聞き手の相槌種類に応じた表情生成システムの基礎検討

鹿摩 大智[1]　岡 哲平[1]　大西 俊輝[2]　東 直輝[1]　石井 亮[3]　深山 篤[3]　宮田 章裕[1,a)]

概要：対話において聞き手の相槌は，円滑なコミュニケーションを実現するための重要な要素の 1 つである．適切な相槌を打つことは人同士の対話だけでなく，人とエージェントの対話にも重要なのは明らかである．近年の対話エージェントの研究では，話し手のマルチモーダル情報に基づいて，返答の振舞いの生成が行われている．しかし，相槌の機能の明示的な分類に基づいていないため，生成される相槌が会話の流れに合わない可能性がある．そこで本稿では，相槌の言語的・機能的側面に基づいた相槌を生成する取り組みの初期検討として，聞き手が相槌を打つ際の表情の生成を行う．具体的には言語的・機能的側面に応じた相槌を打つ際の表情を生成し，ユーザに提示するシステムを提案する．

1. はじめに

対話において聞き手が相槌を打つことは，コミュニケーションを円滑に進めるために必要不可欠である．日本語における相槌を表す言葉は豊富であり [1]，機能的な側面に着目した相槌の分類を行う研究 [2] や，言語的・機能的な側面に着目して話し手の発話意図や種類と聞き手の相槌の関係を分類する研究 [3] が行われている．さらに近年，対話エージェントとインタラクションを行う機会が増えている．このような背景から，人とエージェントの対話においても聞き手が適切な相槌を打つことが重要であると考えられる．相槌を生成する取り組みは数多く行われている [4], [5], [6], [7], [8], [9]．しかし，相槌が持つ言語的・機能的な側面に基づいて相槌を生成する取り組みは我々が調査した範囲では見つかっていない．相槌が持つ言語的・機能的側面に基づかないことで，生成される相槌が会話の流れに合わない可能性がある．そこで我々は，対話における話し手のマルチモーダル情報から相槌の機能の明示的な分類に基づいた聞き手の相槌生成に取り組む．先行研究 [10], [11], [12] では，対話における話し手のマルチモーダル情報から聞き手の相槌の種類を推定する機械学習モデルを構築した．本稿では，相槌が持つ言語的・機能的な側面に基づいた相槌を生成する取り組みの初期検討として，聞き手が相槌を打つ際の表情の生成を行う．

[1] 日本大学文理学部
[2] 日本大学大学院総合基礎科学研究科
[3] 日本電信電話株式会社 NTT 人間情報研究所
[a)] miyata.akihiro@acm.org

2. 関連研究

言語・非言語行動を利用して対話中における相槌の生成を行う研究は多く行われている．本研究は，言語・非言語行動を利用して相槌を生成する研究と関連している．Lala らは韻律的特徴量を用いて 3 つに分類された相槌の種類と相槌が打たれるタイミングを予測したモデルをトレーニングし，相槌の生成を行うモデルを構築した [6]．加えて，言語的特徴量を含めることでモデルの性能が向上することを示唆していた．Dermouche らは視覚的，韻律的特徴量を用いて対話エージェントの応答として笑顔の強度，頭部の動き，視線を生成するモデルを構築した [7]．Jonell らは視覚的，韻律的特徴量を用いて対話におけるエージェントの顔の動きを生成するモデルを構築した [8]．加えて，言語的特徴量を含めることでモデルの性能が向上することを示唆していた．Bucci らは言語的特徴量を用いてエージェントの表情を生成するモデルを構築した [9]．

3. 研究課題

2 章で述べたように，これまで相槌の言語的・非言語的な生成に関する研究事例が行われている．これらの研究事例では，対話における言語もしくは非言語情報を用いること，さらには言語と非言語情報を合わせて用いることで主要なカテゴリの相槌や非言語的な相槌を生成できることが明らかになっている．しかし，これらは相槌の機能の明示的な分類に基づいて生成を行っていないため，生成される相槌が会話の流れに合わない可能性がある．相槌の機能の明示的な分類に基づき，相槌の機能別の特徴を生成した振る舞いに付与することで，より自然な相槌を生成できるの

ではないかと考える．そこで，我々は対話における話し手のマルチモーダル情報から聞き手の相槌を言語的・機能的側面から生成する取り組みを行う．本稿では，上記の取り組みの初期検討として，相槌の明示的な分類に基づく表情生成の効果を明らかにすることを研究課題とする．

