発話言語に基づく身体モーションの自動生成

石井 亮[1] 片山 太一[1] 東中 竜一郎[1] 富田 準二[1]

概要：人間のコミュニケーションにおいて，身体モーションは，発話言語に加えて感情や意図を伝達する重要な機能を持つ．そのため，擬人化エージェントやヒューマノイドロボットを用いた対話システムにおいて，発話に応じて適切な身体モーションを表出し，ユーザとの円滑なコミュニケーションを行うことが望まれている．本研究では，発話言語から得られる多様な言語解析情報を利用して，人間と同様な適切なタイミングで発話に伴う全身の身体モーションを自動生成する技術を提案する．具体的に，発話言語に含まれる単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為といった多様な言語情報と人間のモーションとの共起関係に着目し，これらの発話言語情報を用いて，頷き，頭部姿勢，表情，ハンドジェスチャ，上半身の姿勢を自動生成するモデルを構築した．最初に，2 者対話を収録し，発話および頭部運動，表情，ハンドジェスチャ，身体姿勢情報を含むマルチモーダルコーパスを構築する．次に，構築したコーパスデータを用いて，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を入力として，文節単位ごとにモーションを生成するモデルを，機械学習を用いて構築する．その結果，本研究で用いた多様な言語情報がモーション生成に有用な情報であることが示す．次に，本技術を広く利用できるための試みとして，構築した生成モデルを用いてモーションを容易に生成可能な API を構築し，UNITY 上の CG キャラクタを発話言語のみから自動制御可能なデモシステムを構築した．本システムでは，任意の発話言語を入力すると，音声合成器およびモーション生成 API から合成音とモーション情報を取得し，UNITY 上の CG キャラクの発声およびモーション付きアニメーションを生成する．このデモシステムを用いて，ユーザ主観評価実験を実施した結果，本技術により生成されたモーションを CG キャラクタに付与することで，動作の自然さ，発話と動作の一致度，エージェントへの好感，人間らしさなどの印象が向上することが確認された．

1. 序論

人間のコミュニケーションにおいて，身体モーションは，発話言語に加えて感情や意図を伝達する重要な機能を持つ [2]．そのため，擬人化エージェントやヒューマノイドロボットを用いた対話システムにおいて，発話に応じて適切な身体モーションを表出し，ユーザとの円滑なコミュニケーションを行うことが望まれている．近年，コミュニケーションロボットや CG エージェントが，産業界でも注目され，コミュニケーション相手や窓口案内業務，Q&A サービスでの利用など，多様なサービスに適用されている．実サービスを手掛ける上での大きな課題の一つとして，コミュニケーションロボットや CG エージェントの身体モーションを作成するために，発話毎に人手でモーションシナリオを作成するため膨大な時間がかかることが挙げられる．また，そのモーションをどのようなものにするか，その表出のタイミングについても，作成者の主観により手さぐりに行っている状況であり，必ずしも適切なモーションが作成されているとは言い難い．今後，対話技術が進歩するにしたがって，多様な発話を自動的に応答可能なシステムが実現されれば，そもそも，全ての発話に対して，人手で適切なモーションを付与することは非現実的である．

このような課題に対して，我々は，ヒューマノイドロボットや擬人化エージェントのモーションを，発話内容に基づいて，人間と同様に適切なタイミングで自動生成することに取り組む．このような技術が実現されれば，発話言語のみから自動でモーションが生成されるため，これまで必要であった作成コストは大幅に削減される，あるいは不要になるものと考える．

本研究では，入力として発話言語から得られる多様な言語解析情報を利用して，頷き，頭部姿勢，表情，ハンドジェスチャ，上半身の姿勢といった全身のモーションを，文節ごとに出力する手法を提案する．言語解析情報として，発話言語に含まれる単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を利用する．このような多様な言語情報とモーションの関連性については，これまで検討がなされておらず初めての試みである．

