Vol.2011-HCI-142 No.4
2011/3/17

音声発話に伴う自然な頭部動作の生成における考察

劉超然[*1*2]　石井 Carlos 寿憲[*2]　石黒浩[*1*2]

人が音声発話する際，自然に頭部動作が伴う．これらの頭部動作が意図的に表出したものもあれば，無意識に行うものもある．頭部動作が意思・態度・感情などのノンバーバル情報の伝達に重要な役割を果たしていると考えられる．自然な頭部制御はヒューマンロボットインタラクション，特にヒューマンロボットコミュニケーションにおいて重要である．本研究ではひと同士の音声対話の分析結果をベースとした頭部動作生成モデルを提案した．二つの人間型ロボットに提案モデルを応用し，有効性を評価した．

Generation of head motions during spoken dialogue speech

Chaoran Liu[*1*2]　Carlos T. Ishi[*2]　Hiroshi Ishiguro[*1*2]

Head motions naturally occur in synchrony with speech and may carry paralinguistic information, such as intentions, attitudes and emotions, in dialogue communication. Natural head movements are of vital importance for the Human-Robot Interaction and especially Human-Robot Communication. The paper proposed a simple model for generating head tilts based on rules inferred from the analysis results, and evaluated it in two types of humanoid robot. Subjective scores showed that the proposed model could generate head motions with naturalness comparable to the original motions.

1. はじめに *

人とロボットの間で音声対話を介して円滑なコミュニケーションが成立するためには，言語情報の理解以上に，発話意図や話者の態度・感情などのパラ言語情報の理解も重要となる．

人間同士の対話では，発話をする際，自然に頭部動作が伴う．頭部動作は何らかの意図を示すため意識的に行う場合がある．特に話し相手の発言に対するリアクションとして頻繁に用いられる，例えば頷くことにより，同意や肯定を示し，首を横に振ることにより，否定を示す．一方，多くの場合は，発話に伴い，無意識に頭部動作が生じる．無意識に生じる頭部動作の場合も，発話となんらかの関連が存在すると考えられる．これらの頭部動作は態度，感情などのパラ言語情報を伝達する．

人間の頭部動作に含まれるパラ言語情報を完全に解明するのは難しいと思われる．しかし，自然な頭部を真似することによって，ロボットの存在感が高まり，より高度なヒューマンロボットコミュニケーションが期待出来る．

我々の研究のモチベーションとして，人型ロボット（アンドロイドなど）の遠隔操作において，音声に伴う頭部動作を音声信号から自動的に生成することを目的としている．そこで本研究では，音声に含まれる言語及びパラ言語情報と頭部動作との関連を調べた．

人型ロボットの発話に伴う自然な頭部動作を生成という目標に対し，３つのステップでアプローチしたいを考慮する．第１ステップは発話音声と頭部動作の関連を調べることである．第２ステップは，人型ロボットが自然な動作を再現できるかを調べることである．最後の第３ステップは，発話音声から頭部動作を生成することである．本論文では第２ステップについて述べる．

2. 関連研究

Yehia ら[1]は，英語と日本語の読み上げ文に対し，頭部動作からの F0 の推定は平均 70%以上推定可能だが，F0 からの頭部動作の推定においては英語の場合 50%，日本語においては 30%以下と報告している．Graf ら[2]は，英語の文発話に対し，頭部動作と音声の関連を調べ，文内の単語の強調は頷きが頻繁に伴い，頭部を上げる動作は声を上げることに対応すると報告している．彼らはこれらの動作を「視覚的韻律」(“visual prosody”)と呼んでいる．

*1 大阪大学基礎工学研究科
Graduate School of Engineering Science, Osaka University
*2 (株)国際電気通信基礎技術研究所　知能ロボティクス研究所
ATR Intelligent Robotics and Communication Laboratories

Beskow ら[3]は，スウェーデン語の読み上げ文に対し，強調する単語を変えながら，肯定，質問，迷い，楽しさ，怒りなどのさまざまな表現を変えた場合の発話と，頭部動作と表情を含んだ顔のパラメータとの関連を調べている．結果として，すべての表現において，強調された単語において，強調されていない単語よりも，顔のパラメータの分散が大きかったと報告している．また，岩野ら[4]は，視覚情報を利用して対話理解を向上させることを目的とし，日本語の対話音声における頭部動作の役割を分析している．発話権や発話意図が考慮され，肯定・同意・応答・相槌では縦方向の動作，相手に応答を求める場合は顔を上げる動作が頻繁に生じることを報告している．