4. 提案手法

3 章で述べた研究課題を達成するために，相槌の機能に基づいて生成を行う．そして生成した表情について，第三者への印象評価実験から自然な相槌であるかどうかを確認する．この取り組みを行うために，相槌の明示的な分類ごとに実対話データにおける話者の表情を分析し，分析結果に基づいて表情を生成および提示する必要がある．そこで本稿では，実対話データにおける話者の表情を用いて相槌の明示的な分類に基づいた表情を提示するシステムを提案する．

5. 対話コーパス

本研究では既存の 2 者対話コーパスを利用する．この対話コーパスには，2 者対話データ [13] と相槌ラベル [3] が記録されている．

5.1 2 者対話コーパスについて

2 者対話の参加者は，初対面の日本人男女合計で 26 名（異なるペアを 13 組）である．発話を含んだ相槌のデータをより多く収集するために，アニメ「トムとジェリー」を視聴した一方の参加者（話し手）が他方の参加者（聞き手）に内容を説明するタスクを行っている．発話の単位には Inter-pausal units（IPU）[14] を使用し，沈黙時間が 200ms 未満の連続した音声区間を 1 つとしている．この対話データでは，話し手が 4,940 件，聞き手が 2,865 件の合計 7,805 件の IPU が記録されている．

5.2 相槌ラベルについて

対話に参加していない第三者のアノテータ 3 名が，聞き手の発話ごとに下記に示す 9 種の相槌ラベルを付与している．なお，アノテータは 1 つの発話に対し複数の相槌ラベルを付与することが許可されている．

- N（Neutral word）:「うん」,「はい」,「おお」など話し手への感情を含まない応答.
- P（Positive word）:「うんうん」,「そうそう」,「それいい」,「なるほど」,「たしかに」など話し手への肯定的な応答.
- NP（Non-positive word）:「うーん」,「ふーん」,「はーん」,「あー」,「へー」,「んー」など話し手への否定的または悩んでいるような応答.
- E（Emotional word）:「すごい」,「ふふ」,「あぁ」,「へぇ」，その他短い感嘆詞など感情の動きを表しているような応答.
- A(Anticipation)：話し手の話題を先取りしている応答.
- C（Confirmation）:「えっ」,「はっ」,「あっ」,「なんで」など確認を促す，質問するような応答.
- R（Repetition of speaker's utterance）：話し手の発言を繰り返す応答.
- S（Summary of speaker's utterance）：話し手の要約，および言い換えをしているような応答.
- O(Other)：聞き手の感想や独り言など，他に該当するラベルがない応答.

6. 実装

本稿では，3 章で述べた研究課題を解決する手法として，ユーザが相槌ラベルを選択し，2 者対話データに基づいて生成した相槌種類ごとの表情を提示するシステムを提案する．相槌種類ごとの表情を提示することにより，より自然な相槌を生成できると考えられる．本研究の手順は以下のようになる．

- **Step1:** 特徴量抽出
  対話コーパスから相槌種類ごとに視覚的特徴量を抽出する．
- **Step2:** 顔画像の生成
  抽出した視覚的特徴量から顔画像を生成する．
- **Step3:** 相槌種類ごとの表情を提示するシステムの実装
  ユーザが選択した相槌種類に基づいた顔画像を提示するシステムを実装する．

6.1 特徴量抽出について

本研究では，聞き手の多様な相槌の表情を生成するために，5.1 節の対話データから聞き手の発話時の視覚的特徴量を抽出した．

6.1.1 視覚的特徴量について

対話データの映像から顔画像処理ツールである OpenFace[15] を用いて話し手の Action Units[16] を抽出した．抽出する特徴量として，OpenFace で用いられている各 Action Units(表 1) の強度の分散，中央値，10 パーセンタイル値，90 パーセンタイル値を用いた．

