The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

# 発話言語に基づく頷き動作生成の検討

Generating Head Nods using Linguistic Information

石井亮 Ryo ISHII　片山太一 Taichi KATAYAMA　小林のぞみ Nozomi KOBAYASHI　西田京介 Kyosuke NISHIDA　東中竜一郎 Ryuichiro HIGASHINAKA
富田準二 Junji TOMITA

*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
NTT Media Intelligence Laboratories, NTT Corporation

## 1. はじめに

人間のコミュニケーションにおいて，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 [1]．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非言語行動を表出し，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．我々は，擬人化エージェントやヒューマノイドロボットにおいて，発話内容に基づいて，人間と同様な適切なタイミングで非言語行動を自動生成することに取り組んでいる．

非言語行動の中で，頭部による頷き動作は，対話相手へのフィードバックだけではなく，発言の肯定，強調，リズム取り，発話の意図の提示といった様々な機能を有することが知られている [13, 14]．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頷き動作を付与することで，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，発話に付随する頷きは，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 [11]．

このような背景から，特に発話音声情報を用いて，発話中の頷き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 [2, 3, 5, 8, 9, 10, 17]．しかし，発話の音声情報から精度良く頷き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, 17]．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頷き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．

これに対して本研究では，発話言語から得られる多様な言語解析情報を利用して頷き動作を生成することに取り組む．頷きは，特定の単語，発話の意図などの言語情報との拘わりが強いと考えられる．言語情報から頷き動作を生成した試みとして，文節末の単語と頷きの共起関係が分析されている．具体的に，反復，フィードバック，話者交替などに関連した単語（感動詞や発話の末尾に現れる助詞）に対して，頷きが共起しやすい傾向が示されており，これらの単語情報を用いた，頷きの生成アルゴリズムが提案されている [8]．これは，我々の言語情報に基づく頷き生成のアプローチを支持する先行知見であり，本研究はより多様な言語情報による頷き生成に取り組む．具体的に

連絡先: 石井亮，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，239-0847 神奈川県横須賀市光の丘 1-1，ishii.ryo@lab.ntt.co.jp

図 1: 2 者対話の様子

本研究では，先行研究では扱われていない，発話言語に含まれる発話末以外の単語，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を利用して，頷き生成モデルを構築する．このような多様な言語情報と頷き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．

本研究では，最初に，2 者対話を収録し，発話および頷き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，文節単位ごとに頷きを生成するモデルを構築する．その結果，本研究で用いた多様な言語情報が頷き生成に有用な情報となることを示す．

## 2. マルチモーダルコーパス

2 者対話を対象に，発話言語およびそれに伴う頷きデータを含む，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．参加者は互いに向き合って着座した．対話内容は，発話に伴う頷き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）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 1 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，発話内容から文を分割した．さらに，係り受け解析エンジン [19] 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．
- 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．
- 注視対象：参加者はグラス型の視線計測装置（Tobii Glass2）を装着しており，人手で視野カメラ上の参加者の注視点から注視対象をラベリングした．
- 身体の位置・姿勢：モーションキャプチャ装置（Xsens MVN）を用いて，全身の部位の位置・姿勢を測定した．

人手アノテーションには ELAN [16] 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．

## 3. 頷き生成モデル

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，文節単位ごとに頷きを生成するモデルを構築した．それぞれの言語情報が有効であるかを検証するために，各特徴量を用いた生成モデルと，全ての情報を用いたモデルを構築した．具体的に，文節単位ごとに，対象となる文節およびその前後の文節から得られる言語特徴量を入力として，頷きの有無の 2 値を出力する生成モデルを，決定木アルゴリズム C4.5 [20] を用いて実装した．使用した特徴量は以下の通りである．

- 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 [18] 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 [12, 6] 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を用いて，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，他者のデータからどれだけ頷きを生成できるかを評価した．なお，各施行で頷きの有無データは，データ量が一致するように，データ数の少ない方に合わせてデータ数を削減をした．よって，チャンスレベルは 0.50 となる．性能評価結果の平均値を表 1 に示す．まず，一つの情報のみを使用した際には，文字数，位置，品詞，シソーラス，発話行為の情報のみを用いたモデルの性能は，チャンスレベルよりも高かった．このうち，語彙大系の情報が最も有用であった．一方，単語情報のみを用いたモデルはチャンスレベルよりも低かった

| 特徴量 | 適合率 | 再現率 | F 値 |
|---|---|---|---|
| チャンスレベル | 0.500 | 0.500 | 0.500 |
| 文字数 | 0.561 | 0.556 | 0.558 |
| 位置 | 0.526 | 0.528 | 0.527 |
| 単語 | 0.357 | 0.529 | 0.431 |
| 品詞 | 0.522 | 0.528 | 0.525 |
| 語彙大系 | 0.615 | 0.538 | 0.579 |
| 発話行為 | 0.513 | 0.533 | 0.523 |
| 全て | 0.578 | 0.601 | 0.593 |

表 1: 頷き生成に用いた特徴量とモデルの評価結果

（対応のある t 検定の結果：$p < .05$）．また，全ての情報を利用したモデルの性能は言語情報単体で使用したものよりも性能が高かった（対応のある t 検定の結果：$p < .05$）．これらの結果から，発話言語から得られる，文字数，位置，品詞，語彙大系，発話行為の情報は頷き生成に有効であることが示唆された．また，これらの情報を統合して利用することで，より高精度に頷きを生成することが可能であることが示唆された．