[1] 日本電信電話株式会社 NTTメディアインテリジェンス研究所

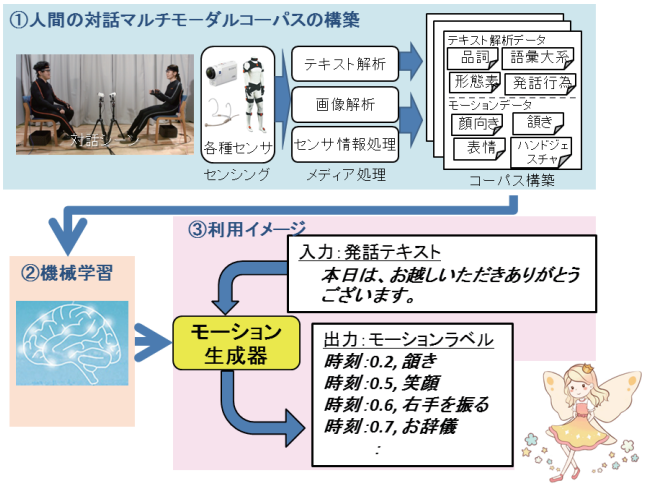

図 1　本研究の流れ

発話言語情報を利用してモーション生成を行う利点として，そもそも，発話言語と身体モーションは脳内で同時に生成されて出力されており，つながりが深いことが知られている．すなわち，このような共起関係を基に，発話言語情報はモーション生成のための情報として有効であると考えられる．

本研究では，最初に，2 者対話を対象にした，発話および頭部運動，表情，ハンドジェスチャ，身体姿勢情報を含むマルチモーダルコーパスの構築した (図 1 の ①)．次に，構築したコーパスデータを用いて，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を入力として，文節単位ごとにモーションを生成するモデルを，機械学習を用いて構築する (図 1 の ②)．その結果，本研究で用いた多様な言語情報がモーション生成に有用な情報であることが示唆された．次に，本技術を広く利用できるための試みとして，構築した生成モデルを用いてモーションを容易に生成可能な API を構築し，UNITY 上の CG キャラクタを発話言語のみから自動制御可能なデモシステムを構築した (図 1 の ③)．本システムでは，任意の発話言語を入力すると，音声合成器およびモーション生成 API から合成音とモーション情報を取得し，UNITY の発声およびモーション付きアニメーションを生成する．このデモシステムを用いて，ユーザ主観評価実験を実施した結果，本技術により擬人化エージェントのモーションの自然さや好感度などの主観評価値が向上することが確認された．

2.　関連研究

人間のコミュニケーションにおいて，頷き，頭部姿勢，表情，ハンドジェスチャ，上半身の姿勢といった身体モーション（非言語行動）は，発話言語に加えて感情や意図を伝達する重要な機能を持つことが知られている [2], [18], [19]．よって，擬人化エージェントやヒューマノイドロボットに適切な身体モーションを付与することで，見た目の自然さの向上だけでなく，会話を促進することが示されている．例えば，発話に付随するモーションは，発話の説得力を強化し，対話相手が発話の内容を理解しやすくする効果がある [13]．このような背景から，特に発話音声情報を用いて，発話中のモーションを生成する試みがなされている．音声情報として，音圧や韻律の特徴が多く利用されている [1], [3], [5], [8], [9], [12], [21], [22]．しかし，発話の音声情報から精度良く身体モーションを生成することは困難であった．特に，日本語では，音声特徴と頷きの共起関係は弱いことが知られている [7], [22]．そのため，音声情報以外の情報も利用し，より適切なタイミングで身体モーションを生成可能な手法が構築されれば，対話システムとユーザ間のより円滑なコミュニケーションが実現されると考えられる．

一方，言語情報によるモーション生成の従来研究では，主に単語情報を用いた，頷きの有無や，限定的なハンドジェスチャ [7], [8], [10] といったごく一部のモーションの生成について取り組まれていた．本研究は，多様な言語情報を用いて，より網羅的な全身のモーションの生成に取り組む，これまでに行われていない初めての試みである．