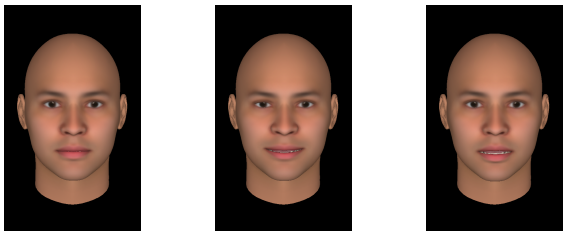
6.2 提案システムについて

6.2.1 顔画像の生成

顔画像の生成には，Action Units に基づいた人間の顔の 3DCG モデルを作成することが可能な FaceGen[17] を用いた．本稿では 5.2 節で述べた相槌ラベルのうち，話し手への感情を含まない応答である N と話し手への肯定的な応答である P，話し手への否定的または悩んでいるような応答である NP を採用した．相槌ラベル別に 3DCG モデルを作成するにあたり，「はい」や「うん」など，発話内容や対話の流れに応じて表情強度に差異がある可能性がある．そ

表 1: Action Units の内容

項目	内容	項目	内容
AU01	眉の内側を上げる	AU14	笑窪を作る
AU02	眉の外側を上げる	AU15	唇の両端を下げる
AU04	眉を下げる	AU17	顎を上げる
AU05	上瞼を上げる	AU20	唇の両端を横に引く
AU06	頬を持ち上げる	AU23	唇を固く閉じる
AU07	瞳を緊張させる	AU25	顎を下げずに唇を開く
AU09	鼻に皺を寄せる	AU26	顎を下げて唇を開く
AU10	上唇を上げる	AU45	瞬きをする
AU12	唇の両端を引き上げる		



(a) ラベル N　(b) ラベル P　(c) ラベル NP

図 1: 生成された顔画像

表 2: 頻出度の高い発話内容

ラベル	頻出度の高い発話内容
N	うん，はい，あー，ふーん，ええ
P	うんうん，はいはい，そうそう，なるほど，そっかそっか
NP	うーん，なんか，えーっと，いや，やっぱ

こで，5.1 節で述べた 2 者対話データから，相槌ラベル別で頻出度の高い発話内容を 5 つ (表 2) 抽出し，OpenFace を用いて相槌の種類ごとに頻出度の高い発話内容を話す聞き手の映像データから Action Units を抽出し，統計量の中央値を算出した．算出された統計量をもとに顔画像の生成を行った．図 1 は 5.1 節の対話データの中から各ラベルが付与された対話を無作為に 1 つ選定し，抽出した特徴量から生成した顔画像である．図 1a はラベル N，図 1b はラベル P, 図 1c はラベル NP が付与された顔画像である．

6.2.2　入出力部

ユーザは 6.2.1 項で採用した N，P，NP のいずれかの相槌ラベルから 1 つを選択する．図 2 は本システムの入力例を示す．選択された相槌ラベルから 6.2.1 項で作成した顔画像をユーザに提示する．図 3 は本システムの出力例を示す．

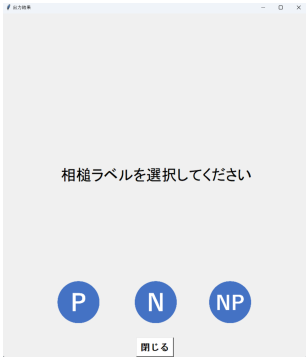


図 2: システムの入力例

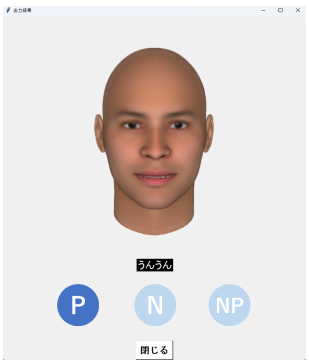


図 3: システムの出力例

7.　おわりに

本稿では相槌が持つ言語的・機能的な側面に基づいて相槌を生成する取り組みの初期検討として，言語的・機能的な側面に基づいた表情の生成を行った．具体的には聞き手の発話時の視覚的特徴量を用いて顔画像を生成し，ユーザが選択した相槌ラベルに応じた顔画像を提示するシステムを実装した．

今後は言語的・機能的な側面に基づいた表情の生成のために，相槌種類の予測と相槌種類に基づいた表情強度を生成する機械学習モデルを構築した上でシステムに組み込み，システムの評価実験に取り組む予定である．また，本稿では対話中における聞き手の発話時の表情のみに焦点を当てたが，自然な相槌を打つエージェントを実現するためには，韻律的・言語的特徴量についても生成を行うことができるのか検討する必要がある．今後さらに研究を進めて，自然な相槌を打つ対話エージェントを実装することが可能なのか検討を行っていきたい．

参考文献

[1] Maynard, S. K.: On back-channel behavior in Japanese and English casual conversation, *Linguistics*, Vol. 24, No. 6, pp. 1079–1108 (1986).