我々の研究では，[5]で提案された発話機能タグに重心を置いた．

- k (keep)：発話権の保持（強い句境界）
- k2 (keep)：弱い句境界（発話権の保持）
- k3 (keep)：発話末を伸ばし、発話の途中であることを表現（発話権の保持）
- f (filler)：「えっとー」「あのー」など、考え中であることを表現する感動詞
- f2 (conjunctions)：「じゃ」などの接続詞（短いフィラーとして捉えられる）
- g (give)：対話相手への発話権の譲渡
- q (question)：発話権の譲渡（対話相手に応答を求める場合）
- bc (backchannels)：「うん」「はい」などの相槌を表現する感動詞
- su (surprise/admiration)：「えー！」「へー」など、驚きや感心などの感情を表現する感動詞
- dn (denial)：「いいえ」「ううん」などの否定を表現する感動詞

前述のように，音声発話の韻律情報と頭部動作の強い相関関係が見られない．しかし，韻律情報と発話機能タグの強い相関関係が報告された[5][6][7]．そして発話機能タグと頭部動作の関係が[8][9]で報告された．この関連関係をベースとした頷き生成モデルが[9]で提案され，有効性が検証された．

3. 首かしげの生成モデル

3.1 首傾げの生成

頭部動作と発話機能の分析結果[8][9]が示したように，人間の自然な動作の中に，弱い句境界において，頭部が動かない傾向が強い．しかし，話者が発話文の文末音素を伸ばしている，つまり考えていると思われるところ(発話機能タグの k3,f)に首傾げは高い確率で見られる．このセッションでは，新たな首傾げ生成モデルを紹介する．

首傾げモデルは，既存の頷きモデルに対し，特定の弱い句境界(k3,f)に，首傾げ動作を加えたモデルである．ベースは[9]で提案された頷きモデルとなり，まず強い句境界(k,g,q)において頷きを生成させる．

それに加え，データベースから首傾げのモーションサンプルを利用し，弱い句境界の k3,f に加えた．この動作の特徴は，始まりに 0.6 秒を掛けて，頭部の roll 軸を元の位置から 15 度まで回転し，終わりは同様のスピードで元の位置まで戻す．生成タイミングの制御は句境界に首傾げを開始させ，その後傾げる角度を保ち，次のフレーズの終了時刻までに頭を元の位置に戻す．

3.2 実験設定

ヒューマノイドの Robovie R2 とアンドロイドの Geminoid F を用いて，首傾げ生成モデルによって生成した頭部動作の自然さを検証する実験を行う．

データベースから無作為に 11 の弱い句境界 k3・f を含む 10 秒から 20 秒発話サンプルを抜き出し，頷き生成モデルと首傾げ生成モデルを用いて，発話毎に 2 種類の動作を生成した．比較対象として，モーションキャプチャーシステムによって記録した話者のオリジナルモーションを加えて，3 種類の動作パタンをヒューマノイド Robovie R2 と女性アンドロイド Geminoid F に再生させた．使用したロボットの写真は下に示す．

Figure 2. Robovie R2 and Geminoid F

話者のオリジナルモーションをロボットにマッピングする際，Robovie R2 は首に 3 つの回転自由度を持っていて，線形的にマッピングすることができる．アンドロイド Geminoid F へのマッピングは，[10]で提案された手法を利用する．ロボットに動作指令を送る時間間隔は，ロボットのハードウェアの制限に従って，Robovie R2 の場合は 100ms で，Geminoid F の場合は 20ms．Robovie R2 では口が動かないため，頭部の回転のみを再生するには 100ms の時間間隔でも支障は生じないと考えられる．Geminoid F

口の動きはモーションキャプチャーが撮った話者の鼻と顎に貼り付けたマーカーの距離の変化から求められる．

41 人の被験者（男性 17 人，女性 24 人）を招いで，11 サンプル・各 3 パタンの動作，計 33 の動作を 2 種類のロボットに再生させ，発話のビデオを被験者に見せ，アンケートを取った．