また，言語情報を利用する利点についていくつかの利点が挙げられる．音声を利用した対話機能を有する対話エージェントにおいて，音声情報を用いたモーション生成を適用する場合，対話システムが発話を行うテキスト情報を対話技術によって取得した後，音声合成を行って，合成音から音響特徴量を抽出するなどの処理が必要となり，処理時間が増大する．これにより，音声情報を用いたモーション生成では，ユーザへのレスポンス時間の遅れが発生してしまうことが問題として挙げられる．一方で，言語情報を利用した際は，そのような処理は不要であり，リアルタイムコミュニケーションに重要であるレスポンス性の担保にも有効なアプローチであると考えられる．また，そもそも音声合成自体も，入力は発話言語であるため，発話言語のみを用いたものと生成精度は変わらないか，むしろ低下する可能性も考えられる．その他，言語情報のみを用いることで，音声を用いないテキストチャットにも適用できるなど，技術がより広く適用されることが期待される．

3.　コーパス

2 者対話を対象に，発話言語およびそれに伴う身体のモーションデータを含む，言語・非言語マルチモーダルコーパスの構築を行った．2 者対話の参加者は，20–50 代の日本人男性・女性であり初対面であった．参加者は計 24 人（12 ペア）であった．参加者は図??のように，互いに向き合って着座した．対話内容は，発話に伴う頷きやハンドジェスチャなどの多様なモーションに関するデータを多く収集するために，アニメーションの説明課題を採用した．参加者はそれぞれ，異なる内容のアニメーション（Tom & Jerry）

生成対象	クラス数	クラス
頷きの回数	6	0, 1, 2, 3, 4, more than 5
頷きの深さ	4	micro, small, medium, large
頭部姿勢 yaw	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
頭部姿勢 roll	9	front, right-micro, right-small, right-medium, right-large, left-micro, left-small, left-medium, left-large
頭部姿勢 pitch	7	front, up-micro, up-small, up-medium, up-large, up-micro, up-small, up-medium, up-large
表情	8	happiness, sadness, surprise, fear, anger, disgust, contempt, normal
ハンドジェスチャ	9	none, iconic, metaphoric, beat, deictic, feedback, compellation, hesitate, others
上半身の姿勢	7	center, forward-small, forward-medium, forward-large, forward-small, forward-medium, forward-large

表 1 モーション生成のパラメータ一覧

図 2 2 者対話の様子

を視聴した後，対話相手にアニメーション内容を説明した．10 分間の対話セッションにおいて，1 人の参加者が対話相手にアニメーションの内容を詳細に説明した．対話相手は，説明者に自由に質問をし，自由に会話を行うことを許可した．発話の収録のために各被験者の胸につけられた指向性ピンマイクを用いた．対話状況の全体的な外観や参加者の様子の収録として，ビデオカメラを用いた．映像は 30Hz で収録された．撮影されたビデオの一例を図 2 に示す．下記に，取得した言語・非言語データを示す．

- 発話： 人手で音声情報から発声言語の書き起こしを行った後，発話内容を確認して文を分割した．さらに，係り受け解析エンジン [11] を利用して，各文を文節に分割した．分割された文節数は 11877 であった．
- 顔向き： 参加者の正面に設置されたビデオカメラから撮影された参加者の正面映像に対して，顔画像処理ツールである OpenFace[17] を利用して，3 次元顔向き情報である yaw，roll，pitch の角度をそれぞれ取得した．それぞれの角度が，10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．
- 表情： 顔向きと同様に，OpenFace を利用して，AU (Action Uncit) の強度を抽出し，その強度の組み合わせを用いて，7 種類の表情の種別（喜び，怒り，嫌悪，悲しみ，恐怖，驚き，平常）から一つを抽出した．算出方法は具体的に，各表情に関連のある AU（喜びは AU6，AU12，怒りは AU4，AU7，AU24，嫌悪は AU4，AU7，AU25，悲しみは AU1，AU4，AU7，AU15，AU17，恐怖は AU1，AU2，AU5，AU26，驚きは AU1，AU2, AU20, AU43，平常は強度を算出しない）の強度の平均値を算出し，これを表情 f の強度値 x_f とした．さらに，対話コーパス全体から得られる表情の強度値を用いて，平均値 μ_f，標準偏差 σ_f を表情毎に算出し，これを用いて Z スコア Z_f を算出し，x_f を正規化した．各表情の Z_f の中で最大であった表情クラスを現在の表情として選択した．なお，全ての表情の Z スコア Z_f が 0.3 未満であった際には，“平常”クラスを割り当てた．
- 頷き： 人手で映像から頷きが発生した区間をラベリングした．連続的に発生した頷きは 1 つの頷きイベントとして扱った．また，回数を 1 回から 5 回（以上）の 5 段階に分類した．また，頷きの深さについて，OpenFace を利用して，頷きの際の頭部姿勢 pitch の開始と，最も深く回転した箇所の角度の差を算出した．その角度が，10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．
- ハンドジェスチャ： ハンドジェスチャが行われている区間を人手アノテーションをおこなった．なお，ジェスチャの一連の動作は，下記の 4 つの状態に分類した．
  - Prep：ホームポジションからジェスチャをするために手を上げる
  - Hold：空中に手をあげたままの状態 (ジェスチャ開始までの待機時間)
  - Stroke：ジェスチャを実施
  - Return：手をホームポジションに戻す

