了するタイミングでモーションの再生も同時終了するとは限らない．これは, 発話終了時点における手の位置がホームポジションに無い状態, つまりジェスチャーの最中である可能性が存在し, そのような場合には, 終わりのモーションの時間分を追加する必要が生じてくるため発生する．発話と同時に終わらないことにより, ジェスチャー再生中に次の発話が起こり, モーションの衝突が起こることが予測される．この対処として, 前モーションと現モーションの比較を行い補完などの新たな処理を行う必要がある．

これらに対処するために, 生成する動作データは, アンドロイドの各ジョイントのアクチュエーターの指令値以外に, 各フレームに対して右手と左手それぞれに, ジェスチャーのどのフェーズに当たるのか, つまり, 準備, ストロークなどの情報を同時に保存しておくことで実装を行った．

5.1　発話衝突時の処理

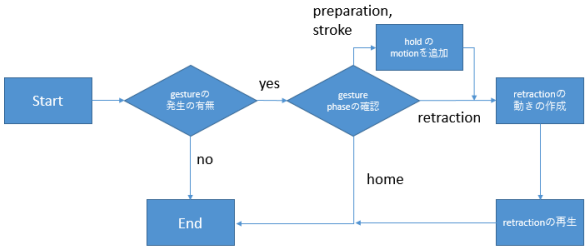

図 6: Processing at speech collision

図 6 は, 発話衝突が起きた際の処理のフローを表している．発話衝突が発生すると, 音声および動作を中断させるためのキャンセル信号が受信される, キャンセル信号を受信した際は, まず動作再生中の有無を確認し, 送信中の場合は, 信号を受信した時点でのフレーム数を用いて, そのジェスチャーのどのフェーズで発話衝突が起こったかの確認を行う．ジェスチャーのフェーズがストロークや準備の場合はホールドの時間を追加した後に終わりのフェーズを, 終わりの場合はそのままモーションを続けるように設計した．終わりのモーションの作成方法は, ジョイントの指令値を扱って作成しており, ホームポジション時のジョイントの指令値とジェスチャー途中の指令値をシグモイド関数を用いて補完することで作成している．ホームポジション時のジョイントの指令値はあらかじめ用意しているものとする．

5.2 動作再生時の処理

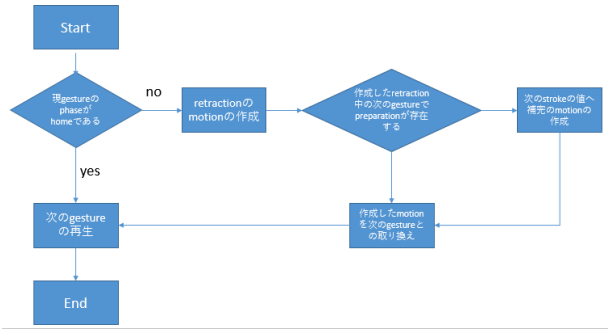

図 7: Processing at the start of gesture

図 7 は再生の信号を受信した処理のフローを表している. 再生信号を受信した時点で再生中のモーションがあるか確認を行い, 再生中の場合は, 発話衝突時と同様に, 動作開始の命令を受信した時点でのジェスチャーのフェーズの確認を行う. フェーズがホーム以外であるならば, まず, その時点での手の位置から, 終わりのモーションの作成を行う. 作成した終わりのモーション時間（フレーム数）分を次に再生をするジェスチャーの開始から変更を行うことで, モーションをつなぎ合わせる. この変更を行う際に, 次のジェスチャーの変更部分のフェーズが準備であった場合は, 終わりのモーションから次のストロークの位置を求めて, 新たに準備のモーションの作成を行うことで補完している.

6 結果

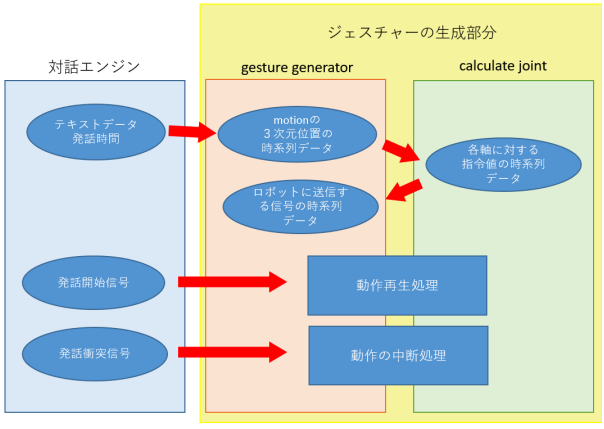

図 8: Online System

図 8 に, 本研究で実装したシステムにおける処理の簡略図を示す. テキスト情報, 発話情報はソケット通信で受け取り,gesture generator でジェスチャーの発生の計算, モーションの 3 次元データを作成する. そして, 同様にソケット通信で 3 次元データの送信を行い, アンドロイドの各アクチュエータの指令値のモーションデータを作成する. この処理で, 準備, 終わりのフェーズの作成を行っている. 作成した指令値のモーションデータは, アンドロイドへ送る byte 型のデータへと変換し, gesture generator にソケット通信で送信している.

表 1 に, 発話テキスト, 発話時間, 作成時間, ジェスチャーの有無についての結果示す. 計算に用いた機器のスペックは CPU:Intel(R)Core(TM) i7-5500U, クロック数: 2.40Ghz である. 今回のシステムでは, テキストに対して音声合成を用いてアンドロイドの発話音声を作成しており, その作成された音声情報をもとに発話時間を扱っている. ジェスチャーの有無について, その発話に対して何もジェスチャーが発生しなければ,None と表記し, 図像, 隠喩, 指示に対しては, それぞれ Icn,Mtp,Dct と表記する. 作成時間は, テキスト情報を受け取った時間を生成開始時間とし, 図 8 の gesture generator がアンドロイドへの指令値へと計算し終え, すべての時系列データを byte 型として格納できた時点を完了としている.