比較させたペアは以下に示す

- 頷きモデル vs. 首傾げモデル
- 首傾げモデル vs. オリジナル
- 頷きモデル vs. オリジナル

ペア内の二つのビデオの順番は一定ではなく，ランダムになっている．

アンケートは 7 段階評価の様式にした．

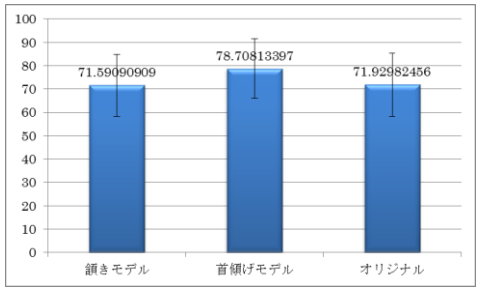
- ロボットの動作が自然か否か

大変自然(7) | 自然(6) | やや自然(5) | どちらとも言えない(4) | やや不自然(3) | 不自然(2) | かなり不自然(1)

- 二つの動作を比較して，どちらが自然?

一つ目の方がずっと自然 | 一つ目の方が自然 | 一つ目の方がやや自然 | どちらとも言えない | 二つ目の方がやや自然 | 二つ目の方が自然 | 二つ目の方がずっと自然

3.3 実験結果



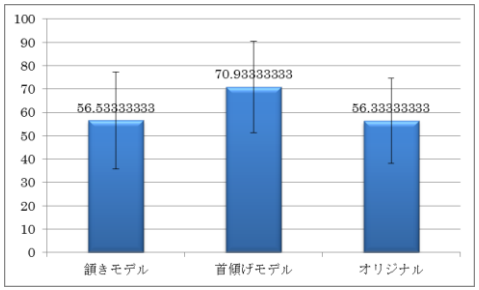


Figure 3. Distributions of the preference scores for each motion type, for Geminoid F (top) and Robovie R2 (bottom)

図 3 は自然さに関するアンケート結果を括弧の中の数字で数値化し，その値を 0-100 の範囲内に標準化した結果を示す．二つのロボットともに自然さに関するスコアが最も高いのは提案の首傾げモデルである．この結果は首傾げを弱い句境界の k3,f で生成することで，ロボットの動きが自然に見えることを示した．首傾げのタイミング制御の正確性の裏づけになると思われる．そして，[9]で報告された頷きモデルの検証実験の結果と同じく，オリジナルモーションが単純な頷きモデルの生成モーションとほぼ同様のスコアになっている．この結果に関しては後ほど議論する．

二つのロボットに同様の動作を再生したものの，自然さに関するスコアが明らかにヒューマノイド Robovie R2 の方が低かった．ロボットの外見による差が出ていると思われるが，もう一つ大きな理由として，Robovie R2 の口が動かないことだと考えられる．Geminoid F の方が口の動きが人間話者の口の動きを再現したため，ロボットが発話する際，被験者が視覚・聴覚両方から認識することができる．一方，Robovie R2 の場合は，聴覚から発話を検知しても，最初に述べたように，頭部動作と音声の韻律情報の関連性が低いため，視覚的に頭部の回転だけからロボットが発話をしているという再確認が難しい．それが故に Robovie R2 での自然さに関する評価が低くなっていると考えられる．

Vol.2011-HCI-142 No.4
2011/3/17

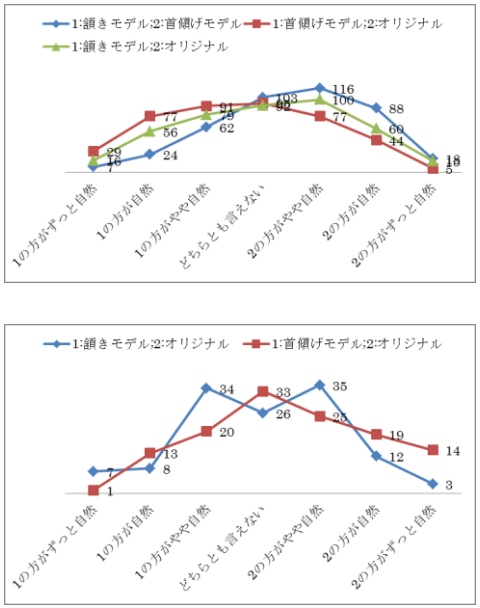


Figure 4. Distributions of the preference scores for each pair of motion types, for Geminoid F (top) and Robovie R2 (bottom)