  ただし，本研究では，取扱いの簡便さから，Prep から Return までの一連の動作を一つのジェスチャイベントとして扱った．さらに，ハンドジェスチャの種類

生成対象	チャンスレベル	提案手法
頷きの回数	0.226	0.428
頷きの深さ	0.304	0.475
顔回転 yaw	0.232	0.329
顔回転 roll	0.297	0.397
顔回転 pitch	0.261	0.378
表情	0.175	0.290
ハンドジェスチャ	0.156	0.303
上半身の姿勢	0.183	0.311

表 2 チャンスレベルと提案手法による推定精度（F 値）

を，McNeil のハンドジェスチャの分類 [14] を基に，下記の 8 種類に分類した．

- Iconic： 情景描写，動作を表現するために使われるジェスチャ．
- Metaphoric： Iconic と同様，絵画的，図形的なジェスチャであるが，指示される内容は抽象的な事柄，概念となる．例えば，時間の流れなど．
- Beat： 発話の調子を整えたり，発言を強調．手を振動させたり，発話に合わせて手を振る．
- Deictic： 指差しなど，直接的に方向や，場所，事物を指し示すジェスチャ．
- Feedback： 他人の発話に同調・同意・呼応して出るジェスチャ．他人の前の発話・ジェスチャに対して，呼応して発言したときに付随するジェスチャ．また，相手のジェスチャを模倣して同じ形のジェスチャを行ったもの．
- Compellation： 相手に呼びかけを行うジェスチャ．
- Hesitate： 言いよどみ時に出るジェスチャ．
- Others： 判断に迷うが，何か意味がありそうなジェスチャ．

- 上半身の姿勢： 本対話シーンでは，参加者は着座しており着座した位置の大きな変化は無かった．そのため，頭部の 3 次元位置を基に，上半身の前後の姿勢を抽出した．具体的に，OpenFace を利用して得られた頭部位置の前後方向の座標位置と中心位置の位置の差を取得した．その位置情報から，上半身の姿勢変化の角度を算出し，それが 10 度以下の時に micro，20 度以下の時に small，30 度以下の時を medium，45 度以上の時を large として分類した．

得られたコーパスデータのパラメータの一覧を表 1 に示す．また，人手アノテーションには ELAN [20] を使用し，上記のすべてのデータを 30Hz の時間分解能で統合した．

4. モーション生成器の構築

構築したコーパスデータを用いて，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を入力として，表 1 に示した 8 項目のモーションにおいてそれぞれ一つのモーションクラスを文節ごとに生成するモデルを，決定木アルゴリズム C4.5 を用いて構築した．すなわち，文節毎に，8 つのモーションラベルが生成される．具体的に，使用した言語特徴量は下記の通りである．

図 3 被験者実験で利用した CG キャラクタ

- 文字数：文節内の文字数
- 位置：文頭，文末からの文節の位置
- 単語：形態素解析ツール Jtag [4] により抽出された文節内の単語情報（Bag-of-words）
- 品詞：Jtag により抽出された文節内の単語の品詞情報
- シソーラス：日本語語彙大系に基づく文節内の単語のシソーラス情報
- 発話行為：単語 n-gram およびシソーラス情報を用いた発話行為推定手法 [6], [15] により文ごとに抽出された発話行為（33 種類）