今回オンライン化に実装した部分は, ジェスチャーファンクションにおける図像, 隠喩, 指示の部分の生成を行った. ジェスチャーが起きない場合は, 約 2.5 秒の発話について 0.2 秒程度の生成時間が必要であり, 発話が長いほど生成する時間も長くなる結果となった.

システムの設計上, 発話情報を受け取ってすぐに再生開始の信号がくるようにはなっていないことに加え, 腕のモーションが多少遅れたとしても, 研究機関内の評価では大きな違和感を感じることは無かった.

7 考察・今後の展望

表 1 より, ジェスチャー生成処理において, 同じ発話内容かつ発話時間であっても, ジェスチャーが起きた場合はさらに時間を要する結果となっているが, これは, ジェスチャーが発生しない場合は,3 次元データは与えられないので指令値の計算を行わず, ホームポジションの指令値を結果として処理するためだと考えられる. 今回は, 発話内容を 1 文受け取り, 形態素解析をすることによって単語の抽出を行ったが, 音声合成の際, 発話テキストを形態素解析する処理が含まれると考えられるので, その解析結果を入力情報として受け取ることで, より生成時間を短くできるのでは無いかと予測できる. また, 発話量が多い場合, つまり発話時間が長ければ長いほど計算時間は長くなってしまう上に, 発話時間に相当するフレーム数のモーションデータを完全に作成し終えるまで, 再生を許していないので, 1 つのテキストにおいても, 作成し終えたモーションから随時再生可能にし, 作成と再生を平行に行うことが理想的である. 生成手法について, 現在の手法では, ジェスチャーファンクションに依存して動きの選択を行っているが, これでは, 単語そのものが持つ特徴をうまく扱えていない. また, どのような対話状況においても, 単語に対するジェ

表 1: 生成時間

NO.	テキスト	発話時間 (msec)	ジェスチャーの有無	生成時間 (msec)
1	はい	704	None	78
2	そうですか	1124	None	109
3	お名前はなんていうのかしら	2414	None	178
4	どちらからいらしたの	1654	None	142
5	どちらからいらしたの	1654	Mtp	208
6	長野と長崎って似ていますよ	2569	None	201
7	果物屋ですか	1394	None	125
8	果物屋ですか	1394	Icn	202
9	あなた何てことをおっしゃるの	2039	Dct	224

スチャーの発生率は同じなどの問題点があるので, これらを考慮したネットワークの作成が必要である.

現時点では, オンライン化のシステムの実装を行った. 今後, このシステムに対する評価実験を行う予定である. また, 今回は紹介していないが, オフラインでは, 拍子の生成を音声の韻律情報から作成を行っている [8]. この技術も, オンライン化することで, より人間らしい動作生成を進めていく.

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村瀬妙子氏, 馬場由美子氏に感謝する.

参考文献

[1] *https://github.com/CMU-Perceptual-Computing-Lab/openpose.*

[2] Kirsten Bergmann and Stefan Kopp. Gnetic — using bayesian decision networks for conference on intelligent virtual agents. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '9, pp. 76–89. Springer-Verlag, 2009.

[3] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. Japanese semcor: A sense-tagged corpus of japanese. *Proceedings of the 6th Global Wordnet Conference(GWC 2012)*, pp. 56–63, 2012.

[4] J. Denavit and R. S. Hartenberg. A kinematic notation for lower-pair mechanisms based on matrices. *Trans. ASME, J. Appl. Mech.*, Vol. 22, No. 2, pp. 215 – 221, 1965.

[5] R. S. Hartenberg and J. Denavit. *Kinematic synthesis of linkages*. McGraw-Hill series in mechanical engineering. McGraw-Hill, 1964.

[6] Carlos T Ishi, Chaoran Liu, Hiroshi Ishiguro, and Norihiro Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, pp. 293–300. IEEE Press, 2010.

[7] Carlos T Ishi, Chaoran Liu, Hiroshi Ishiguro, and Norihiro Hagita. Evaluation of formant-based lip motion generation in tele-operated humanoid robots. In *Intelligent Robots and Systemsv(IROS), 2012 IEEE/RSJ International Conference on*, pp. 2377–2382. IEEE, 2012.

[8] Carlos T Ishi, Daichi Machiyashiki, Ryusuke Mikata, and Hiroshi Ishiguro. A speech driven hand gesture generation method and evaluation in android robots. *IEEE Robotics and Automation Letters*, Vol. 3, No. 4, pp. 3757–3764, 2018.

[9] Carlos T Ishi, Takashi Minato, and Hiroshi Ishiguro. Motion analysis in vocalized surprise expression and motion generation in android robots. *IEEE Robotics and Automation Letters*, Vol. 2, No. 3, 2017.

[10] Adam Kendon. *Gesticulation and speech : Two aspects of the process of utterance in M.* The Relationship of Verbal and Nonverbal Communication, 1980.

[11] Chaoran Liu, Carlos T Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Generation of nodding, head tilting and gazing for human –robot speech interac-

tion. *International Journal of Humanoid Robotics*, Vol. 10, No. 01, p. 1350009, 2013.

[12] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. *International Conference on Computer Vision*, Vol. 1, No. 2, p. 5, 2017.

[13] D. McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, 1992.

[14] Timo Sowa and Ipke Wachsmuth. A model for the representation and processing of shape in coverbal iconic gestures, 2005.

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

[16] 門野友城, 高瀬祐, 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, Vol. 29, pp. 1–3, 2015.