[2] Mukai, C.: The Use of Back-channels by Advanced Learners of Japanese: Its Qualitative and Quantitative Aspects, *Japanese language education around the globe*, Vol. 9, pp. 197–219 (1999).

[3] Morikawa, A., Ishii, R., Noto, H., Fukayama, A. and Nakamura, T.: Determining Most Suitable Listener Backchannel Type for Speaker's Utterance, *Proc. 22nd ACM International Conference on Intelligent Virtual Agents (IVA '22)*, pp. 1–3 (2022).

[4] Kawahara, T., Uesato, M., Yoshino, K. and Takanashi, K.: Toward Adaptive Generation of Backchannels for Attentive Listening Agents, *Proc. 6th International Work-*

shop on Spoken Dialog System (IWSDS '15), pp. 1–10 (2015).

[5] Park, H. W., Gelsomini, M., Lee, J. J. and Breazeal, C.: Telling stories to robots: The effect of backchanneling on a child's storytelling, *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction*, pp. 100–108 (2017).

[6] Lala, D., Inoue, K., Kawahara, T. and Sawada:, K.: Backchannel Generation Model for a Third Party Listener Agent, *Proc. 10th International Conference on Human-Agent Interaction (HAI '22)*, pp. 114–122 (2022).

[7] Dermouche, S. and Pelachaud:, C.: Generative Model of Agent's Behaviors in Human-Agent Interaction, *2019 International Conference on Multimodal Interaction (ICMI '19)*, pp. 375–384 (2019).

[8] Jonell, P., Kucherenko, T., Henter, G. E. and Beskow, J.: Let's Face It: Probabilistic Multi-modal Interlocutor-aware Generation of Facial Gestures in Dyadic Settings, Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents (IVA '20), pp. 1–8 (2020).

[9] Bucci, B., Rossi, A. and Rossi:, S.: Action Unit Generation through Dimensional Emotion Recognition from Text, *31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1071–1076 (2022).

[10] 東　直輝, 大西俊輝, 木下峻一, 石井　亮, 深山　篤, 中村高雄, 宮田章裕: マルチモーダル情報に基づく多様な相槌の生成の基礎検討, 情報処理学会研究報告グループウェアとネットワークサービス (GN), Vol. 2023-GN-119, No. 8, pp. 1–6 (2023).

[11] 東　直輝, 大西俊輝, 木下峻一, 石井　亮, 深山　篤, 中村高雄, 宮田章裕: マルチモーダル情報に基づく多様な相槌の予測の検討, 情報処理学会シンポジウム論文集, マルチメディア, 分散, 協調とモバイル (DICOMO '23), Vol. 2023, pp. 352–358 (2023).

[12] Onishi, T., Azuma, N., Kinoshita, S., Ishii, R., Fukayama, A., Nakao, T. and Miyata:, A.: Prediction of Various Backchannel Utterances Based on Multimodal Information, *Proc. the 23rd ACM International Conference on Intelligent Virtual Agents(IVA '23)* (2023).

[13] Ishii, R., Higashinaka, R. and Tomita:, J.: Predicting Nods by using Dialogue Acts in Dialogue, *Proc. 11th International Conference on Language Resources and Evaluation (LREC '18)*, pp. 2940–2944 (2018).

[14] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A. and Den:, Y.: An Analysis of Turn-Taking and Backchannels Based on Prosodic and Syntactic Features in Japanese Map Task Dialogs, *Language and Speech*, Vol. 41, pp. 295–321 (1998).

[15] Baltrusaitis, T., Zadeh, A., Lim, Y. C. and Morency:, L.-P.: OpenFace 2.0: Facial Behavior Analysis Toolkit, *13th IEEE international conference on automatic face and gesture recognition (FG '18)*, pp. 59–66 (2018).

[16] Ekman, P. and Friesen:, W. V.: Manual for the Facial Action Coding System, *Palo Alto: Consulting Psychologists Press* (1977).

[17] Inversions, S.: FaceGen Modeller, https://facegen.com/modeller.htm. (accessed 2023/12/6).