図 4 は各ペア内二つの動作のどちらの方が自然という問題に対して，各選択肢を選んだ人数を表す．

Geminoid F における結果は，頷きモデルと首傾げモデルの比較で多くの被験者が首傾げモデルの方が自然またはやや自然という答えを選んだ．一方，頷きモデルとオリジナルモーションの比較で，どちらとも言えないと答えた人数が最も多く，各答えを選んだ被験者数の分布がほぼ対称的であった．この結果は各動作パタンの自然さの評価スコアとも一致した．首傾げモデルとオリジナルモーションの比較では，[15]で報告された結果に似ていて，オリジナルモーションの方が自然と答えた被験者もいるが，首傾げモデルの方が自然と答えた被験者ほど多く無い．この結果は，複雑な共振除去制御を施さない条件では，シンプルで且つタイミング正しい動作生成モデルのほうが効果的だという結論に繋がる．

Robovie R2 における結果は，すべて対称的なクラフになっていて，つまり二つの動作の区別が付きにくいという結果になっている．その理由として，前述のように，口の持たないロボットの場合，被験者はロボットが発話をしているというはっきりした視覚情報を得るのが難しい．聴覚情報だけでロボットの存在感が十分に伝えなく，自然さの比較も難しくなる理由となる．

4. おわりに

人とロボットの自然な対話インタラクションを実現するには，ロボットも発話に伴って自然な頭部動作を行うことが重要である．

本研究では，人間の対面対話における頭部動作の分析結果に基づいて，ルールベースの首傾げ生成モデルを提案した．これらの生成モデルを二種類の人間型ロボットに応用し，評価実験によってその自然性を確認した．

今後の予定は，韻律・言語情報から発話中の句境界の種類の推定を検討し，本研究の生成モデルを利用し，音声情報のみから自然な頭部動作の生成を試みる．

謝辞 本研究は，JST CREST の委託により実施したものである。

参考文献

1) H.C. Yehia, T. Kuratate, E. Vatikiotis-Bateson, “Linking facial animation, head motion and speech acoustics,” J. of Phonetics, Vol. 30, pp. 555-568, 2002.
2) H.P. Graf., E. Cosatto, V. Strom, F.J. Huang, “Visual prosody: Facial movements accompanying speech,” Proc. IEEE Int. Conf. on Automatic Face and Gesture Recognition (FGR’02), 2002.
3) J. Beskow, B. Granstrom, D. House, “Visual correlates to prominence in several expressive modes,” Proc. Interspeech 2006 – ICSLP, pp. 1272-1275, 2006.
4) Y. Iwano, S. Kageyama, E. Morikawa, S. Nakazato, K. Shirai, “Analysis of head movements and its role in spoken dialogue,” Proc. ICSLP’96, pp. 2167-2170, 1996.
5) C.T. Ishi, H. Ishiguro, N. Hagita, “Analysis of prosodic and linguistic cues of phrase finals for turn-taking and dialog acts,” Proc. Interspeech’2006 - ICSLP, pp. 2006-2009, 2006.
6) C.T. Ishi, “Perceptually-related F0 parameters for automatic classification of phrase final tones,” IEICE Trans. Inf. & Syst., Vol. E88-D, No. 3, pp. 481-488, 2005.
7) C.T. Ishi, H. Ishiguro, N. Hagita, “Evaluation of prosodic and voice quality features on automatic extraction of paralinguistic information,” Proc. IROS 2006, pp. 374-379, 2006.
8) C.T Ishi, J. Haas, F.P. Wilbers, H. Ishiguro, and N. Hagita, “Analysis of head motions and speech, and head motion control in an android,” Proc. of IROS 2007, 548-553, 2007.
9) C.T Ishi, C. Liu, H. Ishiguro, and N. Hagita, (2010). “Head motion during dialogue speech and nod timing control in humanoid robots,” Proc. of HRI 2010, 293-300, 2010.
10) F.P. Wilbers, C.T. Ishi, H. Ishiguro, “A blendshape model for mapping facial motions to an android,” Proc. IROS 2007, 2007.