24 人の参加者のデータの内，23 人のデータを学習に用いて 1 人のデータで評価を行う 24 交差検定法により評価を行った．これにより，他者のデータのみから，どの程度，実際の人間のモーションを生成できるかを評価した．性能評価結果である F 値の平均値を表 2 示す．なお，チャンスレベルは最も，正解数の多いクラスを全て出力した際の性能を示している．表 2 を見ると，全ての生成対象で，チャンスレベルよりも大幅に精度が向上した（対応のある t 検定の結果：$p < .05$）．この結果から，発話言語から得られる，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を用いた本提案手法は，表 1 に示したような全身のモーション生成に有効であることが示唆された．

5. システム構築と評価実験

提案手法を容易に利用できるツールとして，発話言語を送信すると，発話言語をテキスト解析および生成モデルを利用して，文節ごとにどのようなモーションを行うかのスケジュールを返答する Rest 型の API を構築した．提案するモーション生成手法が，実際のＣＧアニメーションが発話する際のモーション生成に有効であるかを評価するために，CG キャラクタの身体アニメーションを制御して，その印象を評価する被験者実験を実施した．

実装方法として，モーション生成 API を利用して，各文節でのモーション情報を取得すると共に，DNN を用いた

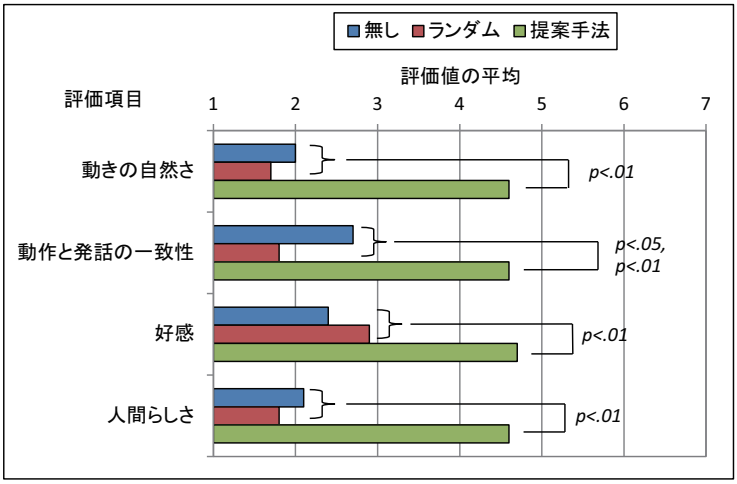

図 4　主観評価の結果

音声合成エンジン（FutureVoice）[16] によって，発話テキストから合成音を作成した．また，UNITY 上で図 4 に示したような CG キャラクタを作成し，表 1 のリストにあるモーションリストに対応するーションを作成した．UNITY 上では，発話音声の再生とモーションスケジュール情報を基に，音声の発生に合わせて，身体モーションのアニメーションが生成される．なおこのとき，表 1 に示した 8 つの生成対象は独立して動作が可能であり，頭部動作についても，全てのパラメータが混合されて生成される．また，リップシンクについても音声に合わせた簡易なリップモーションを生成できるように実装した．

実験条件として，発話に合わせてモーションを行わない条件（無し条件），ランダムに動作が行われる条件（ランダム条件），本提案手法によるモーション生成条件（提案手法条件）の 3 つの条件を設定した．各条件において，キャラクタの自己紹介を行う内容の 5 発話を用意し，それぞれの条件で同じ発話を行わせて，その様子に対する主観評価を実施した．各条件で CG キャラクタの挙動を観察した後，リッカート法による 7 段階の主観評価を実施した．評価項目は，“動きの自然さ”，“発話と動作の一致性”，“好感”，“人間らしさ”を問う 4 項目であった．被験者は 10 名の男女であった．各条件は，順序効果を相殺するため，被験者ごとにランダムに提示された．