## 4. 議論

言語情報を用いた頷き生成モデルの評価の結果，単語情報が頷き生成に有用では無かった．一方で，単語の上位概念クラスであるシソーラス情報は有用であった．その理由として，単語の絶対数に対して学習データ量が十分に多くなかったことが考えられる．一方で，シソーラス情報は，単語を意味や属性でクラス化したものであるため，本研究で使用したデータ数であっても効率的に学習ができたと考えられる．膨大な量のマルチモーダルデータを収集するのはコストがかかり困難であるため，単語情報を利用することは現状では困難であるかもしれない．これに対して，シソーラス情報は比較的少ないデータ量でも十分に学習できる可能性が考えられる．

本研究では，該当文節とその前後の文節から得られる言語情報のみを用いて，頷きを生成した．今後は，文全体の文節から得られる系列情報を用いるなど，アルゴリズムの改良が望まれる．

## 5. まとめ

本研究では，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を利用して，頷き生成モデルを構築した．その結果，文字数，位置，品詞，語彙大系，発話行為の情報は頷き生成に有効であること．また，全ての言語情報を利用することが有効であることが示唆された．

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，頷きの詳細なパラメータの生成に取り組む予定である．また，擬人化エージェントや対話ロボットへ実装し，その効果を検証したい．

## 参考文献

[1] Ray L. BirdWhistell. Kinesics and context. University of Pennsylvania Press, 1970.

[2] Jonas Beskow, Bjorn Granstrom, and David House. Visual correlates to prominence in several expressive modes. In Proceedings of INTERSPEECH, 2006.

[3] Carlos Busso, Zhigang Deng, Michael Grimm, Ulrich Neumann, and Shrikanth Narayanan. Rigid head motion in expressive speech animation: Analysis and synthesis. In IEEE Transactions on Audio, Speech, and Language Processing, pages 1075–1086, 2007.

[4] Takeshi Fuchi and Shinichiro Takagi. Japanese morphological analyzer using word cooccurrence -Jtag. In Proceedings of International conference on Computational linguistics, pages 409–413, 1998.

[5] Hans Peter Graf, Eric Cosatto, Volker Strom, and Fu Jie Huang. Visual prosody: Facial movements accompanying speech. In Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition, pages 381–386, 2002.

[6] Ryuichiro Higashinaka, Kenji Imamura, Toyomi Meguro, Chiaki Miyazaki, Nozomi Kobayashi, Hiroaki Sugiyama, Toru Hirano, Toshiro Makino, and Yoshihiro Matsuo. Towards an open-domain conversational system fully based on natural language processing. In Proceedings of International conference on Computational linguistics, pages 928–939, 2014.

[7] Carlos T. Ishi, Judith Haas, Freerk P. Wilbers, Hiroshi Ishiguro, and Norihiro Hagita. Analysis of head motions and speech, and head motion control in an android. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 548–553, 2007.

[8] Carlos T. Ishi, Hiroshi Ishiguro, and Norihiro Hagita. Head motion during dialogue speech and nod timing control in humanoid robots. In Proceedings of ACM/IEEE International Conference on Human-Robot Interaction, pages 293–300, 2010.

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

[10] Munhall KG, Jones JA, Callan DE, Kuratate T, and Vatikiotis-Bateson E. Visual prosody and speech intelligibility: head movement improves auditory speech perception. 15(2):133–7, 2004.

[11] Manja Lohse, Reinier Rothuis, Jorge Gallego-Pérez, Daphne E. Karreman, and Vanessa Evers. Robot gestures make difficult tasks easier: The impact of gestures on perceived workload and task performance. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'14), pages 1459–1466, 2014.

[12] Toyomi Meguro, Ryuichiro Higashinaka, Yasuhiro Minami, and Kohji Dohsaka. Controlling listening-oriented dialogue using partially observable markov decision processes. In Proceedings of International conference on computational linguistics, pages 761–769, 2010.

[13] Senko Maynard. Interactional functions of a nonverbal sign: Head movement in japanese dyadic casual conversation. Journal of Pragmatics, 11:589–606, 1987.

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

[15] Tomio Watanabe, Ryusei Danbara, and Masashi Okubo. Effects of a speech-driven embodied interactive actor interactor on talker's speech characteristics. In Proceedings of IEEE International Workshop on Robot-Human Interactive Communication, pages 211–216, 2003.

[16] P. Wittenburg, H. Brugman, A. Russel, A Klassmann, and H Sloetjes. Elan a professional framework for multimodality research. In Proceedings of International Conference on Language Resources and Evaluation, 2006.

[17] Hani Camille Yehia, Takaaki Kuratate, and Eric Vatikiotis-Bateson. Linking facial animation, head motion and speech acoustics. Journal of Phonetics, 30(3):555–568, 2002.

[18] 池原 悟, 宮崎 正弘, 白井 諭, 横尾 昭男, 中岩 浩巳, 小倉 健太郎, 大山 芳史, 林 良彦，日本語語彙大系. 岩波書店，1997.

[19] 今村賢治，系列ラベリングによる準話し言葉の日本語係り受け解析．言語処理学会第 13 回 年次大会発表論文集，pp.518-521，2007.

[20] J. R. Quinlan. Improved use of continuous attributes in c4.5. Journal of Artificial Intelligence Research, 4:77-90, 1996.