主観評価の平均値を図 4 に示す．各評価項目内で，一元配置の分散分析を行って，実験条件の要因の効果が有意であるかを検証した結果，いずれも有意差が認められた (自然さ：$F(2,27) = 37.12, p < .01$，動作と発話の一致性：$F(2,27) = 18.34, p < .01$，好感：$F(2,27) = 7.80, p < .01$，人間らしさ：$F(2,27) = 33.76, p < .01$)．さらに，提案手法条件が 2 つの対照条件に対して有意差があるかを対応のある t 検定によって検証した結果，いずれも有意差が認められた．（具体的な有意差は図 3 を参照されたい．） よって，この結果から，本提案手法によって，CG キャラクタの動きの自然さ，動作と発話の一致性，キャラクタへの好感，人間らしさの印象が向上することが示唆された．

6. 議論

提案した発話言語解析情報を用いたモーション生成器の精度は，全体的に 0.290〜0.475 であり，ある程度の精度が得られたものの必ずしも高いものではなかった．これについては，そもそも人間のモーションが必ずしも，特定の発話で発生しなければならないという性質のものではなく，すなわち，そもそもがナイーブに発生されるものである．そのため，人間の実際のモーションを完全に再現することは難しく，またそれを行うことは必須ではない．どの程度，人間のモーションを精度良く再現できれば，ヒューマノイドロボットや擬人化エージェントのモーション生成に十分であるかを別途検討することも興味深い課題である．また，それが十分に検証されなかったとしても，CG キャラクタを用いた評価実験において，提案手法によるモーション生成によって，CG キャラクタの動きの自然さ，動作と発話の一致性，キャラクタへの好感，人間らしさなどの主観項目で評価が向上したことから，本手法による身体モーションの自動生成の有効性は示されたものと考える．

7. まとめ

発話言語に含まれる単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為といった多様な言語情報と人間の非言語モーションとの共起関係に着目し，これらの発話言語情報を用いて，頷き，頭部姿勢，表情，ハンドジェスチャ，上半身の姿勢を自動生成するモデルを構築した．最初に，2 者対話を収録し，発話および頭部運動，表情，ハンドジェスチャ，身体姿勢情報を含むマルチモーダルコーパスを構築した．次に，構築したコーパスデータを用いて，単語，その品詞およびシソーラス，単語位置，発話言語全体の発話行為を入力として，文節単位ごとに

モーションを生成するモデルを，機械学習を用いて構築する．その結果，本研究で用いた多様な言語情報がモーション生成に有用な情報であることが示唆された．次に，本技術を広く利用できるための試みとして，構築した生成モデルを用いてモーションを容易に生成可能な API を構築し，UNITY 上の CG キャラクタを発話言語のみから自動制御可能なデモシステムを構築した．本システムでは，任意の発話言語を入力すると，音声合成器およびモーション生成 API から合成音とモーション情報を取得し，UNITY の発声およびモーション付きアニメーションを生成する．このデモシステムを用いて，ユーザ主観評価実験を実施した結果，本技術により生成されたモーションを CG キャラクタに付与することで，動作の自然さ，発話と動作の一致度，エージェントへの好感，人間らしさなどの印象が向上することが確認された．

本モーション生成技術は，擬人化エージェントやヒューマノイドロボットを用いた対話システムだけでなく，テキストチャット上のアバタや，CG アニメーションにも幅広く適用が可能であり，今後，広く利用されることが期待される．

今後は，多様な系列情報を利用したアルゴリズムの構築や，より多様で詳細な身体モーションパラメータの生成に取り組む予定である．また，本研究では，CG キャラクタの動作の自然さ，動作と発話の一致性，好感，人間らしさといった基本的な評価のみに留まった．モーション生成による，より多様な効果についても検証を実施したい．

参考文献

[1] Beskow, J., Granstrom, B. and House, D.: Visual correlates to prominence in several expressive modes, *INTERSPEECH* (2006).

[2] BirdWhistell, R. L.: *Kinesics and context*, University of Pennsylvania Press (1970).

[3] Busso, C., Deng, Z., Grimm, M., Neumann, U. and Narayanan, S.: Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis, *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 1075–1086 (2007).

[4] Fuchi, T. and Takagi, S.: Japanese Morphological Analyzer using Word Cooccurrence -JTAG, *International conference on Computational linguistics*, pp. 409–413 (1998).

[5] Graf, H. P., Cosatto, E., Strom, V. and Huang, F. J.: Visual Prosody: Facial Movements Accompanying Speech, *IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 381–386 (2002).

[6] Higashinaka, R., Imamura, K., Meguro, T., Miyazaki, C., Kobayashi, N., Sugiyama, H., Hirano, T., Makino, T. and Matsuo, Y.: Towards an open-domain conversational system fully based on natural language processing, *International conference on Computational linguistics*, pp. 928–939 (2014).

[7] Ishi, C. T., Haas, J., Wilbers, F. P., Ishiguro, H. and Hagita, N.: Analysis of head motions and speech, and head motion control in an android, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 548–553 (2007).

[8] Ishi, C. T., Ishiguro, H. and Hagita, N.: Head motion during dialogue speech and nod timing control in humanoid robots, *ACM/IEEE International Conference on Human-Robot Interaction*, pp. 293–300 (2010).

[9] Iwano, Y., Kageyama, S., Morikawa, E., Nakazato, S. and Shirai, K.: Analysis of head movements and its role in spoken dialogue, *International Conference on spoken language*, pp. 2167–2170 (1996).

[10] Kadono, Y., Takase, Y. and Nakano, Y. I.: Generating Iconic Gestures Based on Graphic Data Analysis and Clustering, *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, Piscataway, NJ, USA, IEEE Press, pp. 447–448 (online), available from ⟨http://dl.acm.org/citation.cfm?id=2906831.2906920⟩ (2016).

[11] kenji Imamura: Analysis of Japanese dependency analysis of semi-spoken words by series labeling, *Proceedings of the Annual Meeting of the Association for Natural Language Processing*, pp. 518–521 (2007).

[12] KG, M., JA, J., DE, C., T, K. and E, V.-B.: Visual prosody and speech intelligibility: head movement improves auditory speech perception, Vol. 15, No. 2, pp. 133–7 (2004).

[13] Lohse, M., Rothuis, R., Gallego-Pérez, J., Karreman, D. E. and Evers, V.: Robot Gestures Make Difficult Tasks Easier: The Impact of Gestures on Perceived Workload and Task Performance, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, New York, NY, USA, ACM, pp. 1459–1466 (online), DOI: 10.1145/2556288.2557274 (2014).

[14] McNeill, D.: *Hand and Mind: What Gestures Reveal About Thought*, University Of Chicago Press (1996).

[15] Meguro, T., Higashinaka, R., Minami, Y. and Dohsaka, K.: Controlling listening-oriented dialogue using partially observable Markov decision processes, *International conference on computational linguistics*, pp. 761–769 (2010).

[16] NTT テクノクロス株式会社: FutureVoice.

[17] Schroff, F., Kalenichenko, D. and Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering, *CoRR*, Vol. abs/1503.03832 (online), available from ⟨http://arxiv.org/abs/1503.03832⟩ (2015).

[18] Senko Maynard: Interactional functions of a nonverbal sign: Head movement in Japanese dyadic casual conversation, *Journal of Pragmatics*, Vol. 11, pp. 589–606 (1987).

[19] Senko Maynard: Japanese conversation: Self-contextualization through structure and interactional management, *Norwood, New Jersey: Ablex Publishing Corporation* (1989).

[20] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A. and Sloetjes, H.: ELAN a Professional Framework for Multimodality Research, *International Conference on Language Resources and Evaluation* (2006).

[21] Yamamoto, M. and Watanabe, T.: Development of an Embodied Image Telecasting Method Via a Robot with Speech-Driven Nodding Response, *HCI*, Vol. (8), pp. 1017–1025 (2007).

[22] Yehia, H. C., Kuratate, T. and Vatikiotis-Bateson, E.: Linking facial animation, head motion and speech acoustics, Vol. 30, No. 3, pp. 555–568 (2002).